# APPRENTICE PROGRAM PROBABILITY LECTURE 005: TYPES OF CONVERGENCE, AND A FEW ADDITIONAL THEOREMS

ETHAN NAEGELE

Abstract. These notes were intended for participants in the 2024 REU Apprentice Program at the University of Chicago who learned probability theory under my mentorship. The notes cover a few theorems relating to the different types of convergence and appear commonly in probability. We will conclude by taking a quick look at applications in statistics.

## 1. Introduction

The main aim of these notes are to revisit the different types of convergence and explain how they relate to one another. To recap, in real analysis, we have a good, single notion of convergence for, say, sequences in metric spaces. But if the elements of your sequence and your limit are random, it is not really sensible to impose that we should have $X_n$ close to some $X$; we have to somehow include probability in our definition of convergence. We can say, for example, that as $n$ increases, the probability of $X_n$ and $X$ being close will increase to 1. Or we may only care about the distributions of the variables: maybe those become close. These concepts are different, and we therefore see that there is no single notion of convergence for random variables. We will introduce the different forms of convergence and provide a detailed discussion of how these forms relate to each other.

The goal of these notes is so that you can have all these statements and proofs in one place as a reference since these results very commonly appear in probability and statistics, either explicitly or implicitly.

## 2. Weak Convergence

We start out with the weakest form of convergence, aptly called weak convergence, which actually has a surprisingly deep theory behind it, ultimately culminating in several important results such as the central limit theorem. Weak convergence of a sequence $X_n$ is related to the convergence of CDFs of the $X_n$'s in the real valued case, and as such, is called convergence in distribution for this case. Since we danced around the topic during our discussion of the central limit theorem, we'll take the time now to define it properly.

As a note: when discussing weak convergence, we will routinely use the pushforward measure discussed in the fourth set of notes. In the general setting, we have random variables $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathcal{X}, \mathcal{B}_{\mathcal{X}}, P)$, where $(\mathcal{X}, d)$ is a metric space, $\mathcal{B}_{\mathcal{X}}$ is the Borel $\sigma$-algebra, and $P$ is the pushforward measure defined by $P(B) = \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B))$ for all $B \in \mathcal{B}_{\mathcal{X}}$.

*Date*: July 19, 2024.

The following proposition justifies the interchange between talking about the expectation of the function $g(X)$ and the integral of $g$ with respect to the pushforward measure of $X$ that we will use to define weak convergence in Definition 2.3 below.

**Proposition 2.1.** *Suppose $g$ is Borel measurable and suppose $g$ is either bounded or nonnegative. Then*

$$\mathbb{E}[g(X)] = \int g(x) dP_X,$$

*where $P_X$ is the pushforward measure $X_*\mathbb{P}$.*

*Proof.* You proved this change of variables formula when you completed the exercises in the fourth set of notes.

As a sketch, you can prove this using the "usual machine": prove it for simple functions, then approximate with an increasing sequence of nonnegative simple functions and use the monotone convergence theorem to conclude it for nonnegative functions, then use linearity of expectation to conclude the general case. □

**Definition 2.2.** Let $(\mathcal{X}, d)$ be a metric space. $f : \mathcal{X} \to \mathbb{R}$ is ***Lipschitz*** if there exists $0 < K < \infty$ such that $|f(x) - f(y)| \leq K \cdot d(x, y)$ for all $x, y \in \mathcal{X}$. We denote ***the set of bounded Lipschitz functions*** by

$$BL(\mathcal{X}) := \left\{ f : \mathcal{X} \to \mathbb{R} \mid \sup_x |f(x)| < \infty, \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x, y)} < \infty \right\}.$$

**Definition 2.3.** Let $\{P_n\}_n$ and $P$ be probability measures on $(\mathcal{X}, \mathcal{B}_\mathcal{X})$. We say that a sequence of probability measures $P_n$ ***converges weakly*** to $P$, denoted $P_n \rightsquigarrow P$, if $\lim_{n\to\infty} P_n f = Pf$ for all $f \in BL(\mathcal{X})$, where $P_n f = \int f dP_n$ and $Pf = \int f dP$.

We say that a sequence of random variables $X_n$ ***converges weakly*** to $X$, denoted $X_n \rightsquigarrow X$, if $P_n \rightsquigarrow P$, where $P_n$ and $P$ are the pushforward measures of $X_n$ and $X$, respectively. Equivalently, $X_n \rightsquigarrow X$ if $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ for every $f \in BL(\mathcal{X})$.

If $X_n$, $X$ are real random variables, we can also say that $X_n$ ***converges in distribution*** to $X$. (The reason why will become apparent when we discuss Theorem 2.6.)

This is a much more abstract definition than the one you may be familiar with, but it is helpful in that it extends to general metric spaces $\mathcal{X}$, where simply talking about CDFs does not suffice. It is a type of convergence which only concerns itself with the *distributions* of the random variables, not the specific random variables themselves on a specific probability space. For instance, if $X \sim N(0, 1)$, then $-X \sim N(0, 1)$ also, but of course $X \neq -X$ unless $X = 0$ (and hence $X \neq -X$ almost surely). It is this distinction which indeed makes this type of convergence weaker than others.

In addition, weak convergence starts to capture a more general notion of expectations being preserved under nice transformations of the variables, which then allows one to use linearity of expectation to prove results. These ideas appear, for instance, in proofs of the central limit theorem that don't revolve around characteristic functions.

However, since weak convergence will ultimately tie back to convergence of CDFs in the case of real random variables, we now give some results on CDFs that you may not have seen before but will prove useful when solving problems.

**Proposition 2.4.** *Let $F(x) = \mathbb{P}(X \leq x)$ be the CDF of a random variable $X$. Then the following hold:*

*(1) $F$ is nondecreasing.*
*(2) $F$ is right continuous with left limits.*
*(3) $\lim_{n\to\infty} F(x) = 1$ and $\lim_{n\to-\infty} F(x) = 0$.*
*(4) $F$ is continous at $x$ if and only if $\mathbb{P}(X = x) = 0$.*

*Proof.* (1) is immediate. To prove (2), let us first show right continuity by fixing $x_0 \in \mathbb{R}$ and taking a sequence $x_n \downarrow x_0$. Then

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(X^{-1}((-\infty, x_n])) &= \lim_{n\to\infty} \mathbb{P}(X^{-1}(\cap_{k=1}^{n}(-\infty, x_k])) \\ &= \lim_{n\to\infty} \mathbb{P}(\cap_{k=1}^{n} X^{-1}((-\infty, x_k])) \\ &= \mathbb{P}(\cap_{k=1}^{\infty} X^{-1}((-\infty, x_k])) \\ &= \mathbb{P}(X^{-1}(\cap_{k=1}^{\infty}(-\infty, x_k])) \\ &= \mathbb{P}(X^{-1}((-\infty, x_0])) \\ &= \mathbb{P}(X \leq x_0),\end{aligned}$$

where we use the fact that preimage commutes with set operations and the continuity property of measure on decreasing sets.

