Part 4: Convergence of random variables and limits theorems

Probability Theory: Math 605, Fall 2024

Luc Rey-Bellet
University of Massachusetts Amherst

2024-12-09

1 Convergence of random variables

We think of a (real-valued) random variable as a function $X(\omega)$ and so if we have a sequence of RVs $\{X_n\}$ we can define various types of convergence.

- Convergence almost sure
- Convergence in probability
- Convergence in L^p

In subsequent chapters we are will study another type convergence, namely weak convergence (also called convergence in distribution). It is of quite different type because it based on on the distribution of X and not on the notion of random variable as a function.

1.1 Almost sure convergence

Definition 1.1 (Almost sure convergence) A sequence of RVs $\{X_n\}$ *converges almost surely to a RV* X if

$$\lim_n X_n(\omega) = X(\omega) \text{ a.s}$$

that is $P(\{\omega \,:\, \lim_n X_n(\omega) = X(\omega)\}) = 1$.

Almost sure convergence is pointwise convergence, the limit is unique if of course we identify RV which are equal a.s.

It will be very useful to rephrase almost sure convergence in a different way. At first sight it looks a bit strange but it is a good idea. We explain the idea first for a sequence $\{x_n\}$ of numbers. Consider the function defined for any $\epsilon > 0$

$$i_\epsilon(x) = 1_{(\epsilon,\infty)}(x) = \begin{cases} 1 & x > \epsilon \\ 0 & x \le \epsilon \end{cases}$$

Lemma 1.1

- A sequence $\{x_n\}$ converges to x if and only if $\sum_n i_\epsilon(|x_n - x|) < \infty$ for every ϵ.
- If there is a non-negative sequence ϵ_n such that $\sum_n \epsilon_n < \infty$ and $\sum_n i_\epsilon(|x_n - x_{n+1}|) < \infty$ then x_n converges to some x.

Proof.

- Fix $\epsilon > 0$. If $x_n \to x$ then there exists N such that for any $n \geq N$, $|x_n - x| \leq \epsilon$ which means $i_\epsilon(|x_n - x|) = 0$ for $n \geq N$ and thus $\sum_n i_\epsilon(|x_n - x|) < \infty$. Conversely if $\sum_n i_\epsilon(|x_n - x|) < \infty$ then, since the terms are either 0 or 1, only finitely many terms can be nonzero and thus there exists N such that $|x_n - x| \leq \epsilon$ for $n \geq N$. $\square$

- If this holds there exists N such that $|x_n - x_{n+1}| \leq \epsilon_n$ for $n \geq N$. Taking $n > m \geq N$ gives

$$|x_n - x_m| \leq |x_n - x_{n-1}| + \cdots + |x_{m+1} - x_m| \leq \epsilon_m + \cdots + \epsilon_n \leq \sum_{j=m}^{\infty} \epsilon_j$$

  Since the the sequence ϵ_n is summable $\sum_{j=m}^{\infty} \epsilon_j$ goes to 0 as $m \to \infty$. Therefore the sequence x_n is a Cauchy sequence and thus $x_n \to x$.

Returning to random variables we find

Theorem 1.1 The sequenceof RV X_n converges almost surely if and only if, for every $\epsilon > 0$,

$$\sum_n i_\epsilon \circ |X_n - X| < \infty \quad \text{almost surely} \tag{1.1}$$

Proof.

- Let Ω_0 the set on which convergence holds, then the sum in Equation 1.1 converges for $\omega \in \Omega_0$ (which is independent of ϵ).

- For the converse the only small issue to deal with is that the set on which the sum converges may depend on ϵ. So pick a sequence $\epsilon_k \searrow 0$ and let N_k the value of the sequence in Equation 1.1 and we have $P(N_k < \infty) = 1$. Since $\epsilon_{k+1} \le \epsilon_k, i_{\epsilon_{k+1}} \ge i_{\epsilon_k}$ and so $N_{k+1} \ge N_k$. The events $\{N_k < \infty\}$ are shrinking to

$$\Omega_0 = \{\omega \,:\, \sum_n i_\epsilon \circ |X_n - X| < \infty \text{ for all } \epsilon > 0\}$$

  By sequential continuity $P(\Omega_0) = \lim_k P(N_k < \infty) = 1$ and thus X_n converges to X almost surely. $\square$.

At this point we recall the Borel-Cantelli theorem

Lemma 1.2 (Borel Cantelli Lemma) Suppose B_n is a collection of events. Then

$$\sum_n P(B_n) < \infty \implies \sum_n 1_{B_n} < \infty \text{ almost surely}.$$

which we use to prove the following criteria for almost sure convergence

Theorem 1.2

1. If, for any $\epsilon > 0$, $\sum_n P(|X_n - X| \geq \epsilon) < \infty$ then $X_n \to X$ almost surely.
2. If there exists a sequence $\epsilon_n \searrow 0$ such that $\sum_n P(|X_n - X| \geq \epsilon_n) < \infty$, then $X_n \to X$ almost surely.
3. If there exists a sequence ϵ_n with $\sum_n \epsilon_n < \infty$ and $\sum_n P(|X_n - X_{n+1}| \geq \epsilon_n) < \infty$, then $X_n \to X$ almost surely.

Proof.

1. By Borel-Cantelli we have $\sum_n i_\epsilon(|X_n - X|) < \infty$ a.s. which means a.s convergence.
2. By the Borel-Cantelli Lemma we have $|X_n - X| \leq \epsilon_n$ for all but finitely many n, almost surely. Since $\epsilon_n \to 0$ this means that X_n converges to X a.s
3. By Lemma 1.1 and Borel-Cantelli $\{X_n\}$ is a Cauchy sequence a.s. and thus converges a.s.

1.2 Convergence in Probability

Definition 1.2 (Convergence in probability) A sequence of RVs $\{X_n\}$ *converges in probability* to a RV X if, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}) = 0 .$$

This is a very useful mode of convergence in probability, in particular due to the fact that, that it is weaker than almost sure convergence and thus easier to prove.

Example: Let $\Omega = [0,1)$ and P_0 be Lebesgue measure. Consider the sequence of RV

$$X_1 = 1_{[0,\frac{1}{2})}, X_2 = 1_{[\frac{1}{2},1)}, X_3 = 1_{[0,\frac{1}{3})}, X_4 = 1_{[\frac{1}{3},\frac{2}{3})}, X_5 = 1_{[\frac{2}{3},1)},, X_6 = 1_{[0,\frac{1}{4})}, X_7 = 1_{[1/4,\frac{2}{4})} \cdots \quad (1.2)$$

- We claim that X_n converges to 0 in probability. Indeed for any $\epsilon > 0$, $P(|X_n| > \epsilon) = P(X_n = 1) \to 0$ since the $X_n = 1_{I_n}$ is a charactersitic function of an interval I_n whose measure goes to 0 as n goes to infinity.
- The sequence X_n does not converge a.s. Indeed $\omega \in [0,1)$ belong to infinitely many intervals of the form $[\frac{k}{n}, \frac{k+1}{n})$ and does not belong to infinitely many such intervals. Therefore $\liminf X_n(\omega) = 0 \neq \limsup X_n(\omega) = 1$.
- The sequence $X_n(\omega)$ has (many!) convergent subsequence which converges to 0 almost surely. To do this choose n_k such that the interval I_{n_k} for $X_{n_k} = 1_{I_{n_k}}$ is contained in the interval $I_{n_{k-1}}$ for $X_{n_{k-1}} = 1_{I_{n_{k-1}}}$.

The relation between almost sure convergence and convergence in probability is contained in the following theorem. The third part, while looking somewhat convoluted is cvery useful to streamline subsequent proofs. It relies on the following simple fact: suppose that the sequence x_n is such that every subsequence has a subsubsequence which converges to x then x_n converges to x.

Theorem 1.3 (Almost sure convergence versus convergence in probability)

1. If X_n converges almost surely to X then X_n converges in probability to X.
2. If X_n converges in probability to X then there exists a subsequence X_{n_k} which converges to X almost surely,
3. If every subsequence has a further subsubsequence which converges to X almost surely, then X_n converges to X in probability.

Proof. Item 1.: If X_n converges to X almost surely then $i_\epsilon(|X_n - X|)$ converges to 0 almost surely. By the bounded convergence theorem this implies $E[i_\epsilon(|X_n - X|)] = P(|X_n - X| \geq \epsilon)$ converges to 0.

Item 2.: If X_n converges to X in probability then pick $\epsilon_k = \frac{1}{k} \searrow 0$. Since $P(|X_n(\omega) - X(\omega)| > \epsilon_k) \to 0$ as $n \to \infty$ we can find a subsequence n_k such that $P(|X_{n_k}(\omega) - X(\omega)| > \epsilon_k) \leq \frac{1}{2^k}$. and thus

$$\sum_{k=1}^{\infty} P(|X_{n_k}(\omega) - X(\omega)| > \epsilon_k) \leq \sum_k \frac{1}{2^k} < \infty\,.$$

By part 2. of Theorem 1.2 X_{n_k} converges almost surely to X.

Item 3.: Assume that every subsequence of X_n has a sub-subsequence which converges to X. Fix $\epsilon > 0$ and consider the numerical sequence $p_n(\epsilon) = P(|X_n - X| \ge \epsilon)$. Since this sequence is bounded, by Bolzano-Weierstrass theorem, let p_{n_k} be a convergent subsequence with $\lim_k p_{n_k} = p$. By assumption X_{n_k} has a convergent subsequence $X_{n_{k_j}}$ which converges to X almost surely. This implies, by part 1., that $p_{n_{k_j}}$ converges to 0. This means that for the sequence p_n, every convergent subsequence has a subsubsequence which converges to 0. This implies that p_n converges to 0 and thus X_n converges to X in probability. $\square$.

Based on this we obtain the following continuity theorem

Theorem 1.4 (Continuity theorem for convergence in probability)

1. If X_n converges to X almost surely and f is continous function then $f(X_n)$ converge to $f(X)$ almost surely.
2. If X_n converges to X in probability and f is continous function then $f(X_n)$ converge to $f(X)$ in probability.

Proof. Part 1. is just the definition of continuity.
For part 2. suppose X_n converges to X in probability. Then, by Theorem 1.3, there exists a subsequence X_{n_k} which converges almost surely which implies, by part 1, that $f(X_{n_k})$ converges to $f(X)$ almost surely.

Now we apply part 3. of Theorem 1.3 to the sequence $Y_n = f(X_n)$. Since X_{n_k} converges in probability, by the previous paragraph, every subsequence $f(X_{n_k})$ has a convergent subsubsequence whic converges to $f(X)$ a.s. and thus $f(X_n)$ converges to $f(X)$ in probability.

Using a similar argument we show that convergence in probability is preserved under arithmetic operations.

Theorem 1.5 Suppose X_n converges to X in probability and Y_n converges to Y in probability then $X_n + Y_n$ converges to $X + Y$ in probability, $X_n - Y_n$ converges to $X - Y$ in probability, $X_n Y_n$ converges to XY in probability and $X - n/Y_n$ converges to X/Y in probability (assuming that Y_n and Y are almost surely non-zero).

Proof. All the proofs are the same so let us do the sum. We pick a subsequence such that X_n converges to X almost surely along that subseqence. Then we pick a subsubsequence such that both X_n and Y_n converges almost surely along that subsequence. For that subsequence $X_n + Y_n$ converges to $X + Y$ almsost surely. We now apply this argument to subsequence of $X_n + Y_n$, every such subsequence as a subsubsequence which converges to $X + Y$ almost surely and thus by part 3. of Theorem 1.3 $X_n + Y_n$ converges to $X + Y$ almost surely.

We finish this section by showing that we can use weak convergence to turn the space of RV into a complete metric space. We will use the following metric

$$d(X,Y) = E[\min\{|X - Y|, 1\}]$$

The choice is not unique, often one will find instead $d(X,Y) = E\left[\frac{|X-Y|}{1+|X-Y|}\right]$.

Theorem 1.6 (Convergence in probability and metric)

- X_n to converge to X in probability if and only if $\lim_{n\to\infty} d(X_n, X) = 0$.
- The space of all measurable RV,

$$L^0(\Omega, \mathcal{A}, P) = \{X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B}) \text{ measurable}\}$$

with the metric d is a complete metric space for the metric $d(X, Y)$ (as usual we identify RV which are a.s. equal). Equivalently, for any Cauchy sequence $\{X_n\}$ for convergence in probability there exists a random variable Y such that X_n converges to Y in probbaility.

Proof. It is easy to check that $d(X, Y)$ is a distance.

We also have for $\epsilon \in (0, 1)$ and $x \geq 0$ the inequality

$$\epsilon i_\epsilon(x) \leq \min\{x, 1\} \leq \epsilon + i_\epsilon(x)$$

Replacing x by $|X_n - X|$ and taking expectations and $n \to \infty$ shows that

$$\epsilon P(|X_n - X| \geq \epsilon) \leq d(X_n, X) \leq \epsilon + P(|X_n - X| \geq \epsilon)$$

and this proves the second claim.

Finally assume that $\{X_n\}$ is a Cauchy sequence for the metric d. Then $\lim_{n,m\to\infty} d(X_n, X_m) = 0$ or, as we have just seen, equivalently for any $\epsilon > 0$ we can find N such that $P(|X_n - X_m| \geq \epsilon) \leq \epsilon$ for $n, m \geq N$. Choose now $\epsilon_k = \frac{1}{2^k}$ and find corresponding $N_k \leq N_{k+1} \leq \cdots$. Setting $Y_k = X_{N_k}$, this implies that $\sum_k P(|Y_k - Y_{k+1}| > \epsilon_k) < \infty$ and thus Y_k converges almost surely to a RV Y.

To conclude we show that X_n converges to Y in probability. Since $|X_n - Y| \leq |X_n - X_{N_k}| + |Y_k - Y|$ we have

$$i_\epsilon(|X_n - Y|) \leq i_{\frac{\epsilon}{2}}(|X_n - X_{N_k}|) + i_{\frac{\epsilon}{2}}(|Y_k - Y|)\,.$$

Taking expectations gives

$$P(|X_n - Y| > \epsilon) \leq P(|X_n - X_{N_k}| > \epsilon/2) + P(|Y_k - Y| > \epsilon/2).$$

The first goes to 0 as n goes to ∞ since the sequence X_n is Cauchy and the second term goes to 0 since we have almost sure convergence. Therefore X_n converges to Y in probability. $\square$

1.3 Convergence in L^p.

Convergence in L^p simply uses the norm $\|X\|_p$. Most of the time we use $p = 1$ or $p = 2$.

Definition 1.3 (Convergence in L^p) A sequence of RVs $\{X_n\}$ *converges in L^p to a RV* X if

$$\lim_{n\to\infty} E[|X_n - X|^p] = 0$$

or equivalently $\lim_{n\to\infty} \|X_n - X\|_p = 0$.

