Chapter 3

Concentration and Limits

3.1 Concentration inequalities

In probability theory, concentration inequalities provide bounds on how a random variable deviates from some value (typically, its expected value).

Theorem 3.1 (**Markov's inequality**). *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple and let* $X \in L^1(\mathbb{P})$ *be a non-negative* $\mathbb{R}$*-valued RV. Then,*

$$\mathbb{P}(X \geq \epsilon) \leq \frac{\mathbb{E}(X)}{\epsilon}, \qquad \text{for any} \quad \epsilon > 0 \ . \tag{3.1}$$

Proof. Let $A_\epsilon = [\epsilon, \infty)$ then by Lemma 2.8

$$\mathbb{1}_{A_\epsilon}(x) + \mathbb{1}_{A_\epsilon^c}(x) = 1$$

as such we can write

$$X = X\mathbb{1}_{A_\epsilon}(X) + X\mathbb{1}_{A_\epsilon^c}(X) \geq X\mathbb{1}_{A_\epsilon}(X) \geq \epsilon\mathbb{1}_{A_\epsilon}(X).$$

Now the inequalities are preserved when taking the expectation of both sides (Theorem 2.45), and we get

$$\mathbb{E}[X] \geq \epsilon\,\mathbb{E}[\mathbb{1}_{A_\epsilon}(X)] = \epsilon\,\mathbb{P}(X \in A_\epsilon) = \epsilon\,\mathbb{P}(X \geq \epsilon)$$

□

Let us look at some immediate consequences of Markov's inequality.

Proposition 3.2 (**Chebychev's inequality**). *For* **any** *RV X and any $\epsilon > 0$,*

$$\mathbb{P}(|X| > \epsilon) \leq \frac{\mathbb{E}(|X|)}{\epsilon}$$
$$\mathbb{P}(|X| > \epsilon) = \mathbb{P}(X^2 \geq \epsilon^2) \leq \frac{\mathbb{E}(X^2)}{\epsilon^2}$$
$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \epsilon) = \mathbb{P}((X - \mathbb{E}(X))^2 \geq \epsilon^2) \leq \frac{\mathbb{E}(X - \mathbb{E}(X))^2}{\epsilon^2} = \frac{\mathbb{V}(X)}{\epsilon^2}$$

In the above we interpret the expectations as ∞ if they dont exist. However if we want finite expressions then we would need $X \in L^1(\mathbb{P})$ for the first inequality and $X \in L^2(\mathbb{P})$ for the second and third inequality.

Proof. All three forms of Chebychev's inequality are mere corollaries (careful reapplications) of Markov's inequality. □

Definition 3.3. *We say that the sequence $X_1, X_2, \ldots$ of $\mathbb{R}$-valued RVs is an independent and identically distributed (i.i.d.) sequence of $\mathbb{R}$-valued random variables with distribution F, if for any $n \in \mathbb{N}$ we have that $X_1, \ldots, X_n \sim F$ and that $Z = (X_1, \ldots, X_n)$ is an $\mathbb{R}^n$-valued RV with distribution function*

$$F_Z(x_1, \ldots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n) = \prod_{i=1}^{n} F(x_i).$$

We usually denote this with $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F$.

Remark 3.4. *This can quite easily be extended to sequences i.i.d. random vectors with the trivial modifications.*

One of the most fundamental aspects of statistics is the concept of "concentration of measure". As we saw in Chapter 1 one can motivate the concept of probability as a long-term relative frequency. That is if we toss a fair coin we expect that $N(\mathtt{H}, n) \to \frac{1}{2}$ as $n \to \infty$, however for any finite number of tosses there is a probability that this can deviate from $\frac{1}{2}$ (we could for instance observe the unlikely event that we get all heads). We do however expect that the probability of a large deviation to become smaller as we observe more tosses, this is the phenomenon of concentration of measure. We will begin with a "helper lemma" and then move on to prove a fundamental concentration inequality called Hoeffdings inequality.

Lemma 3.5 (**Hoeffdings lemma**). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and suppose that X is a $\mathbb{R}$-valued RV such that $\mathbb{P}(X \in [a, b]) = 1$ for $a < b$.*

Then, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$$

Proof. First we need to make sure that the left hand side is OK. First we need to make sure that $X \in L^1(\mathbb{P})$ and then we need to make sure that $e^{\lambda X} \in L^1(\mathbb{P})$. However the boundedness condition $a \leq X \leq b$ immediately implies this, since

$$a = \mathbb{E}[a] \leq \mathbb{E}[X] \leq \mathbb{E}[b] = b$$

and the same thing holds for the exponential, i.e.