To see that a left limit exists at $x_0$, note that since $F$ is nondecreasing, we have $\lim_{x\uparrow x_0} F(x) = \sup_{x<x_0} F(x)$.

To prove (3), we follow a similar argument. Let $x_n \downarrow -\infty$. Then

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(X^{-1}((-\infty, x_n])) &= \lim_{n\to\infty} \mathbb{P}(\cap_{k=1}^{n} X^{-1}((-\infty, x_k])) \\ &= \mathbb{P}(\cap_{k=1}^{\infty} X^{-1}((-\infty, x_k])) \\ &= \mathbb{P}(X^{-1}(\cap_{k=1}^{\infty}(-\infty, x_k])) \\ &= \mathbb{P}(X^{-1}(\emptyset)) \\ &= 0.\end{aligned}$$

The limit $\lim_{x\to\infty} \mathbb{P}(X \leq x) = 1$ is similar.

To prove (4), we show that $\lim_{x\uparrow x_0} F(x) = \mathbb{P}(X < x_0)$. Let $x_n \uparrow x_0$. Then

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(X \leq x_n) &= \lim_{n\to\infty} \mathbb{P}(\cup_{k=1}^{n} X^{-1}((-\infty, x_k])) \\ &= \mathbb{P}(\cup_{k=1}^{\infty} X^{-1}((-\infty, x_k])) \\ &= \mathbb{P}(X^{-1}(\cup_{k=1}^{\infty}(-\infty, x_k])) \\ &= \mathbb{P}(X^{-1}((-\infty, x_0))).\end{aligned}$$

Now, since $\mathbb{P}(X \leq x_0) = \mathbb{P}(X < x_0) + \mathbb{P}(X = x_0)$, and $F$ is right continuous at $x_0$, it follows that $F$ is continuous at $x_0$ if and only if $\mathbb{P}(X = x_0) = 0$. □

Part (2) of the result above shows that $F$ is in a class of functions called right continuous with left limits, abbreviated RCLL, and sometimes called càdlàg. As a consequence of property (4), we will call $x$ a *continuity point* if $\mathbb{P}(X = x) = 0$.

2.1. **The Portmanteau Theorem.** A portmanteau is a name for a suitcase that opens into two hinged compartments, and it is also the name for when you smash parts of two words together to make a new one (e.g. smoke + fog = smog). In the same way, the Portmanteau Theorem below is the combining of several different equivalencies that characterize convergence in distribution. Part of the theorem pertains to upper and lower semicontinuous functions, which we introduce now and are in reference to the topological definition of continuity. Its main utility in this context is in relating these functions to open and closed sets.

**Definition 2.5.** A function $g : \mathcal{X} \to \mathbb{R}$ is ***lower semicontinuous (LSC)*** if $\{x \in \mathcal{X} : g(x) > t\} =: \{g > t\}$ is open for all $t \in \mathbb{R}$. Similarly, $f : \mathcal{X} \to \mathbb{R}$ is ***upper semicontinuous (USC)*** if $-f$ is LSC, or, equivalently, if $\{f < t\}$ is open for all $t \in \mathbb{R}$.

**Theorem 2.6** (Portmanteau Theorem)**.** *The following are equivalent:*

*(1)* $P_n \rightsquigarrow P$
*(2)* $\liminf_n P_n g \geq Pg$ *for all LSC* $g$ *bounded from below*
*(3)* $\limsup_n P_n f \leq Pf$ *for all USC* $f$ *bounded from above*
*(4)* $\liminf_n P_n(B) \geq P(B)$ *for all open sets* $B$
*(5)* $\limsup_n P_n(G) \leq P(G)$ *for all closed sets* $G$
*(6)* $\lim_n P_n f = Pf$ *for all bounded functions almost surely continuous with respect to* $P$.
*(7)* $\lim_n P_n(B) = P(B)$ *for all Borel sets* $B$ *with* $P(\partial B) = 0$.

*Proof.* We give a partial proof of this. Once we have shown (1) $\implies$ (2) $\iff$ (3), we can prove (4) and (5) by taking the LSC function $g = \mathbf{1}_B$ for open sets $B$, and by taking the USC function $f = \mathbf{1}_G$ for closed sets $G$.

We can show (4) $\iff$ (5). If $\liminf_n P_n(B) \geq \mathbb{P}(B)$ for all open sets $B$, then if $G$ is closed, we can apply the condition to the open set $G^c$, so

$$\begin{aligned}
\limsup_n P_n(G) &= \limsup_n (1 - P_n(G^c)) \\
&= 1 + \limsup_n (-P_n(G^c)) \\
&= 1 - \liminf_n P_n(G^c) \\
&\leq 1 - P(G^c) \\
&= P(G).
\end{aligned}$$

The other direction is essentially identical.

We prove (2) $\implies$ (6). Define $\bar{f}(x)$ pointwise to be the smallest USC function greater than $f(x)$, and $\mathring{f}(x)$ to be the largest LSC function smaller than $f(x)$. For any continuous $f$, we have $\mathring{f} = \bar{f}$ (the converse is also true). For any bounded $f$, we have

$$\begin{aligned}
P\bar{f} \geq \limsup_n P_n \bar{f} &\geq \limsup_n P_n f \\
&\geq \liminf_n P_n f \\
&\geq \liminf_n P_n \mathring{f} \\
&\geq P\mathring{f}.
\end{aligned}$$

But if $f$ is continuous, we have $\overset{\circ}{f} = \bar{f}$, so we conclude $Pf \geq \limsup_n P_n f \geq \liminf \geq Pf$, implying the limit.

It holds that (6) $\implies$ (1) by definition.