Remarks:

- The limit of a sequence in L^p is unique since $\|X - Y\|_p \le \|X - X_n\|_p + \|Y - X_n\|_p$ by Minkovski inequality.
- Note that if X_n converges to X in L^1 then we have convergence of the first moments. Indeed we have

  $$|E[X_n] - E[X]| \le E[|X_n - X|] \quad \text{and} \quad |E[|X_n|] - E[|X|]| \le E[|X_n - X|]$$

  (the second follows the reverse triangle inequality $||x| - |y|| \le |x - y|$). Therefore $E[X_n] \to E[X]$ and $E[|X_n|] \to E[|X|]$.
- If X_n converges in L^1 then X_n does not need to converge almost surely. See the sequence Equation 1.2 which also converges to 0 in L^1.

- If X_n converges in L^p then X_n converges in probability as well. We have $P(|X - X_n| \geq \epsilon) \leq E[|X_n - X|^p]/\epsilon^p$ by Markov inequality. In particular, by Theorem 1.3 if X_n converges to X in L^p then there exists a subsequence X_{n_k} which converges almost surely to X.

- Conversely convergence in probability does not imply convergence in L^1. Modify the sequence in Equation 1.2 to make it $Y_1 = X_1, Y_2 = 2X_2, Y_3 = 2X_3, Y_4 = 3X_4, \cdots$. This sequence converges in probability to 0 as well! This ensures that $E[Y_n] = 1$ so Y_n does not converge to 0 in L^1. Note also that for any m there are infitely many $m \geq n$ such that $E[|X_n - X_m| = 2$ so the sequence cannot converge.

We prove now a converse which looks a bit like dominated convergence theorem.

Theorem 1.7 (Convergence in L^p versus convergence in probability)

1. If X_n converges to X in L^p then X_n converges to X in probability.
2. If X_n converges to X in probability and $|X_n| \leq Y$ for some $Y \in L^P$ then X_n converges to X in L^p.

Proof. We already discussed 1. (Markov inequality).

For the converse, since X_n converges in probability to X there exists a subseqeunce X_{n_k} which converges almost surely to X.
Since $|X_{n_k}| \le Y$ we see that $|X| < Y$ and thus $X \in L^p$.

The sequence $a_n = E[|X - X_n|^p]$, is bounded since, by Minkowski

$$a_n^{1/p} = \|X - X_n\|_p \le \|X\|_p + \|X_n\|_p \le 2\|Y\|_p \,.$$

Let a_{n_k} be a convergent subsequence. Then since X_{n_k} converges to X in probability there exists a subsubsequence $X_{n_{k_j}}$ which converge to X a.s Then $|X_{n_{k_j}} - X|^p$ converges almost surely to 0 and $|X_{n_{k_j}} - X|^P \le 2^p |Y|^p$ which is integrable. So by DCT $a_{n_{k_j}}$ converges to 0. This implies that a_n converges to 0. $\square$.

1.4 Exercises

Exercise 1.1 Show (by a counterxample) that if f is not continuous, convergence of X_n in probability to X does not imply convergence of $f(X_n)$ to $f(X)$ in probability.

Exercise 1.2 Let $X_1, X_2, \ldots$ be independent Bernoulli random variables with $P(X_n = 1) = p_n$ and $P(X_n = 0) = 1 - p_n$.

- Show that X_n converges to 0 inprobability if and only if $\lim_n p_n = 0$.
- Show that X_n converges to a.s. if and only if $\sum_n p_n < \infty$.

Hint: Use the Borel Cantelli Lemmas

Exercise 1.3

- Suppose X_n converges to X in L^1. Show that $\lim_n E[X_n] = E[X]$.
- Suppose X_n converges to X in L^2. Show that $\lim_n E[X_n^2] = E[X^2]$.

2 The law of large numbers

The law of large numbers is a foundational (and also very intuitive) concept for probability theory. Suppose we are interested in finding the the probability $P(A)$ for some event A which is the outcome of some (random experiment). To do this repeat the experiment n times, for sufficiently large n and an approximate value for $P(A)$ is the proportion of experiments for which the outocme belong to A

$$P(A) \approx \frac{\text{number of times the experiment belongs to } A}{n} \quad \text{if } n \text{ is large enough}$$

This is the basis for the *frequentist approach to probability*, probability of events are obtained by repeating the random experiment.

2.1 Strong of law of large numbers

Consider a probability space $(\Omega, \mathcal{A}, P)$ on which real-valued random variables $X_1, X_2, \cdots$ are defined. We define then the sum

$$S_n = X_1 + \cdots + X_n$$

Law of large numbers stands for the convergence of the average $\frac{S_n}{n}$ to a limit. The convergence can be in probability in which case we talk of a weak law of large numbers or almost sure convergence in which case we talk about a strong law of large numbers.

Example: Normal Consider $X_1, X_2, \cdots, X_n$ independent normal RV with mean 0 and variance σ^2. Then using the moment generating function we have

$$E\left[e^{t\frac{S_n}{n}}\right] = E\left[e^{\frac{t}{n}X_1} \cdots e^{\frac{t}{n}X_1}\right] = E\left[e^{\frac{t}{n}X_1}\right] \cdots E\left[e^{\frac{t}{n}X_1}\right] = \left(e^{\frac{\sigma^2}{2}\frac{t^2}{n^2}}\right)^n = e^{\frac{\sigma^2}{2}\frac{t^2}{n}}$$

We conclude that $\frac{S_n}{n}$ is a normal random variable with variance $\frac{\sigma^2}{n}$.

By Chebyshev we conclude that $P\left(\left|\frac{S_n}{n}\right| \ge \epsilon)\right) \le \frac{\sigma^2}{n\epsilon}$ and so $\frac{S_n}{n}$ converges to 0 in probability. This is not enough to show almost sure convergence since $\sum_n \frac{\sigma^2}{n\epsilon} = \infty$.

By the Chernov bounds however we have (see the example after **?@thm-chernov**) $P\left(\left|\frac{S_n}{n}\right| \ge \epsilon\right) \le e^{-n\epsilon^2/2\sigma^2}$ and since $\sum_n e^{-nt^2/2\sigma^2} < \infty$ Borel-Cantelli Lemma (see Theorem 1.2) implies that $\frac{S_n}{n}$ converges to 0 a.s.

Example: Cauchy Consider $X_1, X_2, \cdots, X_n$ independent Cauchy RV with parameter β. Using the characteristic function (recall the chararcteristic function of a Cauchy RV is $e^{-\beta|t|}$) we find that

$$E\left[e^{it\frac{S_n}{n}}\right] = E\left[e^{\frac{it}{n}X_1} \cdots e^{\frac{it}{n}X_1}\right] = E\left[e^{\frac{it}{n}X_1}\right] \cdots E\left[e^{\frac{it}{n}X_1}\right] = \left(e^{-\beta\left|\frac{t}{n}\right|}\right)^n = e^{-\beta|t|}$$

that is $S_n n$ is a Cauchy RV with same parameter as X_i. No convergence almost sure or in probability seems reasonable here convergence in distribution will be useful here.

We start with a (not optimal) version of the LLN

Theorem 2.1 (The strong law of large numbers) Suppose $X_1, X_2, \cdots$ are independent and identically distributed random variables defined on the probability space $(\Omega, \mathcal{A}, P)$ and with mean $\mu = E[X_i]$ and variance $\sigma^2 = \mathrm{Var(X_i)} < \infty$. Then we have

$$\lim_{n\to\infty} \frac{S_n}{n} = \lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu \quad \begin{cases} \text{almost surely} \\ \text{in probability} \\ \text{ in } L^2 \end{cases}$$

Proof. The proof is literally the same as what we did for discrete random variables and we shall not repeat it here.

A stronger version exists, only existence of a finite mean is needed.

Theorem 2.2 (The strong law of large numbers) Suppose $X_1, X_2, \cdots$ are independent and identically distributed random variables defined on the probability space $(\Omega, \mathcal{A}, P)$ and with mean $\mu = E[X_i]$. Then we have

$$\lim_{n\to\infty} \frac{S_n}{n} = \lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu \quad \begin{cases} \text{almost surely} \\ \text{in probability} \end{cases}$$

Various proofs of this exist. For example we can use a truncation argument of the random variables and argument similar to the previoious theorem workking wiht subsequences. Another more fancy proof use the Martingale convergence theorem.

We prove next that the case $\mu = \infty$ can also be treated.

Theorem 2.3 Suppose $X_1, X_2, \cdots$ are independent indetically distributed non-negative random variables with $E[X_i] = +\infty$. Then we have

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = +\infty \quad \text{almost surely}$$

Proof. We use a truncation argument combined with the monotone convergence theorem. Given $R > 0$ set $Y_n = \min\{X_n, R\}$ which is bounded and thus has finite variance. So by Theorem 2.2 we have, almost surely, for $\mu_R = E[\min\{X_1, R\}]$

$$\lim_{n\to\infty} \frac{Y_1 + \cdots + Y_n}{n} = \mu_R$$

Since $X_n \ge Y_n$ we have for any R

$$\liminf_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} \ge \lim_{n\to\infty} \frac{Y_1 + \cdots + Y_n}{n} = \mu_R$$

But as $R \nearrow \infty$ $\min\{X_1, R\} \nearrow Y_1$ and thus by the monotone convergent theorem $\mu_R = E[\min\{X_1, R\}] \nearrow E[X_1] = \infty$. This concludes the proof. $\square$.

2.2 Sample variance

Example: convergence of the sample variance Suppose $X_1, X_2, \cdots$ are independent and identically distributed RV, the the sample variance is given by

$$V_n = \sum_{i=1}^{n} \left(X_i - \frac{S_n}{n} \right)^2$$

After some calculation one can prove that $E[V_n] = (n-1)\sigma^2$.

We claim that $\frac{V_n}{n} \to \sigma^2$ almost surely. Indeed we have

$$\frac{V_n}{n} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \frac{S_n}{n} \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i^2 - 2X_i \frac{S_n}{n} + \frac{S_n^2}{n^2} \right) = \sum_{i=1}^{n} X_i^2 - \left(\frac{S_n}{n} \right)^2$$

By the law of Large numbers, Theorem 2.2, which we apply to the RV $X_1, X_2, \cdots$ and the RV $X_1^2, X_2^2, \cdots$ (note that $E[X_1^2] = \sigma^2 + \mu^2$) and by continuity, Theorem 1.4, we have

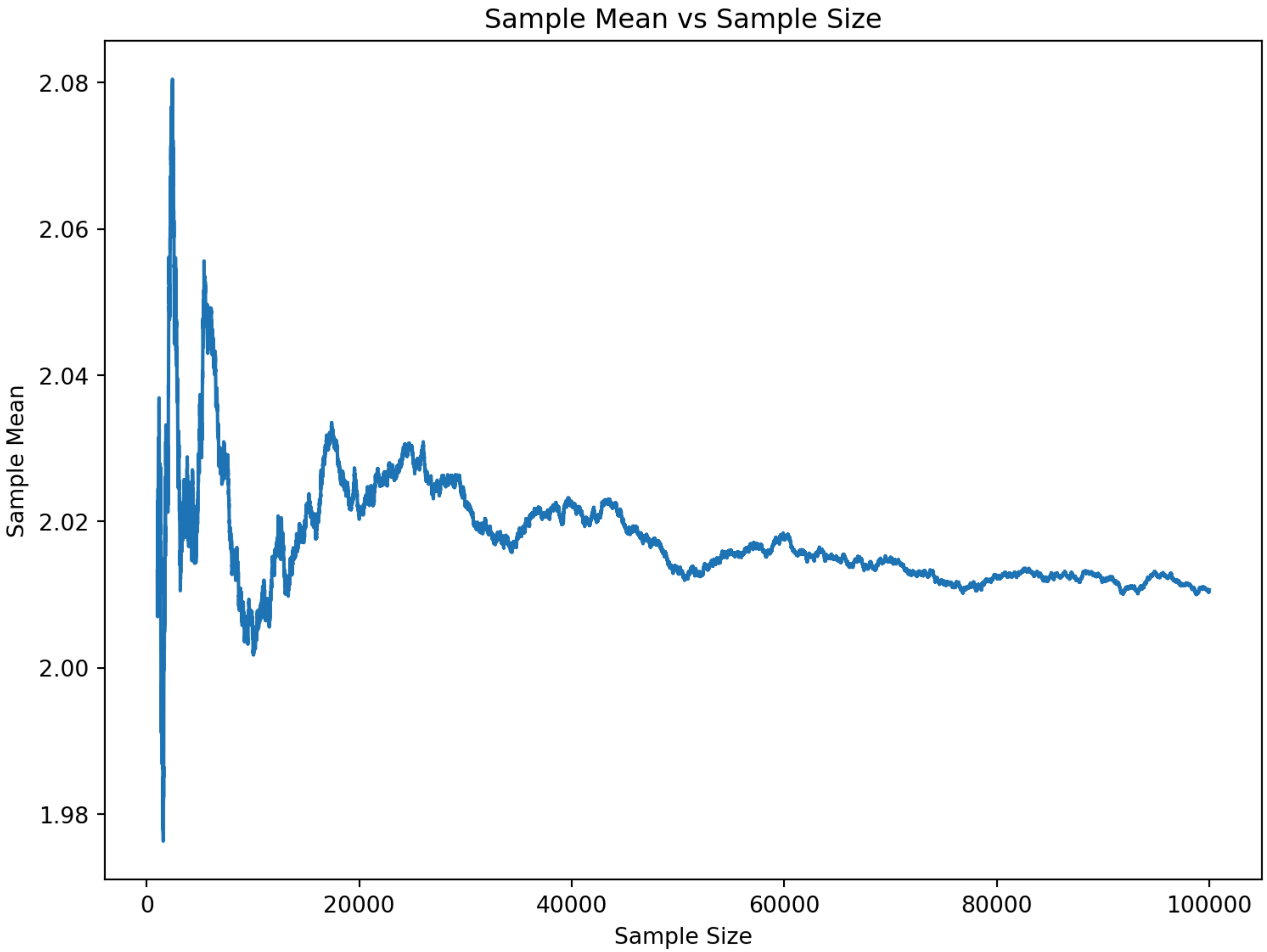
$$\frac{V_n}{n} \to \sigma^2 + \mu^2 - \mu^2 = \sigma^2 \, .$$

