Extra material for MI 2014/15

Ernst Hansen, Niels Richard Hansen, *University of Copenhagen*

Contents

1 Conditional expectations

If X is a real-valued random variable defined on the probability space $(\Omega, \mathbb{F}, P)$, and if $\mathbb{D} \subseteq \mathbb{F}$ is a sub-σ-algebra we can ask the following question: What is the best $\mathbb{D}$-measurable approximation of the random variable X? If X is already $\mathbb{D}$-measurable, then X itself may be the most obvious choice, but if X is not $\mathbb{D}$-measurable, can we define a best $\mathbb{D}$-measurable approximation? It may not be clear why this is an interesting question, but think about the following example. With $X_1, \ldots, X_n$ being n real-valued random variables we let

$$S = X_1 + \ldots + X_n$$

denote their sum, and we let $\mathbb{D} = \sigma(S)$ denote the σ-algebra generated by S. Then any $\mathbb{D}$-measurable random variable is, in fact, a function, $g(S)$, of S. If we observe the sum S but not the individual random variables $X_1, \ldots, X_n$, we might be interested in computing the best prediction of each X_i in terms of the observed sum S. This is equivalent to computing the best $\mathbb{D}$-measurable approximation of X_i. One could suggest that the average S/n is the best prediction given the sum, but how can this be justified? This chapter provides the framework for making it mathematically precise what we mean by "best approximation", and some computational rules are also given. Problem 1.8 deals with the prediction of X_i given the sum S under specific distributional assumptions on $X_1, \ldots, X_n$.

If X has second moment then $X \in \mathcal{L}_2(\Omega, \mathbb{F}, P)$, and since $\mathcal{L}_2(\Omega, \mathbb{D}, P) \subseteq \mathcal{L}_2(\Omega, \mathbb{F}, P)$, we might suggest that the best approximation of X in $\mathcal{L}_2(\Omega, \mathbb{D}, P)$ is the orthogonal projection of X onto $\mathcal{L}_2(\Omega, \mathbb{D}, P)$. This is indeed a sensible idea, but there are a few annoying technical complications. First, $\mathcal{L}_2(\Omega, \mathbb{F}, P)$ is not a Hilbert space since the "norm" is only a pseudo norm, and we have to collect the random variables into equivalence classes to get a Hilbert space. This is a nuisance, and is really a digression. Projections can be defined, as we will see in the proof below, in $\mathcal{L}_2(\Omega, \mathbb{F}, P)$ without reference to equivalence classes. Second, it is convenient to extend the projections to work if X has just first moment.

Theorem 1 (Orthogonal projections in $\mathcal{L}_2$)**.** *Assume that $\mathbb{D} \subseteq \mathbb{F}$ is a sub-σ-algebra. There exists a map*

$$pr : \mathcal{L}_2(\Omega, \mathbb{F}, P) \to \mathcal{L}_2(\Omega, \mathbb{D}, P),$$

which satisfies

$$\int_D X \, dP = \int_D pr(X) \, dP \qquad \text{for every} \;\; D \in \mathbb{D} \tag{1}$$

and

$$\int |pr(X) - pr(\tilde{X})| \, dP \leq \int |X - \tilde{X}| \, dP. \tag{2}$$

Proof: Let

$$\delta = \inf_{Y \in \mathcal{L}_2(\Omega, \mathbb{D}, P)} \int (X - Y)^2 \, dP,$$

and choose a sequence $Y_n \in \mathcal{L}_2(\Omega, \mathbb{D}, P)$ such that $\int (X - Y_n)^2 \, dP \to \delta$. The parallelogram identity gives

$$\int \left(X - \frac{1}{2}(Y_n + Y_m) \right)^2 dP + \int \left(\frac{1}{2}(Y_n - Y_m) \right)^2 dP = \frac{1}{2} \left(\int (X - Y_n)^2 \, dP + \int (X - Y_n)^2 \, dP \right).$$

Since $\frac{1}{2}(Y_n + Y_m) \in \mathcal{L}_2(\Omega, \mathbb{D}, P)$ we have

$$\delta \leq \int \left(X - \frac{1}{2}(Y_n + Y_m) \right)^2 dP,$$

and this implies that

$$\delta + \frac{1}{4} \int (Y_n - Y_m)^2 \, dP \leq \frac{1}{2} \left(\int (X - Y_n)^2 \, dP + \int (X - Y_m)^2 \, dP \right)$$

with the right hand side converging to δ for $n, m \to \infty$. This shows that $Y_1, Y_2, \ldots$ is a Cauchy sequence in $\mathcal{L}_2(\Omega, \mathbb{D}, P)$. By the Riesz-Fischer theorem, $\mathcal{L}_2(\Omega, \mathbb{D}, P)$ is complete and the Cauchy sequence has a limit, that is $Y_n \to Y$ in $\mathcal{L}_2(\Omega, \mathbb{D}, P)$ for $n \to \infty$. We define the map pr by $\mathrm{pr}(X) = Y$.