There are some other steps needed to show the rest of the equivalencies, but this is a start. You will prove some of the remaining parts in the exercises. □

The following establishes the connection between weak convergence as convergence of CDFs and in the sense described above, since $\partial(-\infty, x] = \{x\}$.

**Corollary 2.7.** *For real-valued random variables, $X_n$, $X$ with distributions $P_n$, $P$, we have $X_n \rightsquigarrow X$ if and only if $P_n((-\infty, x]) \to P((-\infty, x])$ for every continuity point of $P$, i.e. for every $x$ such that $P(x) = 0$.*

2.2. **Three Important Weak Convergence Theorems.** To introduce the main theorems for this section, we need to recall the definitions of convergence in probability and almost sure convergence. For completeness, I will include these definitions here.

**Definition 2.8.** A sequence of random variables $X_n$ converges in probability to $X$, denoted $X_n \xrightarrow{\mathbb{P}} X$, if for every $\varepsilon > 0$,

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

**Definition 2.9.** $X_n$ converges to $X$ almost surely, denoted $X_n \overset{a.s.}{\to} X$, if

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

Now, later in these notes, we will find that the other two main forms of convergence enjoy linearity properties. Weak convergence does not have a linearity property in general except for special cases which we show below. We first collect a helpful result relating weak convergence to convergence in probability.

**Proposition 2.10.** *Suppose $Y_n \rightsquigarrow Y$, where $Y = c \in \mathbb{R}$ almost surely. Then $Y_n \xrightarrow{\mathbb{P}} Y$.*

*Proof.* If $X_n$ converges weakly to $X = c$ a.s., then

$$\mathbb{P}(X_n - c > \varepsilon) = \mathbb{P}(X > c + \varepsilon) = 1 - \mathbb{P}(X_n \leq c + \varepsilon),$$

with the rightmost expression converging to $1 - \mathbb{P}(X \leq c+\varepsilon) = 0$. One can similarly show that $\mathbb{P}(X_n - c < -\varepsilon) \to 0$, hence $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$. □

**Theorem 2.11** (Slutsky's Theorem)**.** *Assume that $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$. Then $X_n + Y_n \rightsquigarrow X + c$ and $X_n Y_n \rightsquigarrow cX$.*

*Proof.* By the previous result, we use that $Y_n \xrightarrow{\mathbb{P}} c$. Let $x$ be such that $x - c$ is a continuity point of $F_X$. Choose $\varepsilon$ such that $x - c + \varepsilon$ is again a continuity point. Then

$$\begin{aligned}\mathbb{P}(X_n + Y_n \leq x) &\leq \mathbb{P}(X_n + c \leq x + \varepsilon) + \mathbb{P}(|Y_n - c| > \varepsilon) \\ &\to \mathbb{P}(X \leq x - c + \varepsilon).\end{aligned}$$

Now, take $\varepsilon \downarrow 0$ and use right continuity of the CDF to obtain that

$$\limsup_n \mathbb{P}(X_n + Y_n \leq x) \leq \mathbb{P}(X + c \leq x).$$

One handles the lim inf similarly. One also handles the second statement similarly. □

**Corollary 2.12.** *If $X_n \rightsquigarrow X, Y_n \rightsquigarrow 0$, then $X_n Y_n \xrightarrow{\mathbb{P}} 0$.*

*Proof.* Apply Slutsky's Theorem, then apply Proposition 2.10. □

**Theorem 2.13** (Continuous Mapping Theorem)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then the following hold:*

*(1) If $X_n \rightsquigarrow X$, then $f(X_n) \rightsquigarrow f(X)$.*
*(2) If $X_n \xrightarrow{\mathbb{P}} X$, then $f(X_n) \xrightarrow{\mathbb{P}} f(X)$.*
*(3) If $X_n \xrightarrow{a.s.} X$, then $f(X_n) \xrightarrow{a.s.} f(X)$.*

*Proof.* To prove (1), recall from the Portmanteau Theorem (Theorem 2.6) that $X_n \rightsquigarrow X$ if and only if $\mathbb{E}[h(X_n)] \to \mathbb{E}[h(X)]$ for every almost surely continuous and bounded function $h$, and we wish to show that if $g$ is any almost surely continuous and bounded function, then $\mathbb{E}[g(f(X_n))] \to \mathbb{E}[g(f(X))]$. Then note that $h = g \circ f$ is continuous and bounded. The result follows.

To prove (2), fix $\varepsilon > 0$ and $x \in \mathbb{R}$.

Now, for each $\delta > 0$, put $W_\delta := \{x : \text{ there exists } y \text{ with } |x - y| \leq \delta \text{ but } |f(x) - f(y)| > \varepsilon\}$. We then have

$$\begin{aligned}\mathbb{P}(|f(X_n) - f(X)| > \varepsilon) &\leq \mathbb{P}(|X_n - X| > \delta) + \mathbb{P}(X \in W_\delta) \\ &\overset{n\to\infty}{\longrightarrow} 0 + \mathbb{P}(X \in W_\delta).\end{aligned}$$

Now, by continuity of measure for decreasing sets (and by taking any decreasing sequence $\delta_n \downarrow 0$), we have $\lim_{\delta \downarrow 0} \mathbb{P}(X \in W_\delta) = \mathbb{P}(X \in \emptyset) = 0$. The result follows.

We now prove (3). If $f$ is continuous at $x$, then $f(x_n) \to f(x)$ for any sequence such that $x_n \to x$, so in particular, if $f$ is continuous at $X(\omega)$, then $f(X_n(\omega)) \to f(X(\omega))$ if $X_n(\omega) \to X(\omega)$. The set on which the latter convergence happens is of probability 1, hence so is the convergence $f(X_n(\omega)) \to f(X(\omega))$. □