$$e^{\lambda a} = \mathbb{E}[e^{\lambda a}] \leq \mathbb{E}[e^{\lambda X}] \leq \mathbb{E}[e^{\lambda a}] \leq e^{\lambda a}.$$

Now since $e^{\lambda x}$ is convex we have

$$e^{\lambda x} \leq \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b}$$

for all $a \leq x \leq b$. Hence if we let $Y = X - \mathbb{E}[X]$ we get

$$\mathbb{E}[e^{\lambda Y}] \leq \frac{b-\mathbb{E}[Y]}{b-a}e^{\lambda a} + \frac{\mathbb{E}[Y]-a}{b-a}e^{\lambda b} = \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b}.$$

Now let $h = \lambda(b-a)$, $p = \frac{-a}{b-a}$ and $L(h) = -hp + \ln(1-p+pe^h)$, then

$$\frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} = e^{L(h)}. \tag{3.2}$$

Let us bound L from above, we will do that using basic calculus, first note that

$$L(0) = \ln(1) = 0$$

and

$$L'(h) = -p + \frac{pe^h}{1-p+pe^h} \implies L'(0) = 0.$$

Let us now consider the second derivative,

$$L''(h) = \frac{pe^h}{1-p+pe^h} - \frac{(pe^h)^2}{(1-p+pe^h)^2}$$

this is of the form $y - y^2$ which cannot be larger than $1/4$, as such we get using Taylors theorem, that

$$L(h) \leq \frac{h^2}{8} = \frac{\lambda^2(b-a)^2}{8}$$

from (3.2) we now complete the lemma. □

Theorem 3.6 (Hoeffdings inequality (simple case))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F$ be $\mathbb{R}$-valued RVs such that $\mathbb{P}(X_i \in [a,b]) = 1$, then for any $\epsilon > 0$ we get for $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,*

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Proof. Let $S_n = \sum_{i=1}^n X_i$. Let $s, t > 0$ be positive numbers to be chosen, then using Theorem 3.1 we get

$$\begin{aligned}
\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) &= \mathbb{P}(e^{s(S_n - \mathbb{E}[S_n])} \geq e^{st}) && (3.3)\\
&\leq e^{-st}\,\mathbb{E}(e^{s(S_n - \mathbb{E}[S_n])}) && (3.4)\\
&= e^{-st}\prod_{i=1}^n \mathbb{E}(e^{s(X_i - \mathbb{E}[X_i])}) && (3.5)
\end{aligned}$$

where in the last step we used the independence of $X_1, \ldots$ together with Theorem 2.45. Now using Lemma 3.5 with $\lambda = s$ for each term in the product, we get

$$e^{-st}\prod_{i=1}^n \mathbb{E}(e^{s(X_i - \mathbb{E}[X_i])}) \leq e^{-st} e^{s^2(b-a)^2 n/8} \tag{3.6}$$

Notice now, that the value s was arbitrarily chosen and we can choose it to make the right hand side as small as possible. That is we want to minimize

$$h(s) = s^2\frac{n(b-a)^2}{8} - st. \tag{3.7}$$

This function is minimized at $s^* = \frac{4t}{n(b-a)^2}$, plugging that in we get

$$h(s^*) = s^2\frac{n(b-a)^2}{8} - st = -\frac{2t^2}{n(b-a)^2}. \tag{3.8}$$

Assembling (3.3) and (3.6)–(3.8) we get

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq e^{-\frac{2t^2}{n(b-a)^2}}.$$

Replacing $t = n\epsilon$ we get

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}},$$

which proves the theorem. □

Corollary 3.7. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F$ $\mathbb{R}$-valued RVs such that $\mathbb{P}(X_i \in [a,b]) = 1$, then for any $\epsilon > 0$ we get for $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,*

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}},$$

furthermore

$$\mathbb{P}(\left|\overline{X}_n - \mathbb{E}[\overline{X}_n]\right| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}},$$

Proof. Exercise!! Hint: In Theorem 3.6 we did not assume anything about the sign of X_i. □

Let us look at an application of this, namely that of constructing confidence regions:

Lemma 3.8. *[Estimating p in Bernoulli] Let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} \text{Bernoulli}(p)$. Then for $\alpha \in (0,1)$ we have for $\delta = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)}$*

$$\mathbb{P}(\overline{X}_n - \delta \leq p \leq \overline{X}_n + \delta) \geq 1 - \alpha.$$

Remark 3.9. *In the above, if we fix $\alpha = 0.05$ we get $\delta \approx \frac{1.36}{\sqrt{n}}$.*