For $D \in \mathbb{D}$ define

$$c = \varepsilon \, \mathrm{sgn} \left(\int_D X - \mathrm{pr}(X) \, dP \right)$$

for $\varepsilon > 0$, where sgn is the sign function defined by

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases}$$

Since $\operatorname{pr}(X) + c1_D \in \mathcal{L}_2(\Omega, \mathbb{D}, P)$ it follows that

$$\begin{aligned} \delta = \int (X - \operatorname{pr}(X))^2 \, dP \leq \int (X - \operatorname{pr}(X) - c1_D)^2 \, dP \\ &= \int (X - \operatorname{pr}(X))^2 \, dP - 2c \int_D X - \operatorname{pr}(X) \, dP + c^2 P(D) \\ &= \delta - 2\varepsilon \left| \int_D X - \operatorname{pr}(X) \, dP \right| + \varepsilon^2 P(D) \end{aligned}$$

From this we get that

$$\left| \int_D X - \operatorname{pr}(X) \, dP \right| \leq \frac{\varepsilon}{2} P(D),$$

and as $\varepsilon > 0$ was arbitrary, the left hand side must be 0, and this proves (1).

To prove the final inequality define

$$D_+ = (\operatorname{pr}(X) - \operatorname{pr}(\tilde{X}) \geq 0)$$

and D_- likewise, such that

$$|\operatorname{pr}(X) - p(\tilde{X})| = 1_{D_+}(\operatorname{pr}(X) - \operatorname{pr}(\tilde{X})) - 1_{D_-}(\operatorname{pr}(X) - \operatorname{pr}(\tilde{X})).$$

Observing that $D_\pm \in \mathbb{D}$ we get, using (1), that

$$\begin{aligned} \int |\operatorname{pr}(X) - \operatorname{pr}(\tilde{X})| \, dP &= \int_{D_+} \operatorname{pr}(X) \, dP - \int_{D_+} \operatorname{pr}(\tilde{X}) \, dP - \int_{D_-} \operatorname{pr}(X) \, dP + \int_{D_-} \operatorname{pr}(\tilde{X}) \, dP \\ &= \int_{D_+} X \, dP - \int_{D_+} \tilde{X} \, dP - \int_{D_-} X \, dP + \int_{D_-} \tilde{X} \, dP \\ &= \int_{D_+} X - \tilde{X} \, dP - \int_{D_-} X - \tilde{X} \, dP \\ &\leq \int_{D_+} |X - \tilde{X}| \, dP + \int_{D_-} |X - \tilde{X}| \, dP \\ &= \int |X - \tilde{X}| \, dP. \end{aligned}$$

□

The previous theorem establish just enough knowledge about the orthogonal projection map pr for random variables with second moment to be able to extend the map to random variables with first moment. Observe that (2) means that pr is continuous in the $\mathcal{L}_1$-norm, and that this is really the essential property we use below for extending the map from $\mathcal{L}_2$ to $\mathcal{L}_1$.

Theorem 2. *Let X be a real-valued random variable, defined on a background space $(\Omega, \mathbb{F}, P)$. Let $\mathbb{D} \subseteq \mathbb{F}$ be a sub-σ-algebra. If $E\,|X| < \infty$ there exist a $\mathbb{D}$-measurable real-valued random variable Y such that $E\,|Y| < \infty$ and such that*

$$\int_D X\,dP = \int_D Y\,dP \qquad \textit{for every} \;\; D \in \mathbb{D}\,. \tag{3}$$

The variable Y is unique, if we identify variables that are equal almost everywhere.

Proof: We start by proving uniqueness. Let Y and Y' be two solutions to the problem, meaning that both are $\mathbb{D}$-measurable variables with first moment and that

$$\int_D Y\,dP = \int_D X\,dP = \int_D Y'\,dP \qquad \text{for every} \;\; D \in \mathbb{D}\,.$$

In particular the event $(Y > Y')$ is in $\mathbb{D}$. Hence we see that

$$\int_{(Y>Y')} Y - Y'\,dP = 0\,.$$

But $1_{(Y>Y')}(Y - Y')$ is non-negative and strictly positive on $(Y > Y')$. As the integral is zero, we conclude that $P(Y > Y') = 0$. Similarly, $P(Y < Y') = 0$. So it follows that $P(Y \neq Y') = 0$.