**Theorem 2.14** (Delta Theorem)**.** *Assume $X_n$ is a sequence of random variables and $\theta \in \mathbb{R}$ is such that $\sqrt{n}(X_n - \theta) \rightsquigarrow N(0, \sigma^2)$. Let $g$ be any continuously differentiable function such that $g'(\theta) \neq 0$. Then*

$$\sqrt{n}(g(X_n) - g(X)) \rightsquigarrow N(0, [g'(\theta)]^2 \sigma^2).$$

*Proof.* The main idea is just to perform a Taylor expansion and allow the previous results to perform the rest of the work. We have

$$g(X_n) = g(\theta) + g'(\tilde{\theta}_n)(X_n - \theta) \tag{2.15}$$

for some $\tilde{\theta}_n$ between $X_n$ and $\theta$. We claim that $\tilde{\theta}_n \xrightarrow{\mathbb{P}} \theta$. We have $|\tilde{\theta}_n - \theta| \leq |X_n - \theta|$. Now, $\sqrt{n}(X_n - \theta) \rightsquigarrow N(0, \sigma^2)$, and therefore

$$X_n - \theta = \underbrace{(1/\sqrt{n})}_{\rightsquigarrow 0} \cdot \underbrace{\sqrt{n}(X_n - \theta)}_{\rightsquigarrow N(0,\sigma^2)}.$$

Then, by Slutsky's Theorem, we have $X_n - \theta \xrightarrow{\mathbb{P}} 0$. It then follows that $\mathbb{P}(|\tilde{\theta}_n - \theta| > \varepsilon) \leq \mathbb{P}(|X_n - \theta| > \varepsilon) \to 0$. Now, from (2.15), we obtain

$$\sqrt{n}(g(X_n) - g(\theta)) = \underbrace{g'(\tilde{\theta}_n)}_{\xrightarrow{\mathbb{P}} g'(\theta)} \underbrace{\sqrt{n}(X_n - \theta)}_{\rightsquigarrow N(0,\sigma^2)},$$

where the first convergence on the RHS follows from the Continuous Mapping Theorem. Slutsky's theorem implies that the right side of the above converges to a random variable with distribution $g'(\theta)N(0,\sigma^2) = N(0,[g'(\theta)]^2\sigma^2)$. □

The theorem we proved above can be generalized slightly so that continuous differentiability is not required. For completeness, I will state (but not prove) the more general version as well.

**Theorem 2.16.** *Assume the same setting as Theorem 2.14, except now only assume that $g'(\theta)$ exists and is nonzero. Then again we have*

$$\sqrt{n}(g(X_n) - g(X)) \rightsquigarrow N(0,[g'(\theta)]^2\sigma^2).$$

## 3. Convergence In Probability and Almost Sure Convergence

At this point in the program, we are hopefully no strangers to convergence in probability, and we are certainly no strangers to almost sure convergence. Now, of the three types of convergence we've discussed, weak convergence has the most theory behind it, as you may have guessed. Almost sure convergence is relatively simple and intuitive enough; it is almost everywhere convergence for probability spaces, and, of the three types that we have seen, it is the strongest. Some of its common uses are in results such as the Monotone Convergence Theorem and the Dominated Convergence Theorem that we saw in the second set of notes, and that it is the type of convergence seen in the Strong Law of Large Numbers from the third set of notes. Results such as the strong law and the linearity results we'll discuss below are possible ways for proving almost sure convergence. Convergence in probability, on the other hand, is a sort of halfway point between the two. It tends to serve more as a bridge between either of the other two types of convergence.

Furthermore, convergence in probability only stipulates that the size of the set on which errors of size at least $\varepsilon$, so called $\varepsilon$-errors, are allowed is decreasing to 0, but there may still be a nontrivial set on which $\varepsilon$-errors happen infinitely often; almost sure convergence prevents this behavior.

The following proposition shows the bridge between convergence in probability and almost sure convergence and provides an example to illustrate the aforementioned remark.

**Proposition 3.1.** *The following statements hold:*

*(1) Assume that $X_n \overset{a.s.}{\to} X$. Then $X_n \overset{\mathbb{P}}{\to} X$.*

*(2) There exists a sequence $X_n$ and random variable $X$ such that $X_n \overset{\mathbb{P}}{\to} X$ but $X_n \overset{a.s.}{\not\to} X$.*

*(3) Assume that $X_n \overset{\mathbb{P}}{\to} X$. Then there exists a subsequence $X_{n_k}$ such that $X_{n_k} \overset{a.s.}{\to} X$.*

*Proof.* To prove (1), assume $X_n \overset{a.s.}{\to} X$ and let $\varepsilon > 0$. Define the events $A_n := \{|X_n - X| > \varepsilon\}$. Then $\mathbf{1}_{A_n} \overset{a.s.}{\to} 0$, and for each $n$, $\mathbf{1}_{A_n} \leq 1$, which is integrable, so we may apply the Dominated Convergence Theorem to conclude

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \mathbb{E}[\mathbf{1}_{A_n}] = \mathbb{E}[0] = 0.$$

(2) was an exercise for you during the second set of notes, but we give an example here. Let our probability space be $(\Omega, \mathcal{B}, \lambda)$, and let our sequence of functions $f_n$ be defined as follows: $f_1 = \mathbf{1}_{[0,1/2]}, f_2 = \mathbf{1}_{[1/2,1]}, f_3 = \mathbf{1}_{[0,1/3]}, f_4 = \mathbf{1}_{[1/3,2/3]}, f_5 =$

$\mathbf{1}_{[2/3,1]}, f_6 = \mathbf{1}_{[0,1/4]}, f_7 = \mathbf{1}_{[1/4,2/4]}, \dots$ . The intervals move across $[0,1]$ until hitting 1, where they move back to 0 and start again, this time with decreased length. It is clear that this sequence converges to 0 in probability, but since the subintervals march across $[0,1]$ infinitely often, it is clear that $f_n$ cannot converge to $f$ almost surely.

We prove (3). Now suppose $X_n \xrightarrow{\mathbb{P}} X$, let $n_1 = 1$, and choose $n_j$ by induction so that $n_j > n_{j-1}$ and for $n \geq n_j$, we have $\mathbb{P}(|X_n - X| > 1/j) \leq 1/2^j$. Define the events $A_j := \{|X_{n_j} - X| > 1/j\}$. Then

$$\sum_{j=1}^{\infty} \mathbb{P}(A_j) \leq \sum_{j=0}^{\infty} \frac{1}{2^j} < \infty.$$

Then, by the Borel-Cantelli lemma, we have that $\mathbb{P}(A_j \text{ i.o.}) = 0$. This implies $X_{n_j} \xrightarrow{a.s.} X$. □

We now establish a bridge between convergence in probability and weak convergence.