Proof. We wish to apply Corollary 3.7. Note that $a = 0, b = 1$ in the Bernoulli case, hence for a fix α and fix n we need to solve

$$\alpha = 2e^{-2n\epsilon^2}$$

with respect to ϵ. We get

$$\epsilon = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)}$$

as such we get from Corollary 3.7 that

$$\begin{aligned}\mathbb{P}\left(\overline{X}_n - \frac{1}{\sqrt{n}}\sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)} \leq p \leq \overline{X}_n + \frac{1}{\sqrt{n}}\sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)}\right) \\ = 1 - \mathbb{P}\left(\left|\overline{X}_n - p\right| > \frac{1}{\sqrt{n}}\sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)}\right) \geq 1 - \alpha.\end{aligned}$$

□

Remark 3.10. *Computing multiple intervals for random variables which are not necessarily independent is quite easy using the union bound. Actually, we did it above when going from the one-sided to the two sided inequality, see Corollary 3.7.*

Suppose we have m sequences of random variables $Z_1 = (X_{11}, X_{12}, \dots, X_{1n})$, $\dots, Z_m = (X_{m1}, \dots, X_{mn})$. Where for each i the sequence Z_i is i.i.d. but Z_i and Z_j are not necessarily independent (they could all be the same for instance). Assume that each one of them satisfies

$$\mathbb{P}(|\frac{1}{n}\sum_{j=1}^{n} X_{ij} - \mathbb{E}[X_{1j}]| \geq \epsilon) \leq C_i$$

for every i and for some number C_i. Then

$$\mathbb{P}(|\frac{1}{n}\sum_{j=1}^{n} X_{ij} - \mathbb{E}[X_{1j}]| \geq \epsilon \text{ for some } i) \leq \sum_{i=1}^{m} C_i$$

the complement of this is

$$\mathbb{P}(|\frac{1}{n}\sum_{j=1}^{n} X_{ij} - \mathbb{E}[X_{1j}]| < \epsilon \text{ for all } i) \geq 1 - \sum_{i=1}^{m} C_i.$$

So for instance if all of where Bernoulli(p_i) *then from the above and Lemma 3.8 we get that*

$$\mathbb{P}(\frac{1}{n}\sum_{j=1}^{n} X_{ij} - \delta \leq p_i \leq \frac{1}{n}\sum_{j=1}^{n} X_{ij} + \delta \text{ for all } i) \geq 1 - \alpha.$$

where $\delta = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{2}\ln\left(\frac{2m}{\alpha}\right)}$. This means that m appears in the logarithm, and the increase of δ w.r.t. m is fairly slow. We will use this fact later when we produce multiple intervals for different metrics. This is equivalent to **Bonferroni correction** *in multiple testing.*

So the Hoeffding inequality is actually quite useful (extremely), but the restriction that the random variables are bounded is a heavy restriction. However, if we look at the proof of Theorem 3.6 we note that everything follows from Lemma 3.5, so if the estimate in Lemma 3.5 holds, then so should Theorem 3.6. With this in mind, let us define

Definition 3.11. *A $\mathbb{R}$ valued random variable X is said to be* **sub-Gaussian** *with parameter λ if*

$$\mathbb{E}[e^{s(X-\mathbb{E}[X])}] \leq e^{\frac{s^2\lambda^2}{2}}, \quad \text{for all } s.$$

and a local version

Definition 3.12. *A $\mathbb{R}$ valued random variable X is said to be* **sub-exponential** *with parameter λ if*

$$\mathbb{E}\left[e^{s(X-\mathbb{E}[X])}\right] \leq e^{\frac{s^2\lambda^2}{2}}, \qquad \text{for all } |s| \leq \frac{1}{\lambda}.$$

Both of these can be used in the proof of Theorem 3.6, however in the case of sub-exponential we have to take care of the restriction on s which actually yields a weaker bound.

Theorem 3.13. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F$ be $\mathbb{R}$-valued sub-Gaussian RVs with parameter σ then for any $\epsilon > 0$ we get for $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,*

$$\mathbb{P}(\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right] \geq \epsilon) \leq e^{-\frac{n\epsilon^2}{2\sigma^2}}.$$

For the sub-exponential case we get a weaker bound for the tails. The reason for this is the fact that the bound on $\mathbb{E}^{sX}$ only holds for small s, the resulting estimate thus differentiates between small and big ϵ. We can see in the estimate below that for large ϵ the tail is exponential, i.e. $e^{-\epsilon}$, this in one of the reasons for the name sub-exponential.