To prove existence, observe that $\mathcal{L}_2(\Omega, \mathbb{F}, P)$ is an $\mathcal{L}_1$-dense subset of $\mathcal{L}_1(\Omega, \mathbb{F}, P)$, which is a direct consequence of Corollary 12.11 in Schilling's book. So if $E|X| < \infty$ there is a sequence $X_1, X_2, \ldots$ of variables with $E\,X_n^2 < \infty$, satisfying that

$$\int |X - X_n|\,dP \to 0.$$

By (2)

$$\int |\mathrm{pr}(X_n) - \mathrm{pr}(X_m)|\,dP \leq \int |X_n - X_m|\,dP.$$

Since $X_n \to X$ in $\mathcal{L}_1$, it is in particular an $\mathcal{L}_1$-Cauchy sequence. Consequently the sequence $\mathrm{pr}(X_1), \mathrm{pr}(X_2), \ldots$ is also an $\mathcal{L}_1$-Cauchy sequence. By the Riesz-Fischer theorem $\mathcal{L}_1(\Omega, \mathbb{D}, P)$ is complete, and there exits an $\mathcal{L}_1$-limit Y such that

$$\int |\mathrm{pr}(X_n) - Y|\,dP \to 0.$$

for $n \to \infty$. We claim that this $\mathbb{D}$-measurable limit Y satisfies (3). Indeed, for any $D \in \mathbb{D}$ we have that

$$\begin{aligned}\left|\int_D X\,dP - \int_D Y\,dP\right| &\leq \left|\int_D X\,dP - \int_D X_n\,dP\right| + \left|\int_D X_n\,dP - \int_D \mathrm{pr}(X_n)\,dP\right| \\ &\quad + \left|\int_D \mathrm{pr}(X_n)\,dP - \int_D Y\,dP\right| \\ &\leq \int |X - X_n|\,dP + \int |\mathrm{pr}(X_n) - Y|\,dP\end{aligned}$$

since the middle integral is zero according to (1). As $X_n \to X$ and $\mathrm{pr}(X_n) \to Y$ in $\mathcal{L}_1$, this upper limit can be made arbitrarily small, and hence (3) holds. □

The extension established above of the map pr from $\mathcal{L}_2(\Omega, \mathbb{F}, P)$ to $\mathcal{L}_1(\Omega, \mathbb{F}, P)$ could be denoted pr as well. We prefer the notation commonly used in probability theory where $\mathrm{pr}(X)$ is written $E(X \mid \mathbb{D})$, and is called the conditional expectation of X given $\mathbb{D}$. The content of Theorem 2 is reformulated as the following definition:

Definition 1. *Let X be a real-valued random variable, defined on a background space $(\Omega, \mathbb{F}, P)$, and assume that $E\,|X| < \infty$. Let $\mathbb{D} \subseteq \mathbb{F}$ be a sub-σ-algebra. The* **conditional expectation** *of X with respect to $\mathbb{D}$, denoted $E(X \,|\, \mathbb{D})$, is the $\mathbb{D}$-measurable, integrable real-valued random variable satisfying that*

$$\int_D X\,dP = \int_D E(X \,|\, \mathbb{D})\,dP \qquad \textit{for every} \;\; D \in \mathbb{D}\,. \tag{4}$$

It is common to refer to any specific random variable satisfying the condition of Definition 1 as a *version* of the conditional expectation. Any two versions will be almost everywhere identical, so the distinction between them hardly matters at all.

There are many applications of the concept of conditional expectations. One of the most important is as a tool for computing ordinary expectations; if (4) is applied to $D = \Omega$, the formula reads

$$E\,X = E\Big(E(X \mid \mathbb{D})\Big)\,. \tag{5}$$

So a feasible strategy for computing $E\,X$ is to search for a σ-algebra $\mathbb{D}$ enabling the computation of the right hand side of (5).

Whenever X has second moment the conditional expectation $E(X \mid \mathbb{D})$ has second moment as well, and it is possible to introduce the conditional variance as in the following definition.