**Proposition 3.2.** *(1) Suppose that $X_n \xrightarrow{\mathbb{P}} X$. Then $X_n \rightsquigarrow X$.*

*(2) There exists a sequence $X_n$ and random variable $X$ such that $X_n \rightsquigarrow X$ but $X_n \not\xrightarrow{\mathbb{P}} X$.*

*Proof.* Assume that $X_n \xrightarrow{\mathbb{P}} X$, and let $t$ be such that $\mathbb{P}(X = t) = 0$. Let $\varepsilon > 0$. Then, for any $n$,

$$\begin{aligned}
\mathbb{P}(X_n \leq t) &= \mathbb{P}(X_n \leq t \cap X \leq t + \varepsilon) + \mathbb{P}(X_n \leq t \cap X > t + \varepsilon) \\
&\leq \mathbb{P}(X \leq t + \varepsilon) + \mathbb{P}(X_n \leq t \cap X > t + \varepsilon) \\
&\leq \mathbb{P}(X \leq t + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon) \\
&\overset{n\to\infty}{\longrightarrow} \mathbb{P}(X \leq t + \varepsilon).
\end{aligned}$$

Now, by (right)-continuity of the CDF of $X$ at $t$, we have $\lim_{\varepsilon \downarrow 0} \mathbb{P}(X \leq t + \varepsilon) = \mathbb{P}(X \leq t)$. Hence, $\limsup_{n\to\infty} \mathbb{P}(X_n \leq t) \leq \mathbb{P}(X \leq t)$. Similarly, for any $n$, we have

$$\begin{aligned}
\mathbb{P}(X \leq t - \varepsilon) &= \mathbb{P}(X \leq t - \varepsilon \cap X_n \leq t) + \mathbb{P}(X \leq t - \varepsilon \cap X_n > t) \\
&\leq \mathbb{P}(X_n \leq t) + \mathbb{P}(X \leq t - \varepsilon \cap X_n > t) \\
&\leq \mathbb{P}(X_n \leq t) + \mathbb{P}(|X_n - X| > \varepsilon).
\end{aligned}$$

Hence, taking $n \to \infty$, we obtain $\mathbb{P}(X \leq t - \varepsilon) \leq \liminf_{n\to\infty} \mathbb{P}(X_n \leq t)$. Now, since $\mathbb{P}(X = t) = 0$, we recall from Proposition 2.4 that the CDF of $X$ is therefore continuous at $t$. Hence, we conclude $\lim_{\varepsilon \downarrow 0} \mathbb{P}(X \leq t - \varepsilon) = \mathbb{P}(X \leq t) \leq \liminf_{n\to\infty} \mathbb{P}(X_n \leq t)$. It follows that

$$\limsup_{n\to\infty} \mathbb{P}(X_n \leq t) \leq \mathbb{P}(X \leq t) \leq \liminf_{n\to\infty} \mathbb{P}(X_n \leq t).$$

The statement follows.

To prove (2), we let $X = X_1 = X_2 = X_3 = \cdots$ have an $N(0,1)$ distribution. Then $X_n \rightsquigarrow -X$, but

$$\mathbb{P}(|X_n - (-X)| > \varepsilon) = \mathbb{P}(2|X| > \varepsilon) > 0.$$

The probability does not depend on $n$ and hence cannot decrease to 0 as $n \to \infty$. □

We now turn to some linearity and closure results that allow one to conclude that combinations of convergent sequences are also convergent. This is noteworthy because this is not true in general for weak convergence, as we have seen.

**Proposition 3.3.** *Let $X_n \xrightarrow{\mathbb{P}} X$, and let $Y_n \xrightarrow{\mathbb{P}} Y$. Then the following hold:*

*(1) $X_n + Y_n \xrightarrow{\mathbb{P}} X + Y$.*

*(2) $X_n Y_n \xrightarrow{\mathbb{P}} XY$. In particular, if $Y_n = Y = c \in \mathbb{R}$, then $cX_n \xrightarrow{\mathbb{P}} cX$, implying that linear combinations of sequences convergent in probability also converge in probability.*

*Proof.* To prove (1), we have that

$$\begin{aligned}\mathbb{P}(|X_n + Y_n - (X+Y)| > \varepsilon) &\leq \mathbb{P}(|X_n - X| + |Y_n - Y| > \varepsilon) \\ &\leq \mathbb{P}(|X_n - X| > \varepsilon/2 \cup |Y_n - Y| > \varepsilon/2) \\ &\leq \mathbb{P}(|X_n - X| > \varepsilon/2) + \mathbb{P}(|Y_n - Y| > \varepsilon/2) \\ &\to 0\end{aligned}$$

as $n \to \infty$ by the assumptions. To conclude the second inequality, assume both the inequalities inside the probability are false and conclude the inequality inside the probability on the RHS of the first inequality the first cannot happen.

We now prove (2). Define $\alpha_{u,\varepsilon} = \varepsilon(2u + \varepsilon)$. One may check that

$$\begin{aligned}\mathbb{P}(|X_n Y_n - XY| \geq \alpha_{u,\varepsilon}) \leq \mathbb{P}(|X_n - X| > \varepsilon) + \mathbb{P}(|Y_n - Y| > \varepsilon) + \mathbb{P}(|X| \geq u) \\ + \mathbb{P}(|Y| \geq u).\end{aligned}$$

To check this, again check that if each of the inequalities on the RHS are false, then $|X_n Y_n - XY| = |X_n Y_n - X_n Y + X_n Y - XY| < \alpha_{u,\varepsilon}$. We then deduce from the above that

$$\limsup_{n\to\infty} \mathbb{P}(|X_n Y_n - XY| \geq \alpha_{u,\varepsilon}) \leq \mathbb{P}(|X| \leq u) + \mathbb{P}(|Y| \geq u).$$

Observe that for any $\eta > 0$ and any $u > 0$, we may choose some $\varepsilon' > 0$ such that $\eta \geq \alpha_{u,\varepsilon'}$. Then, for any $\eta > 0$ and any $u > 0$, we have

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(|X_n Y_n - XY| \geq \eta) &\leq \limsup_{n\to\infty} \mathbb{P}(|X_n Y_n - XY| \geq \alpha_{u,\varepsilon'}) \\ &\leq \mathbb{P}(|X| \geq u) + \mathbb{P}(|Y| \geq u).\end{aligned}$$

Since this holds for each $u$, we have

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(|X_n Y_n - XY| \geq \eta) &\leq \inf_{u>0} \left(\mathbb{P}(|X| \geq u) + \mathbb{P}(|Y| \geq u)\right) \\ &= 0.\end{aligned}$$

Since $\eta > 0$ was arbitrary, the result follows. □

The same results hold for almost sure convergence. We state them below for the sake of completeness.