Theorem 3.14. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F$ be $\mathbb{R}$-valued sub-exponential RVs with parameter λ then for any $\epsilon > 0$ we get for $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,*

$$\mathbb{P}(\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right] \geq \epsilon) \leq e^{-\frac{\epsilon^2 n}{2\lambda^2}} \vee e^{-\frac{(\epsilon+1)n}{2\lambda}}.$$

Proof. Proceeding as in the proof of (3.3) we get

$$\mathbb{P}(S_n - \mathbb{E}\left[S_n\right] \geq t) = e^{-st}\prod_{i=1}^n \mathbb{E}(e^{s(X_i - \mathbb{E}[X_i])}) \tag{3.9}$$

Consider now $s \leq \frac{1}{\lambda}$ and apply Definition 3.12 for each term in the product, we get

$$e^{-st}\prod_{i=1}^n \mathbb{E}(e^{s(X_i - \mathbb{E}[X_i])}) \leq e^{-st}e^{\frac{s^2\lambda^2 n}{2}} \tag{3.10}$$

If we proceed as in Theorem 3.6 we consider

$$h(s) = \frac{s^2\lambda^2 n}{2} - st$$

and note again that the function is minimized at $s^* = \frac{t}{n\lambda^2}$. If now $s^* \leq \frac{1}{\lambda}$ we get

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq e^{-\frac{t^2}{2n\lambda^2}} \tag{3.11}$$

However if $s^* > 1/\lambda$, that is if $t > n\lambda$ we get can only take

$$h(1/\lambda) = \frac{n}{2} - \frac{1}{\lambda}t < -\frac{n}{2} - \frac{t}{2\lambda}$$

that is,

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq e^{-\frac{n}{2} - \frac{t}{2\lambda}}. \tag{3.12}$$

Assembling (3.11) and (3.12) we get

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq e^{-\frac{t^2}{2n\lambda^2}} \vee e^{-\frac{n}{2} - \frac{t}{2\lambda}}, \tag{3.13}$$

where $a \vee b = \max(a, b)$. Now again replacing $t = n\epsilon$ we get

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon) \leq e^{-\frac{\epsilon^2 n}{2\lambda^2}} \vee e^{-\frac{(\epsilon+1)n}{2\lambda}}.$$

□

Lemma 3.15. *The following properties hold*

1. *Let X be a sub-Gaussian RV with parameter λ, then αX is sub-Gaussian with parameter $|\alpha|\lambda$.*
2. *Let X be a sub-exponential RV with parameter λ, then αX is sub-exponential with parameter $|\alpha|\lambda$.*
3. *A sub-Gaussian RV X with parameter λ is sub-Exponential with parameter λ.*
4. *A bounded RV X, i.e. $\mathbb{P}(X \in [a, b]) = 1$, then X is sub-Gaussian with parameter $(b-a)/2$. Specifically a Bernoulli RV is sub-Gaussian with parameter $1/2$.*
5. *If X is sub-Gaussian with parameter λ then $Z = X^2$ is sub-exponential with parameter $8\lambda^2$.*
6. *if X, Y are independent and sub-Gaussian with parameter σ_1, σ_2, then $X + Y$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$.*

Proof. Let $Z = X^2 - \mathbb{E}[X^2]$, use the power series representation of the exponential (we have not really gone through the theory for this one, but this is OK by the dominated convergence theorem which is outside the scope of this course)

$$\mathbb{E}[e^{sZ}] = 1 + \sum_{k=2}^{\infty} \frac{s^k \mathbb{E}[Z^k]}{k!}$$

First use the elementary fact that $(a+b)^k \leq 2^{k-1}(a^k + b^k)$ (for $k > 1$ since the power function is convex), we get

$$\mathbb{E}[Z^k] = \mathbb{E}[(X^2 - \mathbb{E}[X^2])^k] \leq 2^{k-1}(\mathbb{E}[X^{2k}] + (\mathbb{E}[X^2])^k)$$

Now by Hölders inequality Theorem 2.49

$$\mathbb{E}[X^2]^k \leq \mathbb{E}[X^{2k}]$$

Thus we get

$$\mathbb{E}[e^{sZ}] \leq 1 + \sum_{k=2}^{\infty} \frac{s^k 2^k \mathbb{E}[X^{2k}]}{k!}$$

Now, note that Theorem 3.13 gives us bounds for the moments of X above, i.e. using the fact that (for the first inequality see Exercise 5.18) we get

$$\begin{aligned}\mathbb{E}[X^k] &= \int_0^\infty \mathbb{P}(|X|^k > t)dt \leq 2\int_0^\infty e^{-\frac{t^{2/k}}{2\lambda^2}} dt = (2\lambda^2)^{k/2} k \int_0^\infty e^{-u} u^{k/2-1} du \\ &= (2\lambda^2)^{k/2} k\Gamma(k/2)\end{aligned}$$