▼ Code

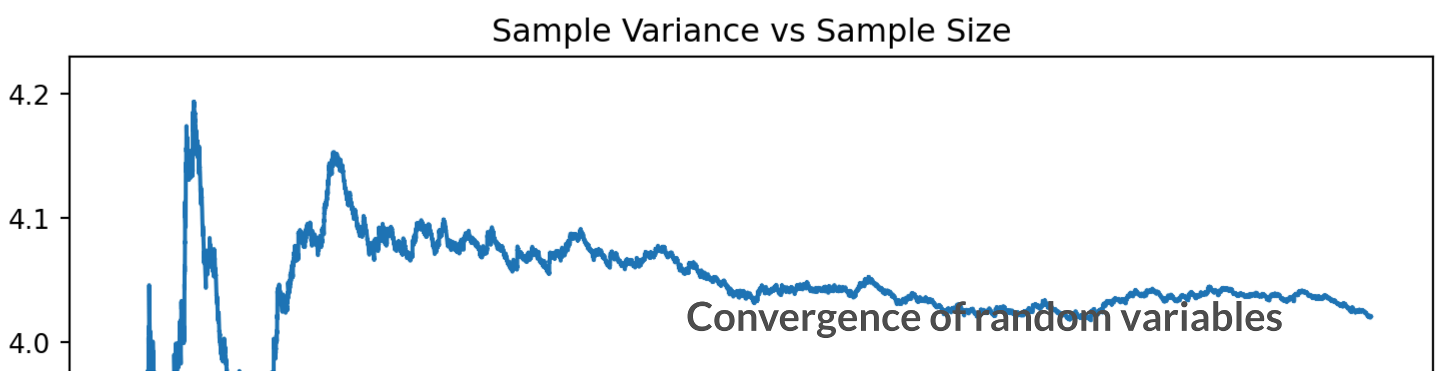
```
import random
import matplotlib.pyplot as plt

# Parameters for the exponential distribution
lambda_parameter = 0.5  # Adjust this to your desired rate parameter

# Initialize variables to track sample mean and sample variance
sample_mean = 0
sample_variance = 0
sample_size = 0

# Number of samples to collect
num_samples = 100000

# Lists to store data for plotting
sample_means = []
sample_variances = []

for _ in range(num_samples):
    # Generate a random sample from the exponential distribution
    sample = random.expovariate(lambda_parameter)

    # Update the sample size
    sample_size += 1

    # Update the sample mean incrementally
```

```
Last 20 sample means =[2.010433725990389, 2.010436103379991, 2.0104400629814756, 2.010425565861256,
2.0104784945739826, 2.010464237002987, 2.010509954990142, 2.0105230324476455, 2.010548268282596,
2.010551752165103, 2.0106193824308938, 2.010604223621517, 2.0106150659410043, 2.0106101808298944,
2.0106085093298613, 2.0105962031333076, 2.010581152622161, 2.0106375647861308, 2.0106260239362452,
2.0106405997359498]
Last 20 sample variances=[4.020619418686085, 4.020579769941937, 4.020541124468071, 4.02052192529738,
4.020761813684171, 4.020741924817882, 4.020910695941878, 4.020887581373271, 4.020911044766209,
```

```
4.0208720448424 93, 4.021289171647335, 4.021271932018198, 4.021243470730342, 4.021205641744572,
4.021165706649759, 4.021140636472159, 4.021123074339779, 4.021401085580041, 4.021374189620114,
4.021355220657325]
```



Sample Mean vs Sample Size
Sample Mean
Sample Size
2.08
2.06
2.04
2.02
2.00
1.98
0
20000
40000
60000
80000
100000



Sample Variance vs Sample Size
4.2
4.1
4.0

2.3 LLN proof of the Weierstrass approximation theorem

A classical result in analysis is that a continuous function $f : [a, b] \to \mathbb{R}$ can be uniformlu approximated by polynomial: for any $\epsilon > 0$ there exists a polynomila $p(x)$ such that $\sup_{x\in[a,b]} |p(x) - f(x)| \le \epsilon$. Many proof of this result exists and we give one here based on the Law of Large Numbers although the statement has nothing to do with probability. Without loss of generality, by rescaling, we can take $[a, b] = [0, 1]$ and we use polynomial naturally associated to binomial random variables, the Bernstein polynomials.

Theorem 2.4 Let $f : [0, 1] \to \mathbb{R}$ be a continuous function. Let $f_n(x)$ be the **Bernstein polynomial of degree** n associated to f, given by

$$f_n(x) = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} f(k/n)\,.$$

Then we have

$$\lim_{n\to\infty} \sup_{x\in[0,1]} |f(x) - f_n(x)| = 0\,.$$

Proof. If X_i are IID Bernoulli with success probability p random then $S_n = X_1 + \cdots + X_n$ is binomial random RV and

$$E\left[f\left(\frac{S_n}{n}\right)\right] = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} f(k/n) = f_n(p)$$

Since $\frac{S_n}{n}$ converges p a.s and in probability we have $f\left(\frac{S_n}{n}\right)$ converges to $f(p)$ and thus taking expectation $f_n(p)$ converges to $f(p)$. We still do need to work harder, though, to establish uniform convergence, in p.

The variance of a binomial is $np(1-p) \le \frac{n}{4}$ is bounded uniformly in $p \in [0,1]$ and thus by Chebyshev

$$P\left(\left|\frac{S_n}{n} - p\right| \ge \delta\right) \le \frac{p(1-p)}{n\delta^2} \le \frac{1}{4n\delta^2}$$

Since f is continuous on a compact interval then f is bounded with $\sup_x |f(x)| = M < \infty$ and f is also uniformly continuous on $[0,1]$. Given $\epsilon > 0$ pick δ such $|x-y| < \delta \implies |f(x) - f(y)| < \epsilon$. We have, for n large enough,

$$\begin{aligned}
|f_n(p) - f(p)| &= \left|E\left[f\left(\frac{S_n}{n}\right)\right] - f(p)\right| \le E\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right|\right] \\
&= E\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right| 1_{\left|\frac{S_n}{n} - p\right| \ge \delta}\right] + E\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right| 1_{\left|\frac{S_n}{n} - p\right| < \delta}\right] \le 2M\frac{1}{4n\delta^2} + \epsilon \le 2\epsilon
\end{aligned}$$

This proves uniform convergence. $\square$.

2.4 Empirical density and Glivenko-Cantelli

Example: sample CDF and empirical measure Suppose $X_1, X_2, \cdots$ are independent and identically distributed RV with common CDF $F(t)$

$$F_n(t) = \frac{\#\{i \in \{1, \cdots, n\} : X_i \le t\}}{n} \to F(t) \quad \text{almost surely}\,. \tag{2.1}$$

$F_n(t) = F_n(t, \omega)$ is called the empirical CDF. Note that $F_n(t, \omega)$ is the (random) CDF for the discrete random variable with distribution

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)}$$

which is called the empirical distribution. The convergence of $F_n(t)$ to $F(t)$ is is just the law of large number applied to $Y_n = 1_{X_n \le t}$ whose mean is $E[Y_n] = P(Y_n \le t) = F(t)$.

A strengthening of the law of large number is that the empirical CDF $F_n(t)$ converges to $F(t)$ uniformly in t.

Theorem 2.5 (Glivenko-Cantelli Theorem) For a RV X with CDF $F(t)$ we have

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \text{ converges to } 0 \quad \text{almost surely.}$$

Proof. We only the prove the case where $F(t)$ is continuous but the proof can be generalized to general CDF by considering the jumps more carefully. The proof relies on the fact that F is increasing which precludes oscillations and control the convergence.

First we show that we can pick a set of probability 1 such that the convergence occurs for all $t \in \mathbb{R}$ on that set. Since a countable union of sets of probability 0 has probability 0, we can pick a set Ω_0 of probability 1 such $F_n(t,\omega)$ converges to $F(t)$ for all *rational $t \in \mathbb{R}$ and all $\omega \in \Omega_0$. For $x \in \mathbb{R}$ and rational s, t with $s \le x \le t$ we have

$$F_n(s,\omega) \le F_n(x,\omega) \le F_n(t,\omega)$$

and therefore

$$F(s) \le \liminf_n F_n(x,\omega) \le \limsup_n F_n(x,\omega) \le F(t)$$

Since $F(t) \searrow F(x)$ as $t \searrow x$ and $F(s) \nearrow F(x)$ as $s \nearrow x$ we conclude that $F_n(t,\omega) \to F(t)$ for all $\omega \in \Omega_0$.

We show next that, for any $\omega \in \Omega_0$, the convergence is uniform in t. Since F is increasing and bounded, given $\epsilon > 0$ we can find $t_0 = -\infty < t_1 < \cdots < t_m = +\infty$ such that $F(t_j) - F(t_{j-1}) \le \frac{\epsilon}{2}$. Using that F_n and F are increasing we have for $t \in [t_{j-1}, t_j]$

$$F_n(t) - F(t) \le F_n(t_j) - F(t_{j-1}) \le F_n(t_j) - F(t_j) + \frac{\epsilon}{2}$$
$$F_n(t) - F(t) \ge F_n(t_{j-1}) - F(t_j) \ge F_n(t_{j-1}) - F(t_{j-1}) - \frac{\epsilon}{2}$$

We can now pick $N_j = N_j(\omega)$ such that $|F_n(t_j) - F(t_j)| \le \frac{\epsilon}{2}$ if $n \ge N_j$ and therefore if $n \ge N = \max_j N_j$ we have, for all $t \in \mathbb{R}$,

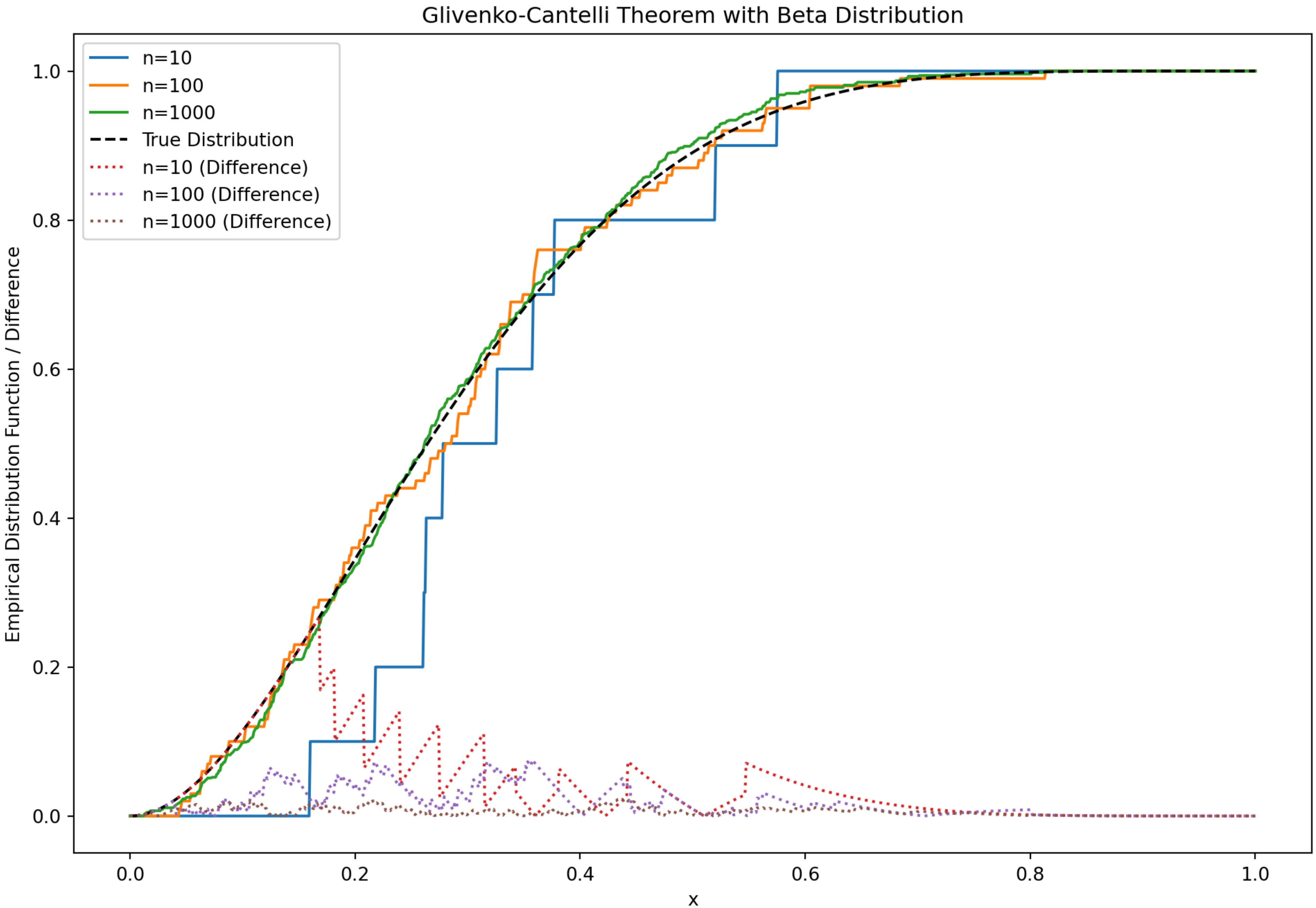
$$|F_n(t, \omega) - F(t)| \le \epsilon$$

for all t and all $n \ge N(\omega)$. This that $\sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$ converges almost surely to 0. $\square$.