Definition 2. *The* **conditional variance** *given $\mathbb{D}$ of a real-valued random X with second moment is*

$$V(X \mid \mathbb{D}) = E\Big(\big(X - E(X \mid \mathbb{D})\big)^2 \mid \mathbb{D}\Big).$$

It follows from Lemma 4 below that $V(X \mid \mathbb{D}) \geq 0$ almost everywhere. The usefulness of conditional variances is intimately linked to the relation

$$V\,X = E\Big(V(X \mid \mathbb{D})\Big) + V\Big(E(X \mid \mathbb{D})\Big)\,, \tag{6}$$

which follows from Corollary 8 below. The derivation is left to the reader as Problem 1.2. There are many cases where hard-to-find unconditional variances can be computed using this conditional approach with a clever choice of $\mathbb{D}$.

In the typical application, the conditioning σ-algebra is not at all arbitrary – it is generated by a relevant random variable Y (which can be real-valued, vector-valued or have values in an arbitrary abstract measurable space). In this case, we will typically write $E(X \,|\, Y)$, repressing the σ-algebra in the notation. Thus in the initial example we would write $E(X_i \mid S_n)$ for the conditional expectation of X_i w.r.t. the sum S. It is a general fact (but not proved in this note) that $E(X \,|\, Y)$ can be written as $g(Y)$ for a suitable real-valued function g, defined on the space where Y has values. This means that $E(X \,|\, Y)$ should be thought of as a function of Y – something that can be computed once Y is known. With this in mind, $E(X \,|\, Y)$ can be understood as the natural **predictor** of X after observation of Y. If X has second moment, this interpretation be most clearly justified by seeing $E(X \,|\, Y)$ as the orthogonal projection on $\mathcal{L}_2(\Omega, \sigma(Y), P)$.

It may seem overly complicated to develop a machinery for conditioning on an abstract σ-algebra if all this machinery is ever used for is to condition on a random variable. But there are considerable notational advantages in the σ-algebra formulation, in particular associated to the fact that several random variables can generate the same σ-algebra. In the σ-algebra formulation we can switch between the conditioning variables without affecting the conditional expectation, and this turns out to be extremely convenient.

If X and X' are two real-valued random variables satisfying that $X = X'$ a.e., it obviously holds that

$$\int_D X\,dP = \int_D X'\,dP \qquad \text{for every } D \in \mathbb{D}\,,$$

if the two variables are integrable. Hence the integral condition defining $E(X \mid \mathbb{D})$ is the same as the integral condition defining $E(X' \mid \mathbb{D})$, and from the uniqueness of the conditional expectation it follows that

$$E(X \mid \mathbb{D}) = E(X' \mid \mathbb{D}) \qquad \text{a.e.}$$

The 'almost everywhere' closing of this formula could be neglected without any problems, but is is customary in probability theory to formulate all results on conditional expectations as holding true almost everywhere, as an admission to the fact that conditional expectations are only determined up to modifications on nullsets.

Example 1. If X itself is $\mathbb{D}$-measurable (as well as integrable), it clearly satisfies the integral condition defining $E(X \mid \mathbb{D})$. Hence we have that

$$E(X \mid \mathbb{D}) = X \qquad \text{a.e.}$$

There is a handful of cases where the conditional expectation can be computed effortlessly, Example 1 being the most obvious. In the exercises there are a few other such cases. But in general some amount of work will be needed, relying on a number of results to be established now.

Lemma 3. *Let X_1 and X_2 be real-valued random variables, defined on a common background space $(\Omega, \mathbb{F}, P)$. Assume that $E\,|X_1| < \infty$ and $E\,|X_2| < \infty$. Let $\mathbb{D} \subseteq \mathbb{F}$ be a sub-σ-algebra. For any $c_1, c_2 \in \mathbb{R}$ it holds that*

$$E(c_1X_1 + c_2X_2 \,|\, \mathbb{D}) = c_1E(X_1 \,|\, \mathbb{D}) + c_2E(X_2 \,|\, \mathbb{D}) \qquad a.e. \tag{7}$$

Proof: Let $E(X_1 \,|\, \mathbb{D})$ and $E(X_2 \,|\, \mathbb{D})$ be versions of the conditional expectations. For any $D \in \mathbb{D}$ it holds that

$$\begin{aligned}\int_D & c_1E(X_1 \,|\, \mathbb{D}) + c_2E(X_2 \,|\, \mathbb{D})\, dP \\ &= c_1 \int_D E(X_1 \,|\, \mathbb{D})\, dP + c_2 \int_D E(X_2 \,|\, \mathbb{D})\, dP \\ &= c_1 \int_D X_1\, dP + c_2 \int_D X_2\, dP \\ &= \int_D c_1X_1 + c_2X_2\, dP\end{aligned}$$