Going back to our problem we get (using that $k\Gamma(k) = k!$)

$$\begin{aligned}\mathbb{E}[e^{sZ}] &\leq 1 + \sum_{k=2}^{\infty} \frac{s^k 2^k (2\lambda^2)^k 2k!}{k!} \\ &\leq 1 + 2\sum_{k=2}^{\infty} (4s\lambda^2)^k \\ &\leq 1 + 2(4s\lambda^2)^2 \sum_{k=0}^{\infty} (4s\lambda^2)^k \\ &\leq 1 + 64s^2\lambda^4 \leq e^{32s^2\lambda^4}\end{aligned}$$

the last sum is a geometric sum and is less than 2 if $8s\lambda^2 < 1$, i.e. $s < \frac{1}{8\lambda^2}$. Thus we see that X^2 is sub-exponential with parameter $8\lambda^2$. □

Distribution	sub-exponential	sub-Gaussian
Gaussian	Yes	Yes
Bernoulli	Yes	Yes
Uniform	Yes	Yes
Bounded	Yes	Yes
Exponential	Yes	No
χ^2	Yes	No
Weibull ($k \geq 1$)	Yes	No
Laplace	Yes	No
Pareto	No	No
Lognormal	No	No

Figure 3.1: Examples of distributions that are sub-exponential and sub-Gaussian

The question is now, what distributions are sub-Gaussian and which are sub-exponential? See Fig. 3.1. We will be using these concentration inequalities in the course to prove that the algorithms we are interested in is actually doing what we want with high probability.

Exercise 3.16. *For the Poisson distribution, we have*

$$\mathbb{E}\left[e^{sX}\right] = e^{\lambda(e^s - 1)}$$

is this sub-Gaussian, sub-exponential or neither?

3.1.1 Random variables that are not exponentially integrable*

Both the sub-Gaussian and sub-exponential rely on the fact that $\mathbb{E}\left[e^{sX}\right] < \infty$, if we rewrite this for a continuous RV we get

$$\int_{-\infty}^{\infty} e^{sx} f_X(x) dx < \infty$$

hence we need either that f_X has finite support or we need that it decays exponentially at infinity. This is why the sub-exponential for instance has the restriction on the size of s, as in that case f_X behaves like $e^{-\frac{1}{\lambda}|x|}$, that is

$$\int_{-\infty}^{\infty} e^{sx} e^{-\frac{1}{\lambda}|x|} dx < \infty, \quad \text{if and only if } s < \frac{1}{\lambda}.$$

One could hope that there is still some concentration of measure in this case. Our first observation is that the exponential integrability implies that all moments exists, i.e. $X \in L^p(\mathbb{P})$ for every $1 \leq p < \infty$. However, what if we only have $X \in L^p(\mathbb{P})$ for some $1 \leq p < \infty$, what can we say then?

Theorem 3.17. *Lets say that* $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F$ *be* $\mathbb{R}$ *valued RVs. Suppose that* $X_i \in L^{2s}(\mathbb{P})$ *and*

$$|\,\mathbb{E}\,[(X_i - \mathbb{E}\,[X_i])^r]| \leq \sigma^2 r!, \text{ for } r = 2, 3, \ldots, 2s$$

for a positive integer $s > 1$. *Then if* $\epsilon \in [0, \sqrt{2}n\sigma^2]$ *and* $s \leq n\sigma^2$ *we have*

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}\,[\overline{X}_n]| \geq \epsilon) \leq \left(\frac{4s\sigma^2}{\epsilon^2 n}\right)^{s/2}.$$

If further, $s \geq \epsilon^2/(2n\sigma^2)$ *then we also have*

$$\mathbb{P}(|\overline{X_n}| \geq \epsilon) \leq 3e^{-\frac{\epsilon^2 n}{12\sigma^2}}.$$

Proof. The proof uses Theorem 3.1 as Theorem 3.6, however now we cannot use the exponential, instead we use powers. The power of $\mathbb{E}\,[|\sum_i X_i|^k]$ has to computed, this can be done by carefully checking the combinatorics of the terms and using the independence assumption. See the FDS book Theorem 12.5. □

3.2 Convergence of Random Variables

This important topic is concerned with the limiting behavior of sequences of RVs. We want to understand what it means for a sequence of random variables $\{X_n\}_{n=1}^{\infty} := X_1, X_2, \ldots$ to converge to another random variable X, when all RVs are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

$$\{X_i\}_{i=1}^{n} := X_1, X_2, X_3, \ldots X_{n-1}, X_n \qquad \text{as } n \to \infty\ .$$

From a statistical or decision-making viewpoint, $n \to \infty$ is associated with the amount of data or information $\to \infty$. More abstractly, we are interested in what happens to the limiting RV $X := \lim_{n\to\infty} X_n$ when given the DFs $F_n(x)$ for each X_n.