Illustration of the Glivenko-Cantelli Theorem (made with Chat GPT)

▼ Code

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import beta

# Generate random data from a Beta distribution
np.random.seed(42)
true_distribution = beta.rvs(2, 5, size=1000)

# Generate empirical distribution function
def empirical_distribution(data, x):
    return np.sum(data <= x) / len(data)

# Compute empirical distribution function for different sample sizes
sample_sizes = [10, 100, 1000]
x_values = np.linspace(0, 1, 1000)

plt.figure(figsize=(12, 8))

for n in sample_sizes:
    # Generate a random sample of size n
    sample = np.random.choice(true_distribution, size=n, replace=True)

    # Calculate empirical distribution function values
    edf_values = [empirical_distribution(sample, x) for x in x_values]

    # Plot the empirical distribution function
```



Convergence of random variables

2.5 The Monte-Carlo method

The **(simple) Monte-Carlo method** is a probabilistic algorithm using sums of independent random variables the law of large numbers to estimate a (deterministic) quantity $\mu \in \mathbb{R}$ (or $\mathbb{R}^d$).
The basic idea is to express μ as the expectation of some random variable $\mu = E[h(X)]$ and then use the law of large numbers to build up an estimator for μ.

Simple Monte-Carlo Sampling Algorithm: To compute $\mu \in \mathbb{R}$

- Find a random variable $h(X)$ such that $\mu = E[h(X)]$.
- $I_n = \frac{1}{n}\sum_{k=1}^n h(X_k)$, where X_k are IID copies of X, is an **unbiased estimator** for μ, that is we have
  - For all n we have $E[I_n] = \mu$ (unbiased).
  - $\lim_{n\to\infty} I_n = \mu$ almost surely and in probability.

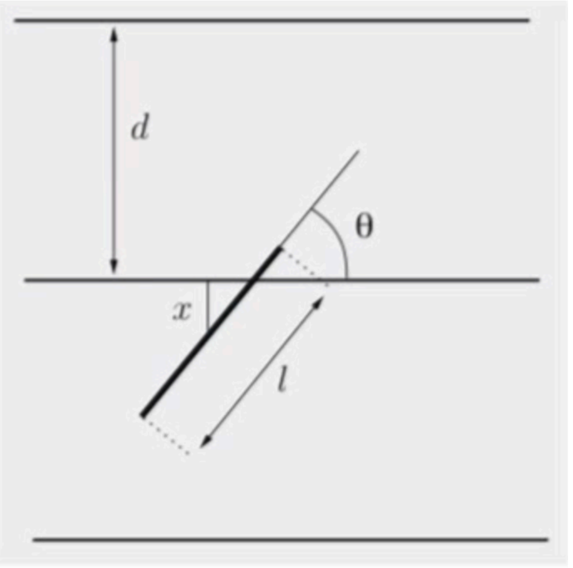
- An interesting part is that there are, in general, many ways to find the random variables $h(X)$.
- Conversely in many problems the random variable $h(X)$ is given but the expection is too diffcult to compute, so we rely on the LLN to compute μ.

2.6 Computing π with Buffon's needles

- This seems to be the first example of a rejection sampling used to solve a mathematical problem, by Le Comte de Buffon (see Bio in Wikipedia).
- A needle of length l is thrown at random on floor made on floorboards of width d and we assume $l \le d$. We want to compute the probability that the needle does intersect two floor boards.
- Denote by X the distance from the center of the needle to the nearest intersection (this is uniformly distributed on $[0, \frac{d}{2}]$) and by Θ the acute angle between the needle and an horizontal line (this is uniformly distributed on $[0, \frac{\pi}{2}]$).
- For the needle to intersect we must have $x \le \frac{l}{2}\sin(\theta)$ and thus

$$P\left(X \le \frac{l}{2}\sin(\Theta)\right) = \int_0^{\frac{\pi}{2}} \int_0^{\frac{l}{2}\sin(\theta)} \frac{2}{d}dx\,\frac{2}{\pi}d\theta = \frac{2l}{d\pi}$$

- So in order estimate π you shall throw n needles on the floors at random and $\pi \approx \frac{2ln}{d}\frac{1}{\#\text{ of needles intersecting two floor boards}}$. No random number generator needed....



Buffon's needles

2.7 Computing π with random numbers

- Enclose a disk of radius 1 in a square of side length 2 and consider the following Bernoulli random variable X.
  - Generate 2 independent vectors V_1, V_2 uniformly distributed on $[-1, 1]$.
  - If $V_1^2 + V_2^2 \leq 1$, set $X = 1$, otherwise set $X = 0$.

2.8 Computing integrals

The goal is compute for example $\int_a^b h(x)dx$. Without loss of generality by rescaling space and replacing f by $cf + d$ we can assume that $[a,b] = [0,1]$ and $0 \le h \le 1$. If $h(x) = \sqrt{1-x^2}$ we recover the previous example.

Monte-Carlo version I: Pick independent random numbers (U_1, U_2) on $[0,1] \times [0,1]$. Define a Bernoulli RV X by

$$X = \begin{cases} 1 & \text{if } U_2 \le h(U_1) \\ 0 & \text{else} \end{cases}$$

Then

$$E[X] = P(U_2 \le h(U_1)) = \int_0^1 \int_0^1 1_{\{x_2 \le h(x_1)\}} dx_2 dx_1 = \int_0^1 \int_0^{h(x_1)} dx_2 = \int_0^1 h(x)dx$$

and so for independent X_i, $\frac{1}{n}\sum_{i=1}^n X_i \to \int_0^1 h(x)dx$ almost surely.

Monte-Carlo version II: Pick U uniform on $[0,1]$ then

$$E[h(U)] = \int_0^1 h(x)dx$$

and so for independent U_i, $\frac{1}{n}\sum_{i=1}^n f(U_i) \to \int_0^1 h(x)dx$ almost surely.

- Monte-Carlo version III:

  Pick V non-uniform on $[0,1]$ with density f (e.g a beta RV with parameter α, β). Then we have

$$\int_0^1 h(x)dx = \int_0^1 \frac{h(x)}{f(x)} f(x)dx$$

  and so if V has distribution f then $\frac{1}{n}\sum_{i=1}^n \frac{h(V_i)}{f(V_i)}$ converges to $E[\frac{h(v)}{f(V)}] = \int_0^1 h(x)dx$.

  This is the idea behind importance sampling: you want to sample more points from regions where h is large and which contribute more ot the integral. We will discuss this in a bit more detail when equipped with the central limit theorem.

- We can generalize this to integral of subsets of $\mathbb{R}^n$ or integral over the whole space.

2.9 Quantitative version of the law of large numbers

- How many sample should we generate to obtain a given precision for the computation of $\mu = E[X]$?
- In Monte-Carlo methods μ itself is unknown, so we would like to make a prediction about μ given the sample mean $\frac{S_n}{n} = \frac{X_1+\cdots+X_n}{n}$.
- Convergence in probability is the way to go: we want to estimate

$$P\left(\left|\frac{S_n}{n} - \mu\right| \le \epsilon\right) = P\left(\mu \in \left[\frac{S_n}{n} - \varepsilon, \frac{S_n}{n} + \varepsilon\right]\right) \ge 1 - \delta$$

which gives a confidence interval for μ in terms of the sample size n and the tolerance ϵ. For example of $\delta = 0.01$, if we have n sample, we can predict tolerance for μ with 99% confidence.

- If we know the variance we could use Chebyshev $P\left(\mu \in \left[\frac{S_n}{n} - \varepsilon, \frac{S_n}{n} + \varepsilon\right]\right) \ge 1 - \frac{\sigma^2}{n\epsilon^2}$. But it may require knowledge of the variance and could be overly pessimistic if we have many moments.
- Even better we could use Chernov bonds which gives exponentiall (in n) bounds

$$P\left(\frac{S_n}{n} - \mu \ge \epsilon\right) = P(S_n \ge n(\mu+\epsilon)) \le \inf_{t \ge 0} \frac{E\left[^{tS_n}\right]}{e^{n(\mu+\epsilon)}} = e^{-n \sup_{t\ge 0}\{t(\mu+\epsilon) - \ln M(t)\}}$$

and a similar bound for $P\left(\frac{S_n}{n} - \mu \le -\epsilon\right)$.

2.10 Hoeffding's bound

- Chernov bounds are very sharp but requires knowledge of the mgf $M_X(t)$.
- One of the main idea behind concentration inequalities: given X bound $M_X(t) \le M(t)$ by the mfg $M(t)$ of a random variable Y which you know explicitly. Mostly here we take Y a Gaussian but one can also uses other ones, Bernoulli, Poisson, Gamma, etc…
- The following elementary bound will be used repeatedly.

Lemma 2.1 (Hoeffding's bound) Suppose $a \le X \le b$ with probability 1. Then for any $\varepsilon > 0$

1. Bound on the variance $\mathrm{Var}(X) \le \dfrac{(b-a)^2}{4}$
2. Bound on the mgf $E\left[e^{tX}\right] \le e^{tE[X]} e^{\frac{t^2(b-a)^2}{8}}$

Proof. For the bound on the variance $a \le X \le b$ implies that $-\dfrac{a-b}{2} \le X - \dfrac{a+b}{2} \le \dfrac{a-b}{2}$ and therefore

$$\mathrm{Var}(X) = \mathrm{Var}\left(X - \frac{a+b}{2}\right) \le E\left[\left(X - \frac{a+b}{2}\right)^2\right] \le \frac{(b-a)^2}{4}\,.$$

Since X is bounded the moment generating function $M(t) = \frac{e^{tX}}{E[e^{tX}]}$ exists for any $t \in \mathbb{R}$. To bound the $M(t)$ let us consider instead its logarithmx $u(t) = \ln M(t)$. We have

$$
\begin{aligned}
u'(t) &= \frac{M'(t)}{M(t)} && = E\left[X \frac{e^{tX}}{E[e^{tX}]}\right] \\
u''(t) &= \frac{M''(t)}{M(t)} - \left(\frac{M'(t)}{M(t)}\right)^2 && = E\left[X^2 \frac{e^{tX}}{E[e^{tX}]}\right] - E\left[X \frac{e^{tX}}{E[e^{tX}]}\right]^2
\end{aligned}
$$

We recognize $u''(t)$ as the variance under the tilted measure Q_t which is defined by $E_{Q_t}[\cdot] = E\left[\cdot \frac{e^{tX}}{E[e^{tX}]}\right]$. with tilted density $\frac{e^{tX}}{E[e^{tX}]}$ and thus by part 1. (applied to Q_t) we have $u''(t) \le \frac{(b-a)^2}{4}$.
Using the Taylor expansion with remainder we have, for some ξ between 0 and t

$$
\ln M(t) = u(t) = u(0) + u'(0)t + u''(\xi)\frac{t^2}{2} \le tE[X] + \frac{t^2(b-a)^2}{8} .
$$

This concludes the proof. $\square$

Remark: The bound on the variance in 1. is optimal. Indeed taking without loss of generality $a = 0$ and $b = 1$ then the variance is bounded by $1/4$ and this realized by taking X to be a Bernoulli with $p = \frac{1}{2}$. This bound says that the RV with the largest variance is the one where the mass is distributed at the end point.

The bound in 2. is optimal only in the sense that it is the best *quadratic* bound on $u(t)$. For example for a Bernoulli with $a = 0$ and $b = 1$ we have $M(t) = \ln(\frac{1}{2}e^t + \frac{1}{2}) = \frac{1}{2}t + \ln\cosh\left(\frac{t}{2}\right)$ which is much smaller (for large t). There is room for better bounds but using Gaussian is computationally convenient.

If we appply the Hoeffding's bound to a sum of random variables we find

Theorem 2.6 (Hoeffding's Theorem) Suppose $X_1, \cdots, X_n$ are independent random variables such that $a_i \leq X \leq b_i$ (almost surely). Then

$$\begin{aligned} P\left(X_1 + \cdots + X_n - E[X_1 + \cdots + X_n] \geq \varepsilon\right) \leq e^{-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}} \\ P\left(X_1 + \cdots + X_n - E[X_1 + \cdots + X_n] \leq -\varepsilon\right) \leq e^{-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}} \end{aligned} \tag{2.2}$$

Proof. Using independence the Hoeffding's bound we have

$$e^{t(X_1 + \cdots + X_n - E[X_1 + \cdots + X_n])} = \prod_{i=1}^n e^{t(X_i - E[X_i])} \leq \prod_{i=1}^n e^{\frac{t^2 (b_i - a_i)^2}{8}} = e^{\frac{t^2 \sum_i (b_i - a_i)^2}{8}}$$

and using Chernov bound (for a Gaussian RV with variance $\sum_i \frac{(b_i - a_i)^2}{4}$) gives the first bound in Equation 2.2.

The second bound is proved similarly. $\square$.

We obtain from this bound a confidence interval for emprical sum

Corollary 2.1 (non-asymptotic confidence interval) Suppose $X_1, \cdots, X_n$ are independent random variables such that $a \leq X \leq b$ (almost surely) and $\mu = E[X_i]$.

$$P\left(\mu \in \left[\frac{S_n}{n} - \varepsilon, \frac{S_n}{n} + \varepsilon\right]\right) \geq 1 - 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}$$

Proof. This is Hoeffding's bound with ε replaced by $n\varepsilon$ and with $\sum_{i=1}^n (b_1 - a_1)^2 = n(b-a)^2$. $\square$.

Using

$$\delta = 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}} \iff \epsilon = \sqrt{\frac{(b-a)^2 \ln\left(\frac{2}{\delta}\right)}{2n}}$$

we get the confidence interval

$$P\left(\mu \in \left[\frac{S_n}{n} - \sqrt{\frac{(b-a)^2 \ln\left(\frac{2}{\delta}\right)}{2n}}, \frac{S_n}{n} + \sqrt{\frac{(b-a)^2 \ln\left(\frac{2}{\delta}\right)}{2n}}\right]\right) \geq 1 - \delta$$

2.11 Bernstein bound

In Hoeffding's bound we use, in an essential way, a bound on the variance. If the variance is small then one should expect the bound to be poor. The Bernstein bound can be used if we have some a-priori knowledge about the variance.

Theorem 2.7 (Bernstein Bound) Suppose X is a random variable such that $|X - E[X]| \le c$ and $\mathrm{var}(X) \le \sigma^2$. Then

$$E[e^{tX}] \le e^{tE[X] + \frac{\sigma^2 t^2}{2(1-c|t|/3)}} .$$

Proof. We expand the exponential and use that for $k \ge 2$, with $\mu = E[X]$,

$$E\left[(X-\mu)^k\right] \le E\left[(X-\mu)^2 |X-\mu|^{k-2}\right] \le E[(X-\mu)^2]c^{k-2} \le \sigma^2 c^{k-2}$$

and get

$$\begin{aligned} E\left[e^{t(X-\mu)}\right] = 1 + \sum_{k=2}^{\infty} \frac{t^k}{k!} E[(X-\mu)^k] &\le 1 + \frac{t^2\sigma^2}{2} \sum_{k=2} \frac{2}{k!} (|t|c)^{k-2} \\ &\le 1 + \frac{t^2\sigma^2}{2} \sum_{k=2}^{\infty} \left(\frac{|t|c}{3}\right)^{k-2} \qquad \text{since } \frac{k!}{2} \ge 3^{k-2} \\ &\le 1 + \frac{t^2\sigma^2}{2(1-\frac{|t|c}{3})} \le e^{\frac{t^2\sigma^2}{2(1-\frac{|t|c}{3})}} \qquad \text{since } 1+x \le e^x \quad \square \end{aligned}$$

To combine this we a Chernov bound we have to solve the following optimzation problem which after some straightforward but lengthy computation gives

$$\sup_{t\geq 0}\left\{\varepsilon t - \frac{at^2}{2(1-bt)}\right\} = \frac{a}{b^2} h\left(\frac{b\epsilon}{a}\right) \qquad \text{where } h(u) = 1 + u - \sqrt{1+2u}$$

Note that we can invert the function h and we have $h^{-1}(z) = z + \sqrt{2z}$. This make the Bernstein bound especially convenient to get explcit formulas. By symmetry we find the same bound for the left tail.

Theorem 2.8 (Bernstein for sum of IID) If $X_1, \cdots, X_N$ are IID random variables with $|X_i| \leq c$ and $\mathrm{Var}(X_i) \leq \sigma^2$ then

$$P\left(\mu \in \left[\frac{S_n}{n} - \frac{c}{3n}\ln\left(\frac{2}{\delta}\right) - \sqrt{\frac{2\sigma^2}{n}\ln\left(\frac{2}{\delta}\right)}, \frac{S_n}{n} + \frac{c}{3n}\ln\left(\frac{2}{\delta}\right) + \sqrt{\frac{2\sigma^2}{n}\ln\left(\frac{2}{\delta}\right)}\right]\right) \geq 1-\delta$$

Proof.

$$P\left(\frac{S_n}{n}-\mu\geq\varepsilon\right)=P\left(X_1+\cdots+X_n-\geq n(\mu+\varepsilon)\right)\leq e^{-n\sup_{t\geq 0}\{\varepsilon-\frac{\sigma^2t^2}{2(1-ct/3)}\}}$$

and we obtain the same bound for $P\left(\frac{S_n}{n}-\mu\geq\varepsilon\right)$.

To obtain a confidence interval we need to solve

$$\delta=2e^{-n\frac{a}{b^2}h\left(\frac{b\varepsilon}{a}\right)}\iff\varepsilon=b\frac{1}{n}\ln\left(\frac{2}{\delta}\right)+\sqrt{2a\frac{1}{n}\ln\left(\frac{2}{\delta}\right)}$$