**Proposition 3.4.** *Let $X_n \xrightarrow{a.s.} X$ and $Y_n \xrightarrow{a.s.} Y$. Then the following hold:*

*(1) $X_n + Y_n \xrightarrow{a.s.} X + Y$.*

*(2) $X_n Y_n \xrightarrow{a.s.} XY$. In particular, if $Y_n = Y = c \in \mathbb{R}$, then $cX_n \xrightarrow{a.s.} cX$.*

*(3) If $X_n \neq 0$ a.s. and $X \neq 0$ a.s., then $1/X_n \xrightarrow{a.s.} 1/X$.*

*Proof.* We leave checking these properties as an exercise for you. □

## 4. Convergence In Mean

This short section is entirely dedicated to convergence in mean and how it relates to the other types of convergence.

**Definition 4.1.** Let $p \geq 1$. We say that $X_n$ ***converges in $p$-th mean*** to $X$, denoted $X_n \xrightarrow{L^p} X$, if $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0$. If $p = 1$, we say $X_n$ ***converges in mean*** to $X$.

**Theorem 4.2.** *Let $X_n$ and $X$ be real-valued random variables.*

*(1) If $1 \leq p \leq q < \infty$, then $X_n \xrightarrow{L^q} X$ implies $X_n \xrightarrow{L^p} X$.*
*(2) If $X_n \xrightarrow{L^p} X$, then $X_n \xrightarrow{\mathbb{P}} X$.*
*(3) There exists a sequence $X_n$ such that $X_n \xrightarrow{a.s.} X$ but $X_n \not\xrightarrow{L^p} X$ for $p = 1$, hence for any $p \geq 1$.*
*(4) There exists a sequence $X_n$ such that $X_n \xrightarrow{L^p} X$ for all $p \geq 1$ but $X_n \not\xrightarrow{a.s.} X$.*

*Proof.* We can use the same proof idea to prove (1) that you diligently used in the fourth problem set to show that $L^q \subseteq L^p$ on probability spaces. Recall that, by Hölder's inequality, we have

$$\begin{aligned}\mathbb{E}[|X_n - X|^p] = \|(X_n - X)^p\|_1 &\leq \|1\|_{\frac{q}{q-p}} \cdot \|(X_n - X)^p\|_{\frac{q}{p}} \\ &= \mathbb{E}[|X_n - X|^q]^{p/q} \\ &\to 0.\end{aligned}$$

Next, (2) follows from Markov's inequality:

$$\begin{aligned}\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - X|^p > \varepsilon^p) &\leq \frac{\mathbb{E}[|X_n - X|^p]}{\varepsilon^p} \\ &\to 0.\end{aligned}$$

For (3), we consider the probability space $([0,1], \mathcal{B}, \lambda)$ and define $f_n = n\mathbf{1}_{[0,1/n]}$. Then $f_n \xrightarrow{a.s.} 0$, but $\int |f_n| = 1$ for each $n$, so $f_n$ cannot converge to 0 in $L^1$.

To prove (4), observe that the typewriter sequence defined in Proposition 3.2 converges to 0 in $L^p$ for any $p$. But as we have already seen in Proposition 3.2, this sequence cannot converge to 0 almost surely. □

The point of Theorem 4.2 above is two-fold: first, showing convergence in $p$-th mean is an effective way of showing convergence in probability, especially for $p = 2$, because it is generally easier to work with a squared random variable than an absolute value. (This trick is used, for example, when showing convergence of the quadratic variation of Brownian motion.) Second, the relationship between convergence in mean and almost sure convergence is not straightforward (the equivalent condition is called *uniform integrability*); however, with additional assumptions about the sequence, one can conclude that almost sure convergence implies convergence in mean. That is why we need results like the Dominated Convergence Theorem, which gives a sufficient condition.

## 5. A Brief Look at Applications In Statistics

5.1. **Confidence Intervals.** The central limit theorem implies that if $X_1, X_2, ...$ are iid with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$, then

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \rightsquigarrow N(0,1). \tag{5.1}$$

A common problem in statistics is in the estimation of $\mu$. It first appears that we would need to know $\sigma$ in order to carry out this calculation, based on (5.1). But we are doing statistics here, and we cannot expect to know $\sigma$! Hence, we need an estimator of $\sigma^2$. Let us define

$$\tilde{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{5.2}$$

This is actually not the typical estimator that is used in statistics because it is biased, i.e. $\mathbb{E}[\tilde{\sigma}^2] \neq \sigma^2$, but we will deal with this later. First, the result below shows that our choice of estimator is still a reasonable one.

**Proposition 5.3.** *If $X_1, X_2, ...$ is a sequence of iid random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$, then*

$$\tilde{\sigma}^2 \overset{a.s.}{\to} \sigma^2,$$

*where $\tilde{\sigma}^2$ is defined as in (5.2).*

*Proof.*

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu + (\mu - \bar{X}))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + \frac{2}{n}(\mu - \bar{X})\sum_{i=1}^{n}(X_i - \mu) + \frac{1}{n}\sum_{i=1}^{n}(\mu - \bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + \frac{2}{n}(\mu - \bar{X})(n\bar{X} - n\mu) + (\mu - \bar{X})^2 \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2}_{\overset{a.s.}{\to}\sigma^2} - \underbrace{(\mu - \bar{X})^2}_{\overset{a.s.}{\to}0} \\
&\overset{a.s.}{\to} \sigma^2.
\end{aligned}$$

The convergences $(\mu - \bar{X})^2 \overset{a.s.}{\to} 0$ and $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \overset{a.s.}{\to} \sigma^2$ follow from the strong law of large numbers. The former follows from the convergence $\bar{X} \overset{a.s.}{\to} \mu$ and the continuous mapping theorem. Linearity of almost sure convergence allows us to conclude the last line from the one before. □

**Corollary 5.4.** *If we define*

$$\hat{\sigma}^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \tag{5.5}$$

*then $\hat{\sigma}^2 \overset{a.s.}{\to} \sigma^2$.*

The estimator (5.5) is the one of interest since it is unbiased.