We need different notions of convergence to characterize such a behavior: two simplest behaviors are that the sequence eventually takes a constant value θ, i.e. X_n approaches $X \sim$ Point Mass(θ) RV, or that values in the sequence continue to change but can be described by an unchanging probability distribution, i.e., X_n approaches $X \sim F(x)$. See https://en.wikipedia.org/wiki/Convergence_of_random_variables.

Definition 3.18 (**Convergence in Distribution** (or Weakly, or in Law))**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple and let* $X_1, X_2, \ldots,$ *be a sequence of RVs*

and let X be another RV. Let F_n denote the DF of X_n and F denote the DF of X. The we say that X_n converges to X in distribution, and write:

$$X_n \rightsquigarrow X$$

if for any real number t at which F is continuous,

$$\lim_{n\to\infty} F_n(t) = F(t).$$

The above limit, by (2.2) in our Definition 2.2 of a DF, can be equivalently expressed as follows:

$$\lim_{n\to\infty} \mathbb{P}(\ \{\omega : X_n(\omega) \le t\}\ ) = \mathbb{P}(\ \{\omega : X(\omega) \le t\}\ ).$$

Convergence in distribution does not in general imply that the sequence of corresponsing probability density functions will also converge. Consider for example RV X_n with density $\mathbb{1}_{(0,1)}(x)(1 - \cos(2\pi n x))$. These RVs converge in distribution to $X \sim \text{Uniform}(0, 1)$, but their densities (PDFs) do not converge at all as evident in Fig. 3.2.

The other way around is however true:

Lemma 3.19. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots,$ be a sequence of RVs and let X be another RV. Let f_n denote the PDF of X_n and f denote the PDF/PMF of X. If*

$$f_n(x) \to f(x), \quad \forall x \in \mathbb{R},$$

then

$$X_n \rightsquigarrow X.$$

From Lemma 3.19 we see that for a discrete sequence of RVs X_n to converge in distribution to another discrete RV X taking values in $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$, it is sufficient to show that $\lim_{n\to\infty} \mathbb{P}(X_n = x) = \mathbb{P}(X = x)$ for each $x \in \mathbb{Z}_+$. We will use this fact to prove why we can approximate Binomial RVs by a Poisson under some limiting conditions.

Theorem 3.20. *Let $X_n \sim \text{Binomial}(n, \lambda/n)$ for $n = 1, \ldots$ and let $Y \sim \text{Poisson}(\lambda)$, then*

$$X_n \rightsquigarrow Y.$$

Proof. Let $X_n \sim \text{Binomial}(n, \theta = \lambda/n)$ and $Y \sim \text{Poisson}(\lambda)$ for a fixed λ. We need to show that

$$\lim_{n\to\infty} \mathbb{P}(X_n = x) = \mathbb{P}(Y = x) = e^{-\lambda}\lambda^x/x!$$