and set $a=\sigma^2$ and $b=c/3$ to obtain the desired bound.

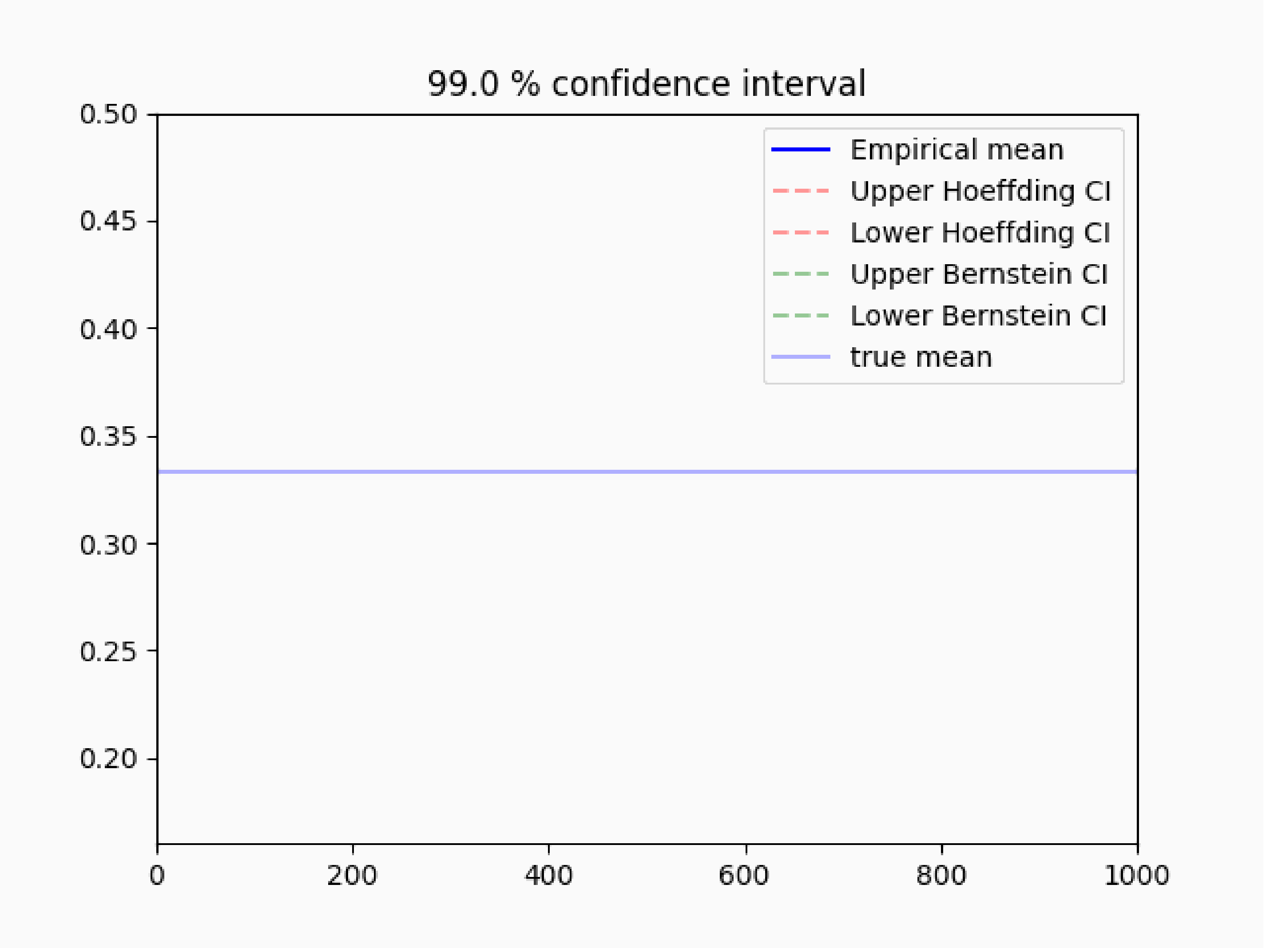
Comparison: Taking $a=0,b=1$

$$\text{(Bernstein)}\quad\frac{c}{3n}\ln\left(\frac{2}{\delta}\right)+\sqrt{2\ln\left(\frac{2}{\delta}\right)}\frac{\sigma}{\sqrt{n}}\quad\text{versus}\quad\sqrt{2\ln\left(\frac{2}{\delta}\right)}\frac{\sigma_{max}}{\sqrt{n}}\quad\text{(Hoeffding)}$$

You should note that this bound can be substantially better than Hoeffding's bound provided $\sigma\leq\sigma_{max}$, if n is large then the $1/n$ term is much smaller than the $1/\sqrt{n}$ term and so Bernstein bound becomes better. See the illustration on the next page.

There is an even more sophisticated version of this bound where one uses the sample variance to build a bound. One needs then to estimate the probability that the sample varaince is far from the true variance which itself requires more sophisticated inequalities using martingales.

As an illustration we plot the empirical mean as well as Hoeffding's and Bernstein's confidence interval for computing the mean beta RV so in Hoeffding's the maximum variance is $\frac{1}{4}$ and we can take $c = 1$ in Bernstein. Here we take the parameter $\alpha = 1$ and $\beta = 2$ so the true mean is $\frac{\alpha}{\alpha+\beta} = \frac{1}{3}$ and the true variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{1}{18}$ which is quite a bit smaller than $\frac{1}{4}$.



confidence intervals with Bernstein and Hoeffding

2.12 Exercises

Exercise 2.1 Suppose X_i are independent and identically distributed normal random variable with mean 1 and variance 3. Compute

$$\lim_{n\to\infty} \frac{X_1 + X_2 + \cdots + X_n}{X_1^2 + X_2^2 + \cdots + X_n^2}$$

Exercise 2.2 Suppose X_i are independent and indentically distributed with $E[X_i] > 0$. Show that $\lim_{n\to\infty} S_n = \lim_{n\to\infty} X_1 + \cdots + X_n = +\infty$ almost surely.

Exercise 2.3 Suppose X and Y are two random variables with finite variance and you want to estimate the correlation coefficient

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V[Y]}}\,.$$

Use the law of large number to find an estimator for ρ using n independent copies of (X_i, Y_i).

Exercise 2.4 Suppose X_i are independent and indentically distributed and strictly positive (i.e $P(X_i > 0) = 1$). Show that, almost surely,

$$\lim_{n \to \infty} (X_1 X_2 \cdots X_n)^{\frac{1}{n}} = \alpha \, .$$

and compute α. This is a simple model used in financial application where X_i describe the change of your investment in any given day: if you start with F_0 your fortune is $F_0 X_1$ after the first day, $F_0 X_1 X_2$ after the second day, and so on.....

3 Weak convergence aka convergence in distribution

3.1 Weak convergence of probability measures

In the notion of weak convergence we do not view random variables as map $X : \Omega \to \mathbb{R}$ and work exclusively with their distribution P^X (that is probability measures on $\mathbb{R}$). Actually one can talk about weak convergence of random variables even if they do not live on the same probability space!

Definition 3.1 (Weak convergence of probability measures and convergence in distribution for random variables)

1. The sequence (P_n) of probability measures on $\mathbb{R}$ **converges weakly to** μ if

$$\lim_{n\to\infty} \int f dP_n = \int f dP$$

for any $f : \mathbb{R} \to \mathbb{R}$ *bounded and continuous.*

2. The {sequence of RVs (X_n) **converges to** X **in distribution** if the distribution P^{X_n} of X_n converges weakly to the distribution P^X of X, i.e.,

$$\lim_{n\to\infty} E[f(X_n)] = E[f(X)]$$

for any $f : \mathbb{R} \to \mathbb{R}$ *bounded and continuous.*

Remark: We can generalize this easily to RV taking values in $\mathbb{R}^n$ or some metric space (so we can talk about bounded continuous function). We stick with $\mathbb{R}$ for simplicity.

3.2 Simple properties and some examples

Theorem 3.1 (weak means weak)

1. If X_n converges to X **in L^p or in probability, or almost surely** then X_n converges to X **weakly**.
2. If X_n converges indistribution to a constant RV $X = a$ then X_n converges to a in probability.

Proof. We show that convergence in probability implies convergence in distribution. Since almost sure convergence and convergence in L^p implies convergence in probability, this will prove 1. But as we have proved in Theorem 1.4 if X_n converges to X in probability, the continuity of f implies that $Y_n = f(X_n)$ converges to $Y = f(X)$ in probability. Since f is bounded the random variables Y_n and Y are bounded (and so in any L^p). As we proved in Theorem 1.7 this implies that $\lim_{n\to\infty} E[f(X_n)] = E[f(X)]$.

For the (partial) converse statement in 2. we take assume that X_n converges weakly to a constant $X = a$ and consider the continuous function $f(x) = \min\{|x-a|, 1\}$. Then we have

$$E[\min\{|X_n - X], 1\}] = E[f(X_n)] \to E[f(X)] = f(a) = 0$$

By Theorem 1.6 this implies thatX_n converges to X in probability.

Examples:

1. If X_i are identically dsitributed RV (i.e. $P^{X_n} = P$ for all n then X_i converges in distribution but X_n does not need to converges in any other sense except if $X = X_1 = X_2 = \cdots$. An example is when X_i are IID Cauchy RV with parameter β then $\frac{S_n}{n} = \frac{1}{n}(X_1 + \cdots + X_n)$ are also Cauchy RV with paramter β. Then both X_n and $\frac{S_n}{n}$ both converges in distribution to a Cauchy RV paramter β but these sequences do not converges in any other sense.

2. The measure $P^n = \frac{1}{n}\sum_{i=1}^n \delta_{\frac{i}{n}}$ converges weakly to the Lebesgue measure on $[0,1]$. Indeed for f continuous, the Riemman sum

$$\int f dP_n = \frac{1}{n}\sum_{i=1}^n f\left(\frac{i}{n}\right) \to \int f dx\,.$$

3. If $P_n = \delta_{x_n}$ then P_n converges weakly to P if and only if $P = \delta_x$ and $x_n \to x$.

4. *Convergence of the quantile functions:* Let Q_n be the quantile function for X_n and Q the quantile function for X and let P_0 be Lebesgue measure on $[0,1]$. Then $P^{X_n} = P_0 \circ Q_N^{-1}$ and $P^X = P_0 \circ Q^{-1}$.
If the quantiles Q_n converges to Q for P_0 almost all $\omega \in [0,1]$ then for f continuous and bounded $f \circ Q_n$ converges to $f \circ Q$ for P_0 almost all $\omega \in [0,1]$. By the bounded convergence theorem

$$E[f(X_n)] = \int_{\mathbb{R}} f(x) dP^{X_n}(x) = \int_{[0,1]} f(Q_n(\omega)) dP_0(\omega) \to \int_{[0,1]} f(Q(\omega)) dP^0(\omega) = E[f(X)]$$

and thus X_n converges in distribution to X.

3.3 Convergence of sets probabilities

Basic question If P_n converges weakly to P and $A \subset \mathbb{R}$ is some measurable set what does this imply for the convergence of $P_n(A) = \int 1_A dP_n$?

Notation If A is a set, then we denote by $\overline{A}$ the closure of A, by $\mathring{A}$ the interior of A, and by ∂A the boundary of A. We have

$$\overline{A} = A \cup \partial A \quad \mathring{A} = A \setminus \partial A \quad \partial A = \overline{A} \setminus \mathring{A}$$

Theorem 3.2 (Weak convergence and set covergence) The following are equivalent

1. P_n converges weakly to P.
2. For any A *closed* we have $\limsup_n P_n(A) \le P(A)$.
3. For any A *open* we have $\liminf_n P_n(A) \ge P(A)$.
4. For any A with $P(\partial A) = 0$ we have $\lim_n P_n(A) = P(A)$.

We will prove

$$1. \implies 2. \iff 3. \implies 4. \implies 1$$

Proof. • Asssume 1. hold and A is a closed set. We consider the ϵ-neighborhood of A:

$$A_\epsilon = \{x, d(x, A) < \epsilon\} \qquad \text{where} \quad d(x, A) = \inf\{d(x, y)\,;\, y \in A\}$$

Since A is closed $A_\epsilon \searrow A$ and so by sequential continuity $P(A_\epsilon) \searrow P(A)$.
Consider now function $f(x) = \min\left\{1 - \frac{d(x,A)}{\epsilon}, 0\right\}$ which is bounded and continuous and satisfies $1_A \le f(x) \le 1_{A_\epsilon}$. From this we see that

$$P_n(A) \le \int f dP_n \qquad \text{and} \qquad \int f dP \le P(A_\epsilon)$$

Weak convergence means $\int f dP_n \to \int f dP$ and thus $\limsup_n P_n(A) \le P(A_\epsilon)$. Since this holds for all ϵ we have proved 2.

• 2. and 3. are equivalent since complements of closed sets are open and vice versa and $\liminf(1 - r_n) = 1 - \limsup_n r_n$ for any sequence r_n in $[0, 1]$.

• Assume 3. and then also 2. hold. Let A be a Borel set, since $\overline{A} \supset A \supset \mathring{A}$ we have

$$P(\overline{A}) \ge \limsup_n P_n(\overline{A}) \ge \limsup_n P_n(A) \ge \liminf_n P_n(A) \ge \liminf_n P_n(\mathring{A}) \ge P(\mathring{A})$$

If $P(\partial A) = 0$ then $P(\overline{A}) = P(\mathring{A})$ and this implies that $\lim_n P_n(A) = P(A)$.

• Take f bounded and continuous so that $a < f(x) < b$ and consider the probability $P \circ f^{-1}$ on (a, b). Since atoms are countable we pick a partition $a_0 = a < a_1 < a_2 < \cdots < a_m = b$ so that $a_i - a_{i-1} < \epsilon$ and a_i is not an atom for $P \circ f^{-1}$.
Set $A_i = f^{-1}((a_{i-1}, a_i])$ and define the simple functions

$$g = \sum_i a_{i-1} 1_{A_i}\,, \qquad h = \sum_i a_i 1_{A_i}\,.$$

By construction we have

$$f - \epsilon \le g \le f \le h \le f + \epsilon. \tag{3.1}$$

Since f is continuous if $x \in \partial A_i$ (A_i is collection of intervals) then $f(x) = a_{i-1}$ or $f(x) = a_i$ which are not atoms. Therefore $P(\partial A_i) = 0$ and thus $P_n(\partial A_i) \to 0$ as $n \to \infty$. This implies that $\int gdP_n \to \int gdP$ and $\int hdP_n \to \int hdP$ and thus by Equation 3.1 we have

$$\begin{aligned}\int fdP - \epsilon \le \int gdP = \lim_n \int gdP_n \le \liminf_n \int fdP_n \\ \le \limsup_n \int fdP_n \le \lim_n \int hdP_n = \int hdP \le \int fdP + \epsilon\end{aligned}$$

Since ϵ is arbitrary we $\lim_n \int fdP_n = fdP. \quad \square$.

3.4 Uniqueness of limits

- Suppose P and Q are two probability measures on $\mathbb{R}$ and suppose that $\int f dP = \int f dQ$ for all bounded continuous functions. Then we claim that $P = Q$.
- We can provide a simple proof using Theorem 3.2. Take $P_n = P$ for all n then P_n converges weakly to Q and so for any open set A we have $\liminf P_n(A) = P(A) \geq Q(A)$. Exchanging the role of P and Q we get $Q(A) \geq P(A)$ and thus P and Q coincide on all open sets and such sets form a p-system which generate the σ-algebra.
- As a consequence limits in weak convergence are unique, if P_n converges weakly to P and Q then $P = Q$.

3.5 Convergence of distribution and quantile functions

For random variables weak convergence is called convergence in distribution and the following theorem explain why.

Theorem 3.3 (Convergence in distribution and convergence of the distribution functions) The following are equivalent

1. The random variables X_n converges to X in distribution.
2. The CDF $F_{X_n}(t) \to F(t)$ at every continuity point of F.
3. The quantile functions $Q_{X_n}(z) \to Q_X(z)$ at every continuity point of Q.

Proof. • $1. \implies 2.$: Suppose t is a continuity point of $F_X(t)$ then x is not an atom for P. From Theorem 3.2 part 4. we see that $F_{X_n}(t) = P^{X_n}((-\infty, t]))$ converges to $P^X((-\infty, t])) = F(t)$.

• $2. \implies 3.$: Suppose z is a point of continuity for Q and let $t = Q(z)$. Fix $\epsilon > 0$, and choose $s \in (t-\epsilon, t)$ and $r \in (t, t+\epsilon)$ to be continuity points of F. Since Q is continuous at z, then F is not flat at level z and thus $F(s) < z < F(r)$. Since $F_n(s) \to F(s)$ by assumption we have $F_n(s) < z$ and thus $Q_n(z) > s > t - \epsilon$ for all but finitely many n. This means that $\liminf Q_n(z) > t - \epsilon$. A similar argument shows that $\limsup Q_n(z) < t + \epsilon$. Thus $\lim_n Q_n(z) = Q(z)$ and 3. holds.

• $3. \implies 2.$: The quantile function being increasing has only countable many discontinuities and thus continuous Lebesgue almost everywhere. As we have seen in the example in Section 3.2 this implies convergence in dsitribution.

Example If X_i are independent and identically distributed random variables with commmon CDF $F(t)$ then the Glivenko-Cantelli theorem implies that for almost all ω

$$F_n(t) = \frac{1}{n}\sum_{k=1}^{n} 1_{\{X_k(\omega)\le t\}} \to F(t) \quad \text{for all } t.$$

That is the empirical (random) measure $\frac{1}{n}\sum_{k=1}^{n}\delta_{X_k(\omega)}$ converges weakly to P^X for almost all ω. In this example the convergence occurs for all t even if F has dsicontinuities.

Example Consider the random variables X_n with distribution function

$$F_n(t) = \begin{cases} 0 & t \le -\frac{1}{n} \\ \frac{1}{2} + \frac{n}{2}t & -\frac{1}{n} < t < \frac{1}{n} \\ 1 & t \ge \frac{1}{n} \end{cases}$$

Then $F_n(t)$ converges to 0 for $t \neq 0$ and $F_n(0) = \frac{1}{2}$ for all n. So $F_n(t)$ converges to $F(t) = 1_{[0,\infty)}(t)$ at all continuity pooints of F. So a uniform random variable on $[-\frac{1}{n}, \frac{1}{n}]$ converges weakly to the random varibale $X = 0$.

Example: extreme value and maxima of Pareto distribution Consider $X_1, \cdots, X_n$ to be independent Pareto RV each with cumulative distribution function $F(t) = 1 - \frac{1}{t^\alpha}$ for $t \geq 1$ and 0 for $t \leq 1$. We are interested in the distribution of the maximum

$$M_n = \max_{m \leq n} X_n$$

for the limit of large n. We have, by independence

$$P(M_n \leq t) = P(X_1 \leq t, \cdots, X_n \leq t) = F(t)^n = \left(1 - \frac{1}{t^\alpha}\right)^n$$

This suggest the scaling $t = n^{1/\alpha} y$ so that