*Proof.*

$$\hat{\sigma}^2 = \underbrace{\frac{n}{n-1}}_{\overset{a.s.}{\to} 1} \cdot \underbrace{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}_{\overset{a.s.}{\to} \sigma^2}.$$

Now apply Proposition 3.4. □

Now that we know our variance estimator works, we can justify the handwaving done in introductory statistics classes when we use the "plug-in" estimator for $\sigma$ in our confidence intervals. This approach is called the Wald interval.

**Proposition 5.6** (Wald Interval)**.** *If $X_1, X_2, \ldots$ is a sequence of iid random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$, and $z_q$ is the $q$-quantile of the standard normal distribution, then*

$$\lim_{n\to\infty} \mathbb{P}(-z_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\hat{\sigma}} \leq z_{1-\frac{\alpha}{2}}) = 1 - \alpha.$$

*Proof.*

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\hat{\sigma}} = \underbrace{\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}}_{\rightsquigarrow N(0,1)} \cdot \underbrace{\frac{\sigma}{\hat{\sigma}}}_{\overset{a.s.}{\to} 1}.$$

Apply Slutsky's Theorem to conclude the LHS converges in distribution to $N(0,1)$. □

There is an alternative to the Wald approach. This is desirable since the approximation of $\sigma$ by $\hat{\sigma}$ will likely make the convergence in distribution slower when using the Wald interval. In addition, there may be cases where you don't want to have the width of your interval to vary with your $\sigma$ estimate. This is where the variance stabilizing approach enters.

**Proposition 5.7** (Variance Stabilizing Transform)**.** *Assume the same setting as Proposition 5.6. If there exists a continuously differentiable function $g$ such that $[g'(\mu)]^2 = 1/\sigma^2$, then*

$$\sqrt{n}(g(\bar{X}) - g(\mu)) \rightsquigarrow N(0,1),$$

*and therefore*

$$\lim_{n\to\infty} \mathbb{P}(-z_{1-\frac{\alpha}{2}} \leq \sqrt{n}(g(\bar{X}) - g(\mu)) \leq z_{1-\frac{\alpha}{2}}) = 1 - \alpha.$$

*Proof.* This follows from the delta theorem Theorem 2.14. □

It is best to see how the variance stabilizing CI operates by example.

**Example 5.8** (Variance Stabilizing)**.** Let $X_1, X_2, \ldots \overset{iid}{\sim} \mathrm{Poisson}(\lambda)$. Then $\mathbb{E}[X_i] = \mathrm{Var}[X_i] = \lambda$, our goal of estimation. We then have

$$\sqrt{n}(\bar{X} - \lambda) \rightsquigarrow N(0, \lambda).$$

Observe, then, that choosing $g(x) = 2\sqrt{x}$ and applying the variance stabilizing transform, we obtain

$$2\sqrt{n}(\sqrt{\bar{X}} - \sqrt{\lambda}) \rightsquigarrow N(0,1).$$

It follows that

$$\lim_{n\to\infty} \mathbb{P}\left(2\sqrt{n}|\sqrt{\bar{X}} - \sqrt{\lambda}| \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

After rearranging, one obtains an approximate $1-\alpha$ confidence interval for $\lambda$ as

$$\lambda \in \left[ \left( \sqrt{\bar{X}} - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}} \right)^2, \left( \sqrt{\bar{X}} + \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}} \right)^2 \right].$$

Empirically, the variance stabilizing tends to perform a little better than the Wald interval.

5.2. **The Kolmogorov-Smirnov Distance and Test.** Suppose again that we have a sequence of iid random variables $X_1, X_2, \ldots$ . Again, we are doing statistics here. We don't know much about the distribution that generates these random variables. How can we approximate the distribution from our samples alone?

**Definition 5.9.** Define the ***empirical distribution function*** (EDF) of a sample of random variables $X_1, \ldots, X_n$ by

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}.$$

The EDF gives us a low-resolution approximation of the CDF of our distribution, which we denote by $F$ (we assume throughout this section that the variables are iid). It approximates $F(x)$ by simply counting the proportion of observations from our sample that fell at our below $x$. It is a naive guess at $F$, the best that we can hope for. However, it turns out to have nice properties.

**Theorem 5.10** (Glivenko-Cantelli)**.** *In the setting above, the following hold:*

*(1) For each $x$, we have $F_n(x) \overset{a.s.}{\to} F(x)$.*
*(2) Convergence is in fact uniform, i.e.*

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \overset{a.s.}{\to} 0.$$

*Proof.* Pointwise convergence follows from the strong law of large numbers since $\mathbb{E}[\mathbf{1}_{\{X_i \leq x\}}] = \mathbb{P}(X_i \leq x) = F(x)$. The result to show is uniform convergence. We will prove this for the case when $X$ is continuous because the general case, though also true, requires dealing with the jump points of $F$.

Consider a partition $x_0 = -\infty < x_1 < x_2 < \cdots < x_m = \infty$ such that $F(x_j) - F(x_{j-1}) = 1/m$ for each $j$. For each $x$, let $j$ be such that $x \in [x_{j-1}, x_j]$. Then

$$F_n(x) - F(x) \leq F_n(x_j) - F(x_{j-1}) = F_n(x_j) - F(x_j) + \frac{1}{m},$$

$$F_n(x) - F(x) \geq F_n(x_{j-1}) - F(x_j) = F_n(x_{j-1}) - F(x_{j-1}) - \frac{1}{m}.$$

Hence,

$$\limsup_{n\to\infty} \sup_x |F_n(x) - F(x)| \leq \limsup_{n\to\infty} \max_{j \in \{1,\ldots,m\}} |F_n(x_j) - F(x_j)| + \frac{1}{m}$$
$$\overset{a.s.}{\to} \frac{1}{m},$$

with the almost sure convergence following from the almost sure pointwise convergence in (1). Since $m$ was arbitrary, it follows that the lim sup is 0, hence $\sup_x |F_n(x) - F(x)| \overset{a.s.}{\to} 0$. □

Now that we have seen that EDFs converge to their respective CDFs uniformly, we are approaching a mathematical notion of distance between a sampled distribution and an actual distribution. The Kolmogorov-Smirnov distance between distributions $F$ and $G$ is defined as

$$KS(F,G) = \sup_x |F(x) - G(x)|.$$

We can define $KS(X,Y)$ for random variables $X$ and $Y$ according to their respective distribution functions.