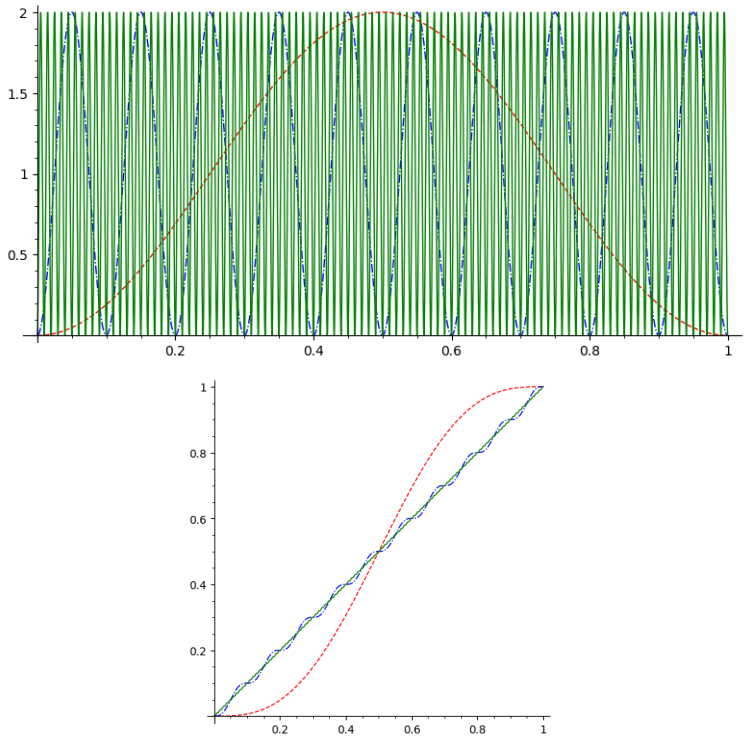

Figure 3.2: PDF $f_{X_n}(x) := \mathbb{1}_{(0,1)}(x)(1 - \cos(2\pi n x))$ of the RV X_n [the left sub-figure] and its DF $F_n(x) := \int_{-\infty}^{x} \mathbb{1}_{(0,1)}(v)(1 - \cos(2\pi n v))dv$ [the right sub-figure], for $n = 1$ [red '- -'], $n = 10$ [blue '-.'], and $n = 100$ [green '-'], respectively. One can see clear convergence of the DFs F_n to $\mathbb{1}_{(0,1)}(x)x$, the DF of the Uniform$(0, 1)$ RV, while the corresponding PDFs $f_n(x)$ keep oscillating wildly with n across $[0, 2]$ about $\mathbb{1}_{(0,1)}(x)$, the PDF of the Uniform$(0, 1)$ RV X. Thus giving a counter-example to the claim that convergence in DFs does not imply convergence in PDFs.

for any $x \in \{0, 1, 2, 3, \ldots, n\}$.

$$\mathbb{P}(X_n = x) \quad = \quad \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

First note that

$$\binom{n}{x} \left(\frac{\lambda}{n}\right)^x = \frac{n!}{(n-x)!n^x} \frac{\lambda^x}{x!}$$

By Stirlings formula this now converges to $\frac{\lambda^x}{x!}$. The last term

$$\left(1 - \frac{\lambda}{n}\right)^{n-x} = \left(1 - \frac{\lambda}{n}\right)^{n} \left(1 - \frac{\lambda}{n}\right)^{-x}$$

Is the product of two terms, where the first tends to $e^{-\lambda}$ and the second tends 1. We thus get

$$\lim_{n\to\infty} \mathbb{P}(X_n = x) = \mathbb{P}(Y = x)$$

which according to Lemma 3.19 gives $X_n \rightsquigarrow Y$. □

The second notion of convergence of RVs is convergence in probability.

Definition 3.21 (**Convergence in Probability**). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots,$ be a sequence of RVs and let X be another RV. We say that X_n converges to X in probability, and write:*

$$X_n \xrightarrow{\mathbb{P}} X$$

if for every real number $\epsilon > 0$,

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

Once again, the above limit, by (2.1) in our Definition 2.1 of a RV, can be equivalently expressed as follows:

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

Definition 3.22 (**Convergence Almost Surely** (or with Probability 1)). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots,$ be a sequence of RVs and let X be another RV. We say that X_n converges to X almost surely (or with probability 1/strongly) if*

$$\mathbb{P}\left(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right) = 1,$$

denoted as

$$X_n \overset{a.s.}{\to} X$$

This means that the values of X_n approach the value of X, in the sense that events for which X_n does not converge to X have probability 0,

$$\mathbb{P}\left(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\}\right) = 0,$$

Another notion which is quite useful is the mean-square convergence (or just L^2 convergence) which is a special case of

Definition 3.23 (**Convergence in L^p**). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots, \in L^p(\mathbb{P})$ be a sequence of RVs and let $X \in L^p(\mathbb{P})$ be another RV. We say that X_n converges to X in $L^p(\mathbb{P})$ if*

$$\|X_n - X\|_{L^p(\mathbb{P})} \to 0.$$

Recall the definition of the $L^p(\mathbb{P})$ norm, Section 2.6.4, hence the above is equivalent to

$$\mathbb{E}\left[|X_n - X|^p\right] \to 0.$$

Other notions of convergence are termed sure convergence or pointwise convergence, such as convergence in mean. But the above types of convergence are elementary.

3.2.1 Properties of Convergence of RVs**

We will merely state some properties (without proofs that are hyper-linked for the curious student as they are advanced for this course) and relations between the three notions of convergence with some examples to better appreciate the subtleties among them. Just remember that subtle implication relations exist between the notions.

Theorem 3.24. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots,$ be a sequence of RVs and let X be another RV. The following are equivalent: Let $f : \mathbb{R} \to \mathbb{R}$ be a function*

- $\mathbb{E}\left[f(X_n)\right] \to \mathbb{E}\left[f(X)\right]$ *for all bounded, countinuous f.*
- $\mathbb{E}\left[f(X_n)\right] \to \mathbb{E}\left[f(X)\right]$ *for all bounded, Lipschitz countinuous f.*
- $X_n \rightsquigarrow X$.
- *For any "good enough" set $A \subset \mathbb{R}$, $\mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A)$.*

Remark 3.25. *For those interested, good enough in the above means that $\mathbb{P}(X \in \partial A) = 0$. This is equivalent to the convergence happening only on the points of continuity of the distribution function, as in Definition 3.18.*