$$P(M_n \leq n^{1/\alpha} y) = \left(1 - \frac{y^{-\alpha}}{n}\right)^n \to e^{-y^\alpha} \quad \text{as } n \to \infty.$$

Thus we proved that $M_n / n^{1/\alpha}$ converges in distribution to a distribution which is called a *Frechet distribution*.

3.6 Convergence of densities

We show next that convergence of the densities $f_n(x)$ implies convergence in distribution.

Theorem 3.4 Suppose X_n is sequence of random variables with densities $f_n(x)$ and $f_n(x)$ converges Lebesgue almost everywhere to a density $f(x)$. Then X_n converges in dsitribution to the random variables X with density $f(x)$.

Proof. The cumulative distribution function $F(t) = \int_{-\infty}^t f_n(x)dx$ is continuous for every t and we would like to show that $F_n(t)$ converges to $F(t) = \int_{-\infty}^t f(x)dx$ for every t. However we cannot use dominated convergence theorem since there is no dominating function for f_n.

We prove instead that for any h bounded and continuous we have $\lim_n E[h(X_n)] = E[h(x)]$ using that f_n is non-negative and is normalized. Since h is bounded we set $\alpha = \sup_x |h(x)|$ and consider the two non-negative function

$$h_1(x) = h(x) + \alpha \ge 0 \quad h_2 = \alpha - h(x) \ge 0$$

We now apply Fatou's Lemma to the sequence of non-negative functions $h_1(x)f_n(x)$ and $h_2(x)f_n(x)$. We have for $i = 1, 2$

$$E[h_i(X)] = \int f(x)h_i(x)dx \le \liminf_n \int f_n(x)h_i(x)dx = \liminf_n E[h_i(X_n)]$$

From this we obtain

$$E[h(x)] + \alpha \leq \liminf_n E[h(X_n)] + \alpha \quad \text{and} \quad \alpha - E[h(x)] \leq \alpha - \limsup_n E[h(X_n)]$$

and thus $E[h(X)] = \lim_n E[h(X_n)]$. $\square$.

Example Suppose X_n is a sequence of normal random variables with mean μ_n and variance σ_n. If $\mu_n \to \mu$ and $\sigma_n \to \sigma > 0$ then X_n converges to a normal random variable with mean μ and variance σ^2. This follows from the fact that the density of a normal random variable is a continous function of μ and σ^2.

3.7 Some remarks on convergence in total variation

The convergence of densities implies in fact a stronger mode of covergence than weak convergence.

We have not used the fact that h is continuous in the proof and thus we proved here that

$$\lim_n E[h(X_n)] = E[h(X)] \quad \text{for all } h \text{ bounded and measurable}$$

In particular we can take $h = 1_A$ for any measurable set A and we have

$$\lim_n P(X_n \in A) \to P(X \in A) \quad \text{for all measurable sets} A, .$$

This convergence is much stronger that weak convergence and is called convergence in total variation.

3.8 Tightness and Prohorov Theorem

Basic question: We prove next a *compactness result* with respect to weak convergence: given a collection of probability measures P^n on $\mathbb{R}$ when can we expect to have the existence of a (weakly) convergent subsequence?

We first need a new concept.

Definition 3.2 (Tightness) A collection of probability measure $\mathcal{P} = \{P_i\}_{i \in I}$ is **tight** if for any $\epsilon > 0$ there exists R such that

$$P_i([-R, R]) \geq 1 - \epsilon \quad \text{for all } i \in I$$

The following theorem is (a version of) Prohorov theorem which actually holds on more general spaces than $\mathbb{R}$ (actually any complete separable metric space).

Theorem 3.5 (Prohorov theorem on $\mathbb{R}$) If a collection of probability measure $\mathcal{P}$ is tight then any sequence of measure $\{P_n\}$ with $P_n \in \mathcal{P}$ has a subsequence which converges weakly to some probability measure P.

Proof. The proof use the cumulative distribution function $F_n(t) = P_n((-\infty, t])$. Since for any $t \in \mathbb{R}$ we have $0 \leq F_n(t) \leq 1$. by Bolzano-Weierstrass there exists a subsequence F_{n_k} such that $F_{n_k(t)}$ converges. Of course the subsequence n_k will depend a prioiri on t.

To construct the limit for any t we use a diagonal sequence argument: consider an enumaration $r_1, r_2, \cdots$ of the rational $\mathbb{Q}$.

- For r_1 there exists a subsequence $n_{1,k}$ such that the limit exists and we set

$$G(r_1) = \lim_{k\to\infty} F_{n_{1,k}}(r_1)\,.$$

- For r_2 there exists a sub-subsequence $n_{2,k}$ of $n_{1,k}$ such that the limit exists and we set

$$G(r_2) = \lim_{n\to\infty} F_{n_{2,k}}(r_2)\,.$$

  and note that we also $G(r_1) = \lim_{n\to\infty} F_{n_{2,k}}(r_1)$.

- Continuing in this way for r_j we have subsequence $n_{j,k}$ and we set

$$G(r_j) = \lim_{n\to\infty} F_{n_{j,k}}(r_j)$$

  and note that $F_{n_{j,k}}(r)$ converges for $r = r_1, r_2, \cdots, r_j$.

- Finally consider the diagonal sequence $n_k = n_{k,k}$ and we have for all j

$$G(r_j) = \lim_{n\to\infty} F_{n_k}(r_j)$$

  since n_k is a subsequence of $n_{j,k}$ for $k \geq j$.

We now define a function F on $\mathbb{R}$ by setting

$$F(t) = \inf\{G(r), r \geq t \text{ rational}\}$$

Since G is non-decreasing, F is also non-decreasing and it is right continuous by construction.

We now use the tightness hypothesis and choose R so that $P_n([-R, R]) \geq 1-\epsilon$ for all n simultaneoussy. This implies that

$$F_n(t) \leq \epsilon \text{ for } t \leq -R \quad \text{and} \quad F_n(t) \geq 1-\epsilon \text{ for } t \geq R\,.$$

The same holds for the function G and finally also for the function F and we have

$$F(t) \leq \epsilon \text{ for } t \leq -R \quad \text{and} \quad F(t) \geq 1-\epsilon \text{ for } t \geq R\,.$$

Since $0 \leq F \leq 1$, F is right-continuos and decreasing and ϵ is arbitrary this shows that $F(t) \to 0$ as $t \to -\infty$ and $F(t) \to 1$ as $t \to +\infty$ and F is the cumulative distribution function for some probability measure P.

To conclude we need to prove that $F_{n_k}(t)$ converges to $F(t)$ for all continuity points of F. Assuming that $F(t_-) = F(t)$ we see that there exists $r, s \in \mathbb{Q}$ such that

$$F(t) - \epsilon < G(r) \leq F(t) \leq G(s) \leq F(t) + \epsilon$$

If k is large enough we have

$$F(t) - 2\epsilon < F_{n_k}(r) \leq F_{n_k}(t) \leq F_{n_k}(s) \leq F(t) + 2\epsilon$$

Thus

$$F(t) - 2\epsilon < F(r) \leq \liminf_k F_{n_k}(t) \leq \limsup_k F_{n_k}(t) \leq F(s) \leq F(t) + 2\epsilon$$

and since ϵ is arbitrary $\lim_k F_{n_k}(t)$ exists and must be equal to $F(t)$. By Theorem 3.3 this shows that P_n converges weakly to P.

3.9 Weak convergence and characteristic function

A fundamental result to prove the central limit theorem is the following

Theorem 3.6 (Lévy continuity theorem) Let P_n be a sequence of probability measure on $\mathbb{R}$ and $\widehat{P}_n(t)$ their Fourier transforms.

1. P_n converges weakly to P implies that $\widehat{P}_n(t)$ converges pointwise to $\widehat{P}(t)$
2. If $\widehat{P}_n(t)$ converges pointwise to a function $h(t)$ which is continuous at 0 then $h(t) = \widehat{P}(t)$ is the Fourier transform of a measure P and P_n converges weakly to P.

Proof. For 1. just note that e^{itx} is bounded and continuous and thus weak convergence implies convergence to the charatersitic function for every t.

For 2. we show first that if $\widehat{P}_n(t)$ converges to a function $h(t)$ which is continuous at 0 then the sequence P_n is *tight*. By Fubini theorem

$$\int_{-\alpha}^{\alpha} \widehat{P}_n(t)dt = \int_{-\infty}^{\infty}\int_{-\alpha}^{\alpha} e^{itx} dt dP_n(x) = \int_{-\infty}^{\infty}\int_{-\alpha}^{\alpha} \cos(tx) dt dP_n(x) = \int_{-\infty}^{\infty} \frac{2}{x}\sin(\alpha x) dP_n(x)$$

Next using the following easy bound

$$2\left(1-\frac{\sin(v)}{v}\right)\begin{cases}\geq 1 & \text{if } |v|\geq 2\\ \geq 0 & \text{always}\end{cases}$$

we obtain

$$\frac{1}{\alpha}\int_{-\alpha}^{\alpha}(1-\widehat{P}_n(t))dt=\int_{-\infty}^{\infty}2\left(1-\frac{\sin(\alpha x)}{\alpha x}\right)dP_n(x)\geq\int_{\alpha|x|\geq 2}dP_n(x)=P_n\left(\left[-\frac{2}{\alpha},\frac{2}{\alpha}\right]^c\right) \tag{3.2}$$

Now since $\widehat{P}_n(0)=1$ for all n we have $h(0)=1$ and so by the *assumed* continuity of h we can choose α sufficiently small so that

$$\frac{1}{\alpha}\int_{-\alpha}^{\alpha}|1-h(t)|dt\leq\frac{\epsilon}{2} \tag{3.3}$$

By the bounded convergence theorem we have

$$\lim_{n\to\infty}\frac{1}{\alpha}\int_{-\alpha}^{\alpha}|1-\widehat{P}_n(t)|dt=\frac{1}{\alpha}\int_{-\alpha}^{\alpha}|1-h(t)|dt$$

and so combining Equation 3.2 with Equation 3.3 we find that for n large enough

$$P_n\left(\left[-\frac{2}{\alpha},\frac{2}{\alpha}\right]^c\right)\leq\epsilon\,.$$

This ensures that the sequence $\{P_n\}$ is tight.

To conclude we invoke Theorem 3.5: for any subsequence P_{n_k} there exists a subsubsequence $P_{n_{k_j}}$ which converges weakly to some probability measure P. By part 1. this implies that $\lim_j \widehat{P}_{n_{k_j}}(t)$ converges $\widehat{P}(t)$ which must then be equal to $h(t)$. This shows that $h(t)$ is the characteristic function for the probability measure P and this shows that the limit is the same for any choice of subsequence n_k. This implies that P_n converges weakly to P. $\square$.

Example If Z is Poisson then $E\left[e^{itZ}\right] = e^{\lambda(e^{i\lambda t}-1)}$. Take Z_n Poisson with $\lambda = n$ and set $Y_n = \frac{Z_n - n}{\sqrt{n}}$

$$\begin{aligned} E[e^{itY_n}] = &\quad E\left[e^{i\frac{t}{\sqrt{n}}(Z-n)}\right] = e^{-it\sqrt{n}} E\left[e^{i\frac{t}{\sqrt{n}}Z}\right] = e^{-it\sqrt{n}} e^{n\left(e^{i\frac{t}{\sqrt{n}}}-1\right)} \\ = &\quad e^{-it\sqrt{n}} e^{n\left(i\frac{t}{\sqrt{n}} - \frac{t^2}{2n} + O(n^{-3/2})\right)} = e^{-\frac{t^2}{2}} e^{O(n^{-1/2})} \end{aligned}$$

So Y_n converges weakly to a standard normal.
This is exactly the kind of computation that we will use to prove the central limit theorem in the next section.

3.10 Exercises

Exercise 3.1 (Continuity Theorem for convergence in distribution) Show that if X_n converges in distribution to X and f is a continuous function then $f(X_n)$ converges in distribution to $f(X)$.

Exercise 3.2 (Convergence of distributions for discrete random varables) Suppose X_n and X takes values in $\mathbb{Z}$. Show that X_n converges to X if and only if $P(X_n = j)$ converges to $P(X = j)$ for all $j \in \mathbb{Z}$.
Hint: For the "if" direction pick a finite set $\Lambda \subset \mathbb{Z}$ such that $\sum_{j \in \Lambda} P(X = j) \geq 1 - \epsilon$.

Exercise 3.3 (Criterion for tightness) Suppose ϕ is a non-negative function with $\lim_{|x| \to \infty} \phi(x) = +\infty$. Show that if $C = \sup_n E[\phi(X_n)] < \infty$ then the sequence of random variable X_n is tight (that is the family of distribution P^{X_n} is tight).

Exercise 3.4

1. Show that if X_n converges to X in distribution and Y_n converges to Y in distribution and X_n and Y_n are independent for all n then $X_n + Y_n$ converges to $X + Y$ in distribution. *Hint*: Use the characetristic function.
2. Show with a counterexample that the assumption that X_n and Y_n are independent can, in general, not be dropped in part 1.

Exercise 3.5 Given independent identically distributed random variables $X_1, X_2, \cdots, X_n$ with a common distribution function $F(x) = P(X_j \le x)$ let $M_n = \max_{1\le k\le n} X_k$ be the maximum.

1. Assume that for any finite x we have $F(x) < 1$ (this means that the X_j are unbounded). Show that

$$\lim_{n\to\infty} M_n = +\infty \text{ almost surely.}$$

Hint: Fix an arbitrary R and consider the event $A_n = P(Y_n \le R)$. Apply then Borel-Cantelli Lemma.

2. Assume that we have $F(x_0) = 1$ and $F(x) < 1$ if $x < x_0$ (this means that X_j are bounded). Show that

$$\lim_{n\to\infty} M_n = x_0 \text{ almost surely.}$$

Hint: Argue as in 1.

3. Suppose that X_j are an exponential random variable with distribution function $F(x) = 1 - e^{-x}$. From part 1. we know that M_n diverges almost surely. In order to characterize this divergence show that

$$\lim_{n\to\infty} P(M_n - \log n \le x) = e^{-e^{-x}}$$

The random variable Z with distribution function $P(Z \le x) = e^{-e^{-x}}$ is called a Gumbell distribution.

4 Central limit theorem

The LLN asserts that for IID random variables the empirical mean \$ \$converge to the mean $\mu = \backslash\mathrm{E}[X_i]$. The central limit theorem describes the small flucttations around the mean. Informally it says that if the X_i are independent and identically distributed and have finite variance then

$$\frac{S_n}{n} \approx \mu + \frac{\sigma}{\sqrt{n}} Z \quad \text{as } n \to \infty$$

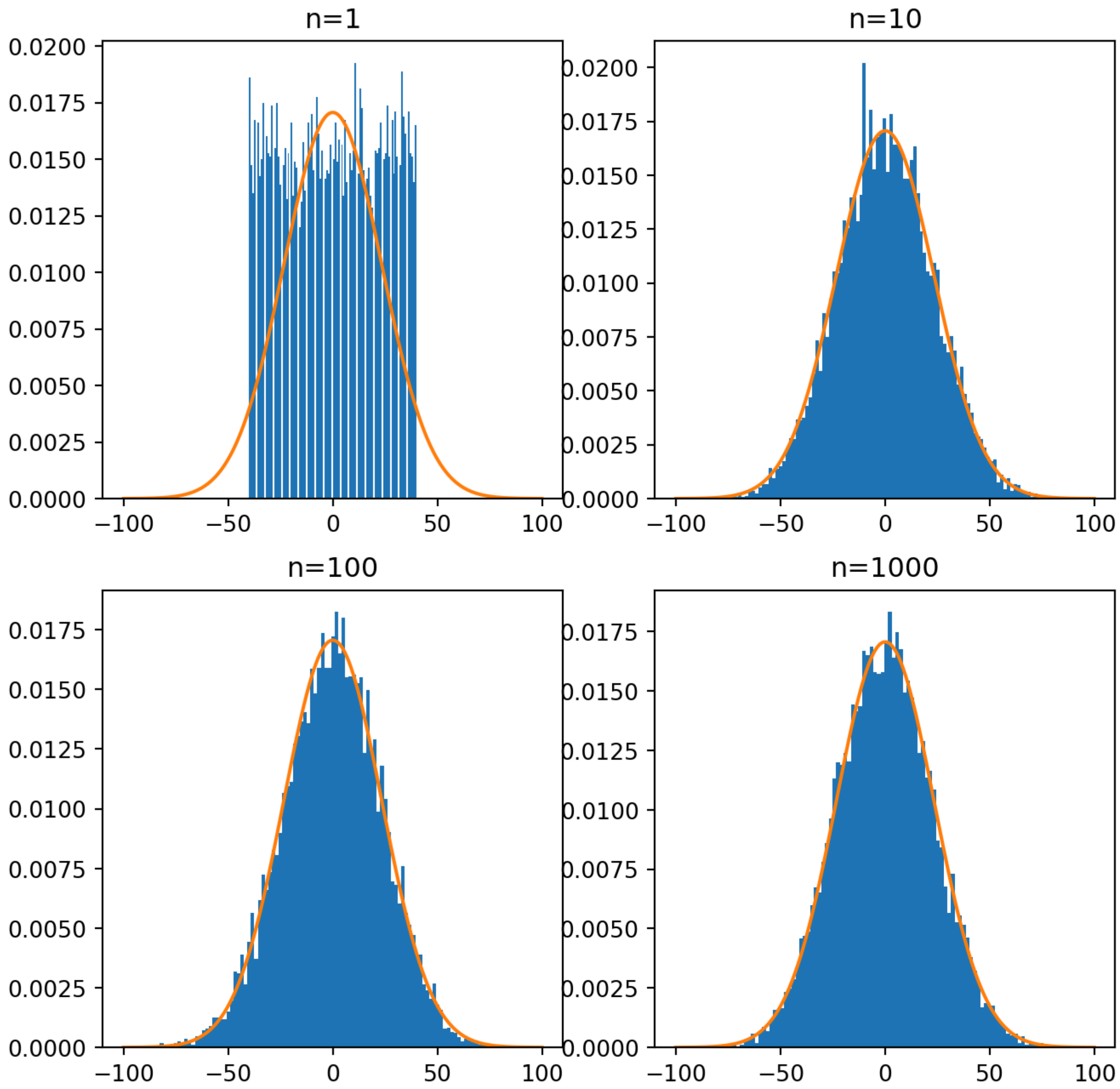
where Z is a standard normal random variable.

4.1 Empirical finding

We take X_k to be uniform on $\{-40, -39, \cdots, 40\}$ with mean $\mu = 0$ and variance $\sigma^2 = \frac{81^2-1}{12}$. For various value of n we generate m IID samples of $\frac{S_n}{n}$ and then rescale them by $\sqrt{n}$ to obtain a variance which is independent of n. We plot then an histogram of the values obtained, comparing with the pdf of a normal distribution with mean 0 and variance σ^2

▼ Code

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

# number of sample in the sample mean
num = [1, 10, 100, 1000]
# list of sample means
means = []

# number of realizations of the sample means

num_re = 10000

# Generating num random numbers from -40 to 40
# taking their mean and appending it to list means.
for j in num:
    x = [np.mean(
        np.random.randint(
            -40, 41, j)) for _i in range(num_re)]
    means.append(x)

k = 0
xrange = np.arange(-100,100,.1)

# plotting all the rescaled means in one figure
fig, ax = plt.subplots(2, 2, figsize=(8, 8))
```