**Example 5.11.** One can show that, under certain conditions, the central limit theorem holds in the Kolmogorov-Smirnov distance, in the sense that if $S_n := \sum_{i=1}^n X_i$, then

$$\lim_{n\to\infty} KS\left(\frac{S_n}{\sqrt{n}}, N(0,1)\right) = 0.$$

(See the *Berry-Esseen theorem.*)

If we want to have a distance between $F_n$ and $F$, we have the distance not between two distributions but between a random variable and a distribution. But the Glivenko-Cantelli theorem above tells us that it makes sense to define $KS(F_n, F)$ as well, in which case we have a random variable, which we use as a test statistic against the null hypothesis $H_0 : X_i \sim F$. If $n$ is large, then $KS(F_n, F)$ should be large under the alternative hypothesis. Once one finds the distribution of $KS(F_n, F)$, one has a test against $H_0$. Unfortunately, the null distribution of this variable turns out to be complicated, and is related to the Brownian bridge, which is a Brownian motion path conditioned on the right endpoint being 0. If $B$ is the Brownian bridge between 0 and 1, then under $H_0$, we have$\sqrt{n}KS(F_n, F) \rightsquigarrow \sup_{t\in[0,1]}|B(F(t))|$.

Without getting bogged down any further in the specifics, the point here is that we can get a reliably accurate picture of the true distribution function, and we can check our empirical distribution against a theoretical distribution, all from samples alone, all underpinned by some powerful results from probability theory.

## Sources

Much of the content I have discussed is in Chapter 21 of Richard Bass's *Real Analysis for Graduate Students*. These notes were also heavily inspired by content from STAT 381.

The proof I have used for Proposition 3.3 is found here.

The proofs for the continuous mapping theorem, delta theorem, and the Glivenko-Cantelli Theorem are found here, here, and here.

## 6. Exercises

**Exercise 1.** Check the details of Proposition 3.4.

**Exercise 2.** Suppose that $X_n$ is a Poisson random variable with parameter $n$ for each $n$. Prove that

$$\frac{X_n - n}{\sqrt{n}} \rightsquigarrow N(0,1).$$

**Exercise 3.** Prove that $\mathbb{E}[|X|] < \infty$ if and only if $\sum_{n=1}^\infty \mathbb{P}(|X| \geq n) < \infty$.

**Exercise 4.** Suppose $\{X_i\}$ is an iid sequence of random variables such that $S_n/n$ converges a.s., where $S_n = \sum_{i=1}^n X_i$.

(1) Prove that $X_n/n \overset{a.s.}{\to} 0$.
(2) Prove that $\sum_{n=1}^\infty \mathbb{P}(|X_n| \geq n) < \infty$.
(3) Prove that $\mathbb{E}[|X_1|] < \infty$.

**Exercise 5.** Suppose $\{X_i\}$ is an iid sequence of random variables such that $\mathbb{E}[X_i] = \mu < \infty$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Show that

$$\sqrt{n}\left(\frac{\mu^2}{\bar{X}} - \bar{X}\right) \rightsquigarrow N(0, 4\sigma^2).$$

**Exercise 6.** For $X_n \sim \mathrm{Poisson}(n)$ as in Exercise 2, show that

$$\sqrt{X_n} - \sqrt{n} \rightsquigarrow N(0, \frac{1}{4}).$$

**Exercise 7.**

(1) Give an example of sequences of random variables $X_n, Y_n$ such that $X_n \rightsquigarrow X$, $Y_n \rightsquigarrow Y$, and $X_n + Y_n \rightsquigarrow X + Y$ and neither sequence converges to a constant. This shows that one can replace the conditions in Slutsky's theorem with other conditions.
(2) Give an example of sequences of random variables $X_n, Y_n$ such that $X_n \rightsquigarrow X$, $Y_n \rightsquigarrow Y$, and $X_n + Y_n \not\overset{d}{\to} X + Y$. This shows that linearity does not hold for weak convergence without other assumptions.

**Exercise 8.** I did not show anything about the seventh condition in the Portmanteau theorem, but you can!

(1) Show that $\overset{\circ}{\mathbf{1}}_B = \mathbf{1}_{\overset{\circ}{B}}$, where $\overset{\circ}{B}$ means the interior of a set and $\overset{\circ}{f}(x) = \sup\{g(x) \leq f(x) : g \text{ is LSC}\}$.
(2) Suppose that $\liminf_n P_n(B) \geq P(B)$ for all open sets $B$. Show $\lim_n P_n(B) = P(B)$ for all $B \in \mathcal{B}_\mathcal{X}$ such that $P(\partial B) = 0$.

**Exercise 9.** An outline of how to show that (7) $\implies$ (1) in the Portmanteau theorem for the real-valued case is as follows:

(1) Show $\{x \in \mathbb{R} : P(\{x\}) > 0\}$ is countable.
(2) Given $h \in BL(\mathbb{R})$ with $0 \leq h \leq 1$, show there exists disjoint intervals $(a_1, b_1], ..., (a_m, b_m]$ such that a) $P(\{a_i\}) = P(\{b_i\}) = 0$ for each $i$ and the set $B = \cup_{i=1}^m (a_i, b_i]$ satisfies $P(B) \geq 1 - \varepsilon$; c) the function oscillates by less than $\varepsilon$ on each $(a_i, b_i]$.
(3) Choose arbitrary points $x_i \in (a_i, b_i]$ for each $i$. Define $g_\varepsilon(x) = \sum_{i=1}^m h(x_i)\mathbf{1}_{(a_i, b_i]}(x)$. Show that $|g_\varepsilon(x) - h(x)| \leq \varepsilon + \mathbf{1}_{B^c}(x)$ for all $x \in \mathbb{R}$.
(4) Show that $P_n h \rightsquigarrow Ph$.

**Exercise 10.** In the continuous mapping theorem, the assumption can be slightly weakened to say that $f$ is almost surely continuous with respect to the distribution of $X$, i.e. $f$ is continuous at $X(\omega)$ for almost every $\omega$. Verify that the proof given still goes through with minimal modifications.