- Convergence almost surely implies convergence in probability[1]

$$\boxed{X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{\mathbb{P}} X \ .}$$

- By the Borel-Cantelli Lemma [2], convergence in probability does not imply almost sure convergence in the discrete case [3]

- Convergence in probability implies convergence in distribution [4]

$$\boxed{X_n \xrightarrow{\mathbb{P}} X \implies X_n \rightsquigarrow X \ .}$$

- Convergence in distribution to a constant θ implies convergence in probability to θ: [5]

$$X_n \rightsquigarrow \text{Point Mass}(\theta) \implies X_n \xrightarrow{\mathbb{P}} \text{Point Mass}(\theta) \ .$$

- Convergence in L^p for $1 \leq q \leq p < \infty$ implies convergence in L^q. (Follows from Theorem 2.49)

- Convergence in L^p for $1 \leq p < \infty$ implies convergence in probability. Follows from Theorem 3.1.

- In general, convergence in distribution does not imply convergence in probability.

3.3 Law of Large Numbers

Theorem 3.26. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots, \in L^2(\mathbb{P})$ be a sequence of i.i.d. RVs with $\mathbb{E}[X_i] = \mu$. Then*

$$\overline{X}_n \xrightarrow{\mathbb{P}} \mu.$$

Proof. We need to prove that for a fixed $\epsilon > 0$ that

$$\lim_{n \to \infty} \mathbb{P}(|X_n - \mu| > \epsilon) \to 0.$$

[1] https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_almost_surely_implies_convergence_in_probability

[2] https://en.wikipedia.org/wiki/Borel%E2%80%93Cantelli_lemma

[3] https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_in_probability_does_not_imply_almost_sure_convergence_in_the_discrete_case

[4] https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_in_probability_implies_convergence_in_distribution

[5] https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_in_distribution_to_a_constant_implies_convergence_in_probability

But note that from Theorem 3.1 we have

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\mathbb{E}\left[|\overline{X}_n - \mu|^2\right]}{\epsilon^2} = \frac{1}{n}\frac{\mathbb{E}\left[|X_1 - \mu|^2\right]}{\epsilon^2}. \tag{3.14}$$

The last step used

$$\begin{aligned}
\mathbb{E}\left[|\overline{X}_n - \mu|^2\right] &= \mathbb{E}\left[\frac{1}{n^2}|\sum_i X_i - n\mu|^2\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j}(X_i - \mu)(X_j - \mu)\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i}(X_i - \mu)^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[|X_1 - \mu|^2\right],
\end{aligned}$$

where in the second to last step in the above we used the independence assumption as $\mathbb{E}\left[(X_i - \mu)(X_j - \mu)\right] = \mathbb{E}\left[(X_i - \mu)\right]\mathbb{E}\left[(X_j - \mu)\right] = 0$ if $i \neq j$. (This is the famous "variance of the sum is the sum of the variance" for independent random variables).

Now (3.14) completes the proof. □

Also a stronger result is true, but we will not give the proof

Theorem 3.27 (Strong law of large numbers)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots, \in L^2(\mathbb{P})$ be a sequence of i.i.d. RVs with $\mathbb{E}\left[X_i\right] = \mu$. Then*

$$\overline{X}_n \overset{a.s.}{\to} \mu.$$

3.4 Central Limit Theorem

What if we scale the sum of X_i's by $\sqrt{n}$ instead of n?

Theorem 3.28. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots, \in L^2(\mathbb{P})$ be a sequence of i.i.d. RVs with $\mathbb{E}\left[X_i\right] = \mu$ and $\mathbb{V}\left[X_i\right] = \sigma^2$. Then if we denote*

$$Z_n := \frac{\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]}{\sqrt{\mathbb{V}\left[\overline{X}_n\right]}}$$

we get

$$Z_n \rightsquigarrow Z$$

where $Z \sim N(0, 1)$.

Rewriting this in $L^p(\mathbb{P})$ space notation we get for $Y_n = X_n - \mathbb{E}[X_n]$ that

$$Z_n := \frac{\overline{Y}_n}{\|\overline{Y_n}\|_{L^2(\mathbb{P})}}$$

that is, we normalized Y_n to have L^2 norm 1, $\|Z_n\|_{L^2(\mathbb{P})} = 1$. The central limit theorem tells us that there is a $Z \in L^2(\mathbb{P})$ that is the distributional limit of Z_n. (Warning we cannot expect stronger convergence, this is due to non-compactness of the unit ball in L^2).

Proof. We will skip the proof of the CLT, as it is not particularly useful for us. □