n=1
n=10
n=100
n=1000
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
−100
−50
0
50
100

4.2 The central limit theorem

Theorem 4.1 (Central Limit theorem) Suppose the random variables X_i are IID RVs with $E[X_i] = \mu$ and $\mathrm{Var(X_i)} = \sigma^2$ for all i. Then

$$Y_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to a standard normal random variable Z.

- To understand and remember the the scaling, note that

$$E[Y_n] = 0 \qquad Var(Y_n) = \frac{1}{n\sigma^2}\mathrm{Var}(S_n) = 1$$

- Often, using one version of the convergence in distribution we write that for any a, b we have

$$\lim_{n\to\infty} P\left(a \le \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \le b\right) = \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$$

Proof. We have done most of the work already! By Theorem 3.6 it enough to prove that the characteristic function $E[e^{itY_n}]$ converges to $E[e^{itZ}] = e^{-t^2/2}$ for all t. We denote by ϕ the characteristic function of the random variables $\frac{X_i-\mu}{\sigma}$. We have, using independence,

$$\phi_{Y_n}(t) = E\left[e^{itY_n}\right] = E\left[e^{i\frac{t}{\sqrt{n}\sigma}\sum_{k=1}^n (X_k-\mu)}\right] = \prod_{k=1}^n E\left[e^{i\frac{t}{\sqrt{n}}\frac{X_k-\mu}{\sigma}}\right] = \phi(\frac{t}{\sqrt{n}})^n$$

Since X_i has finite variance by **?@thm-differentiabilityft**, $\phi(t)$ is twice continuously differentiable and we have

$$\phi'(t) = iE\left[\left(\frac{X_k-\mu}{\sigma}\right)e^{it\frac{X_k-\mu}{\sigma}}\right] \qquad \phi"(t) = -E\left[\left(\frac{X_k-\mu}{\sigma}\right)^2 e^{it\frac{X_k-\mu}{\sigma}}\right]$$

and so $\phi'(0)=0$ and $\phi"(0)=-1$ a Taylor expansion around 0 gives

$$\phi(t) = 1 - \frac{t^2}{2} + t^2 h(t) = 1 - \frac{t^2}{2}(1-h(t)) \quad \text{with } \lim t \to 0 h(t) = 0$$

We have then

$$\phi_{Y_n}(t) = \phi\left(\frac{t}{\sqrt{n}}\right)^n = \left(1 - \frac{t^2(1-h(t/\sqrt{n}))}{n}\right)^n \to e^{-t^2/2}$$

where we have used that if $c_n \to c$ then $(1+c_n/n)^n \to e^c$ (by L'Hopital rule). $\square$

4.3 Variations on the CLT

Modifying the proof slightly one can find

Theorem 4.2 Let X_i be independent random variables with $E[X_i] = 0$ for all i and variance $\sigma_i^2 = \mathrm{Var}(X_i)$. Assume $\sup_i \sigma_i^2 < \infty$ and $\sum_i \sigma_i^2 = \infty$. Then

$$\frac{S_n}{\sqrt{\sum_{j=1}^n \sigma_j^2}} \to Z \quad \text{in distribution}$$

where Z is a standard normal.

and also there is a multidimensional version

Theorem 4.3 (multi-dimensional central limit theorem) Let X_i be IID $\mathbb{R}^d$-valued random variables. Let $\mu = E[X_i]$ the vector or means and let Q be the covariance matrix $Q = \mathrm{Cov}(X_i, X_i)$. Then

$$\frac{S_n - n\mu}{\sqrt{n}} \to Z \quad \text{in distribution}$$

where Z is Gaussian with mean vector μ and covariance matrix Q.

4.4 Confidence intervals (version 1)

- We build build confidence interval for $\frac{S_n}{n}$, since by the Central limit theorem $\frac{\sqrt{n}}{\sigma}(\frac{S_n}{n} - \mu)$ is asymptotically normal.

- To build a α-confidence interval we let z_α the number defined by

$$\alpha = \frac{1}{\sqrt{2\pi}} \int_{-z_\alpha}^{z_\alpha} e^{-\frac{x^2}{2}}\, dx \qquad \text{for example} \qquad \begin{cases} z_{.90} = 1.645\ldots \text{ (90\% confidence interval)} \\ z_{.95} = 1.960\ldots \text{ (95\% confidence interval)} \\ z_{.99} = 2.576\ldots \text{ (99\% confidence interval)} \end{cases}$$

- By the CLT $P\left(\mu \in \left[\frac{S_n}{n} - z_\alpha \frac{\sigma}{\sqrt{n}}, \frac{S_n}{n} + z_\alpha \frac{\sigma}{\sqrt{n}}\right]\right) \lessapprox \alpha. \quad \text{as } n \to \infty.$

Approximate α Confidence Interval

$$P\left(\mu \in \left[\frac{S_n}{n} - \epsilon, \frac{S_n}{n} + \epsilon\right]\right) \lessapprox \alpha \quad \text{provided} \quad n \geq z_\alpha \frac{\sigma^2}{\epsilon^2}$$

- The issue with that formula is that we are trying to compute μ so there is no reason to believe that the variance σ^2 should be known! To remedy this issue we will use later the estimator for σ^2 built from our samples $X_1, X_2, \cdots$.

4.5 Applications of the CLT to Monte-Carlo method

In the spirit of the Monte-Carlo method the CLT provides a method to compare different MCMC methods to compute a number μ. The idea is simple: given two Monte-Carlo estimator to compute μ

$$\frac{1}{n}\sum_{k=1}^{n} X_i \to \mu \quad \text{and} \quad \frac{1}{n}\sum_{k=1}^{n} Y_i \to \mu$$

choose the one with the smallest variance since by the central limit theorem the estimator with smallest variance will be more concentrated around μ.

Example: comparing estimator to compute integrals Given a function h (without loss of generality with $0 \le h \le 1$ and defined on $[0,1]$) we have the estimators for $\mu = \int_0^1 h(x)dx$

$$\frac{1}{n}\sum_{k=1}^{n} h(U_i) \to \int_0^1 h(x)dx \quad \text{where } U_i \text{ uniform on} [0,1]$$

and

$$\frac{1}{n}\sum_{k=1}^{n} X_i \to \int_0^1 h(x)dx \quad \text{where } X_i = \left\{ \begin{array}{ll} 1 & \text{if } U \le f(V) \\ 0 & \text{if } U > f(V) \end{array} \right. \quad \text{where } U, V \text{ uniform on} [0,1]$$

Computing the variances we find

$$\mathrm{Var}(h(U)) = \int_0^1 h(x)^2 dx - \left(\int_0^1 h(x) dx\right)^2$$

and

$$\mathrm{Var}(X) = \mu(1-\mu) = \int h(x) - \left(\int_0^1 h(x) dx\right)^2$$

and since $0 \le h(x) \le 1$ we have $h^2(x) \le h(x)$ and thus $\mathrm{Var}(h(U)) \le \mathrm{Var}(X)$.

Importance sampling: Suppose we are trying to compute with a Monte-Carlo method (using a RV X with density $f_X(x)$) the integral $E[h(X)] = \int h(x) f_X(x) dx$.

Suppose for example that $h(x) = 1_{\{x \ge 4\}}$ and X is standard normal. Then $E[h(X)] = P(X < 4) = 0.00003$ which is tiny. To have a meaningful estimate for μ, the CLT gives $S_n/n \approx \mu + \frac{\sigma}{\sqrt{n}} Z$ we must have $\frac{\sigma}{\sqrt{n}} \ll \mu$ or $n \gg \frac{\sigma^2}{\mu^2}$.

The naive estimator using the Bernoulli RV $Y = 1_{\{X \ge 4\}}$ has variance

$$\mathrm{Var}(Y) = P(X \ge 4)(1 - P(X \ge 4)) \approx P(X \ge 4) = \mu$$

and so we need $n \gg \mu^{-1}$ samples.

The idea behind importance sampling is that in the previous estimator most samples are "lost". Indeed most samples gives $X < 4$ where $h(x) = 0$. Instead we should change the sampling distribution so that most samples are greater than 4. The general principle is to use another density $g_Y(y)$ for another random variable Y and write

$$E[h(X)] = \int h(x) f_X(x) dx = \int \frac{h(x) f_X(x)}{g_Y(x)} g_Y(x) dx = E\left[\frac{h(Y) f_X(Y)}{g_Y(Y)}\right]$$

which gives us another estimator whose variance is

$$E\left[\left(\frac{h(Y) f_X(Y)}{g_Y(Y)}\right)^2\right] - E\left[\frac{h(Y) f_X(Y)}{g_Y(Y)}\right]^2 = E\left[\left(\frac{h(Y) f_X(Y)}{g_Y(Y)}\right)^2\right] - E\left[h(X)\right]^2$$

The potential gain is in the first term. For the example at hand we pick Y to be a shifted exponential with pdf $g_Y(x) = e^{x-4}$ for $x \geq 4$ which ensures that all samples are exceeding 4 and thus will contribute something. To see if we gain something let us estimate

$$E\left[\left(\frac{h(Y) f_X(Y)}{g_Y(Y)}\right)^2\right] = \int h(x)^2 \frac{f_X^2(x)}{g_Y(x)} dx = \int_4^\infty \frac{1}{2\pi} e^{-x^2/2} e^{-x^2/2+x-4} dx$$

$$= \int_4^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-(x-1)^2/2} e^{-7/2} dx \leq \frac{e^{-7/2} e^{-9/2}}{\sqrt{2\pi}} \int_4^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \frac{e^{-8}}{\sqrt{2\pi}}$$

so we gain a factor $\frac{e^{-8}}{\sqrt{2\pi}} = 0.0000133\ldots$ Impressive!

4.6 Slutsky theorem and applications

Slutsky is a very useful theorem with many applications. First we need a technical result (useful in its own right) which tells us that we only need to consider Lipschitz bounded functions to check for convergence in distribution.

Recall g is Lipschitz continuous if there exists a constant k such that $\{|g(x) - g(y)| \le k\|x-y\|$ for all x, y. Lipschitz functions are uniformle continuous and, functions which are differentiable with a bounded derivative $\sup_x |g'(x)| < \infty$ are Lipschitz continuous.

Theorem 4.4 (Portmanteau Theorem) The sequence X_n converges to X in distribution if and only if $\lim_{n\to\infty} E[f(X_n)] = E[f(X)]$ for all functions f which are *bounded and Lipschitz continuous*.

Proof. Suppose f is bounded with $\alpha = \sup_x |f(x)|$ so that $-\alpha \le f(x) \le \alpha$. Then consider the functions

$$h_k(x) = \inf_y \{f(y) + k\|x-y\|\} \quad \text{and} \quad H_k(x) = \sup_y \{f(y) - k\|x-y\|\}$$

- The functions h_k and H_k are bounded and increasing/decreasing sequences. We hav

$$-\alpha \le \inf_y f(y) \le \inf_y \{f(y) + k\|x-y\|\} = h_k(x) \le f(x) \le \sup_y \{f(y) -$$

and therefore we have $-\alpha \le h_k \le h_{k+1} \le f(x) \le H_{k+1} \le H_k \le \alpha$.

Theorem 4.5 (Slutsky's Theorem) If X_n converges to X in distribution and $|Y_n - X_n|$ converges to 0 in probability then Y_n converges to X in distribution.

Proof. Use Theorem 4.4 consider a bounded Lipschitz function f with $\sup_x |f(x)| \leq M$ and $|f(x) - f(y)| \leq K|x - y|$ for all x, y.

For any $\epsilon > 0$ we have

$$\begin{aligned} |E[f(X_n)] - E[f(Y_n)]| &\leq E[|f(X_n) - f(Y_n)|1_{\{|X_n - Y_n| < \epsilon\}}] + E[|f(X_n) - f(Y_n)|1_{\{|X_n - Y_n| \geq \epsilon\}}] \\ &\leq K\epsilon + 2MP(|X_n - Y_n| \geq \epsilon) \end{aligned}$$

Consequently

$$\begin{aligned} |E[f(Y_n)] - E[f(X)]| &\leq |E[f(Y_n)] - E[f(X_n)]| + |E[f(X_n)] - E[f(X)]| \\ &\leq K\epsilon + 2MP(|X_n - Y_n| \geq \epsilon) + |E[f(X_n)] - E[f(X)]| \end{aligned}$$

As $n \to \infty$ the right hand side converges to ϵ since the second term goes to 0 since $|Y_n - X_n|$ converges to 0 in probability and $|E[f(X_n)] - E[f(X)]|$ converges to 0 since X_n converges to X in distribution. Since ϵ is arbitary this concludes the proof. $\square$

In many applications the following result, also called Slutsky Theorem, is very useful.

Theorem 4.6 (Slutsky's Theorem) Suppose X_n converges to X in distribution and Y_n converges to c in probability. Then

1. $X_n + Y_n \to X + c$ in distribution.
2. $X_n Y_n \to cX$ in distribution.
3. $X_n / Y_n \to X / c$ in distribution (provided $c \neq 0$)

Proof.

- Consider the random variables (X_n, c). We show that (X_n, c) converges to (X, c) in distribution. Indeed for any bounded continous function $f(x, y)$ consider the function $g(x) = f(x, c)$. Since X_n converges to X in distribution then $E[g(X_n)] = E[f(X_n, c)]$ converges to $E[g(X)] = E[f(X, c)]$.
- Now $|(X_n, Y_n) - (X_n, c)| = |Y_n - c|$. So if Y_n converges to c in probability then (X_n, Y_n) converges to (X_n, c) in probability.
- Using Theorem 4.5 we conclude that (X_n, Y_n) converges to (X_n, c) in distribution.
- We now can use the continuity theorem for convergence in distribution (see Exercise 3.1) using the continuous functions $h(x_n, Y_n) = X_n + Y_n$ or $X_n Y_n$ or X_n / Y_n.

The central limit theorem states that $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ converges to a standard normal Z. If σ is not known can we replace it by the estimator for the variance estimator $V_n = \frac{1}{n}\sum_k (X_k - \frac{S_n}{n})^2$? The answer is yes, by applying Slutsky theorem

Theorem 4.7 (CLT using the empirical variance) Suppose the random variables X_i are IID RVs with $E[X_i] = \mu$ and $\mathrm{Var(X_i)} = \sigma^2$ for all i. Then

$$Y_n = \frac{S_n - n\mu}{\sqrt{nV_n}}$$

converges in distribution to a standard normal random variable Z.

Proof. This follows from Theorem 4.5 since $\sqrt{V_n}$ converges to σ in probability by the law of large numbers (and continuity theorem) and $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ converges to Z in distribution.

This is the standard way the CLT is used in statistical applications. For example……

4.7 The δ-method

Another nice application is the so-called δ-method which is some kind of non-linear version of the CLT. To this it is convenient to rewrite the CLT as

$$\sqrt{n}\left(\frac{S_n}{n} - \mu\right) \to Y \quad \text{in distribution}$$

where Y is normal with variance σ^2. By the continuity theorem if g is a continuous function then $g\left(\frac{S_n}{n}\right)$ converges to $g(\mu)$ almost surely and it is natural to ask whether we have a central limit theorem. The answer is yes provided g is differentiable and it is provided by the following theorem. We only show the 1d version.

Theorem 4.8 (δ-method) Suppose Y is normal with mean 0 and variance σ^2 and we have

$$\sqrt{n}\left(\frac{S_n}{n} - \mu\right) \to Y \quad \text{in distribution.}$$

Assume $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable with $g'(\mu) \neq 0$ then

$$\sqrt{n}\left(g\left(\frac{S_n}{n}\right) - g(\mu)\right) \to Y' \quad \text{in distribution,}$$

where Y' is normal with mean 0 and variance $\sigma^2 g'(\mu)^2$.

Proof. Taylor expansion around μ gives

$$g(\mu+h) = g(\mu) + g'(\mu)h + hr(h) \quad \text{with } \lim_{h\to 0} r(h) = 0$$

Applying this to $h = \frac{S_n}{n} - \mu$ we find

$$\sqrt{n}\left(g\left(\frac{S_n}{n}\right) - g(\mu)\right) = \sqrt{n}g'(\mu)\left(\frac{S_n}{n} - \mu\right) + \sqrt{n}\left(\frac{S_n}{n} - \mu\right) h\left(\frac{S_n}{n} - \mu\right) \tag{4.1}$$

Here comes Slutsky's Theorem in action. On one hand $\frac{S_n}{n} - \mu$ converges to 0 in probability and since h is continuous at 0 then $h\left(\frac{S_n}{n} - \mu\right)$ converges to 0 in probability as well by Theorem 1.4. Since $\sqrt{n}\left(\frac{S_n}{n} - \mu\right)$ converges to Y in probability Theorem 4.6 implies that the last term converges to 0 in distribution. But as we have seen in Theorem 3.1 convergence to a constant in distribution implies convergence in probability. We can now apply Theorem 4.5: the first term on the right hand side of Equation 4.1 converegs in distribution to a normal with variance $\sigma^2|g'(\mu)|^2$ and the second term converges in probability to 0, therefore the left hand side of Equation 4.1 converges in distribution a normal with variance $\sigma^2|g'(\mu)|^2$. $\square$.

Example

Suppose we are interested in the distribution of $Y_n = e^{S_n/n} = \left(\prod_{k=1}^{n} e^{X_k}\right)^n$, a type of model used in financial applications. Then we have $g'(\mu) = e^{\mu}$ and the delta method gives

$$\sqrt{n}\left(e^{\frac{S_n}{n}} - e^{\mu}\right) \to Y' \quad \text{in distribution}$$

with Y' is normal with zero mean and variance $\sigma^2 e^{2\mu}$.

Example

Suppose we have Bernoulli random variables $X_1, X_2, \cdots, X_n$ and we are interested in the *odds of success* that is the ratio $\frac{p}{1-p}$. (For gambling, often the odds of sucess are given instead of the probability of success). An estimator for the odds is given by

$$Y_n = \frac{\frac{S_n}{n}}{1 - \frac{S_n}{n}}$$

so we can apply the δ method with $g(x) = \frac{x}{1-x}$ so $g'(x) = \frac{1}{(1-x)^2}$. The delta methods tells us that

$$\sqrt{n}\left(Y_n - \frac{p}{1-p}\right) \to Y \quad \text{in distribution}$$

where Y is normal with variance $p(1-p)g'(p)^2 = \frac{p}{(1-p)^3}$.

4.8 Exercises

Exercise 4.1 Let $(X_j)_{j\geq 1}$ be independent, double exponential random variables with parameter 1 (that is, the common density is $\frac{1}{2}e^{-|x|}$ for $-\infty < x < \infty$. Show that

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{\sum_{j=1}^n X_j}{\sum_{j=1}^n X_j^2}\right) = Z \quad \text{in distribution}$$

where Z is normal with mean 0 and variance $\frac{1}{2}$.

Exercise 4.2 Suppose X_i are IID random variables with $E[X_i] = 1$ and $V[X_i] = \sigma^2$. Show that

$$\frac{2}{\sigma}\left(\sqrt{S_n} - \sqrt{n}\right)$$

converge, in distribution, to a standard nornal random variables.
Hint: $a^2 - b^2 = (a+b)(a-b)$.

Exercise 4.3 Show that

$$\lim_{n\to\infty} e^{-n}\left(\sum_{k=0}^{n} \frac{n^k}{k!}\right) = \frac{1}{2}.$$

Hint: Let (X_j) be i.i.d. Poisson random variables with parameter $\lambda = 1$. Let $S_n = \sum_{j=1}^{n} X_j$ and apply the Central limit theorem