So $c_1E(X_1 \,|\, \mathbb{D}) + c_2E(X_2 \,|\, \mathbb{D})$ is a $\mathbb{D}$-measurable random variable satisfying the relevant integral condition to be a version of $E(c_1X_1 + c_2X_2 \,|\, \mathbb{D})$. □

Lemma 4. *Let X be a real-valued random variable, defined on a background space $(\Omega, \mathbb{F}, P)$. Assume that $E\,|X| < \infty$. Let $\mathbb{D} \subseteq \mathbb{F}$ be a sub-σ-algebra. If $X \geq 0$ a.e., then $E(X \mid \mathbb{D}) \geq 0$ a.e.*

Proof: Let $E(X \,|\, \mathbb{D})$ be a specific version of the conditional expectation. Consider the $\mathbb{D}$-measurable event $\big(E(X \,|\, \mathbb{D}) < 0\big)$. We have that

$$\int_{\left(E(X \,|\, \mathbb{D}) < 0\right)} E(X \,|\, \mathbb{D})\, dP = \int_{\left(E(X \,|\, \mathbb{D}) < 0\right)} X\, dP \geq 0\,,$$

since X is non-negative. But $1_{\left(E(X \,|\, \mathbb{D}) < 0\right)} E(X \,|\, \mathbb{D})$ is non-positive, and strictly negative on $\big(E(X \,|\, \mathbb{D}) < 0\big)$. The only way the integral can avoid being strictly negative is if $\big(E(X \,|\, \mathbb{D}) < 0\big)$ has measure 0, implying that $E(X \mid \mathbb{D}) \geq 0$ a.e. □

We can combine Lemma 7 and Lemma 4 to a monotonicity property for conditional expectations: if $X \geq Y$ are two real-valued random variables with first moment, then $E(X \mid \mathbb{D}) \geq E(Y \mid \mathbb{D})$ a.e. for any sub-σ-algebra $\mathbb{D} \subseteq \mathbb{F}$. This follows from considerations on $X - Y$.

Theorem 5 (Tower property)**.** *Let X be a real-valued random variable, defined on a background space $(\Omega, \mathbb{F}, P)$. Assume that $E\,|X| < \infty$. Let $\mathbb{D} \subseteq \mathbb{E} \subseteq \mathbb{F}$ be two nested sub-σ-algebras. Then it holds that*

$$E\Big(E(X \mid \mathbb{E}) \mid \mathbb{D}\Big) = E(X \mid \mathbb{D}) \qquad a.e. \tag{8}$$

Proof: Let $E(X \mid \mathbb{E})$ and $E(X \mid \mathbb{D})$ be specific versions of the conditional expectations. By definition $E(X \mid \mathbb{E})$ is an integrable real-valued random variable, so the left hand side of (8) makes sense.

The claim in (8) is that $E(X \mid \mathbb{D})$ is a $\mathbb{D}$-measurable random variable (this is obviously satisfied) which has the same integrals over $\mathbb{D}$-sets as $E(X \mid \mathbb{E})$. To check this integration property, take an event $D \in \mathbb{D}$. Since $\mathbb{D}$ and $\mathbb{E}$ are nested, we see that $D \in \mathbb{E}$ automatically. Hence

$$\int_D E(X \mid \mathbb{E})\, dP = \int_D X\, dP = \int_D E(X \mid \mathbb{D})\, dP\,,$$

exactly as desired. □

The tower property is surprisingly useful for computing conditional expectations via a two-step procedure. It is often possible to find an intermediate σ-algebra $\mathbb{E}$ such that $E(X \mid \mathbb{E})$ is easy to compute. And with a bit of luck the result is itself $\mathbb{D}$-measurable, trivializing the outer conditional expectation in (8). Even when $E(X \mid \mathbb{E})$ is not $\mathbb{D}$-measurable, calculation of the outer conditional expectation is frequently an easier task than a direct attempt at computing $E(X \mid \mathbb{D})$.

Theorem 6. *Let $X, X_1, X_2, \ldots$ be real-valued random variables, defined on a common background space $(\Omega, \mathbb{F}, P)$, and let $\mathbb{D} \subseteq \mathbb{F}$ be a sub-σ-algebra. Suppose that $X_n \geq 0$ almost everywhere for every n, and that $X_n \nearrow X$ almost everywhere. If $E\,|X| < \infty$ it holds that*

$$E(X_n \mid \mathbb{D}) \nearrow E(X \mid \mathbb{D}) \qquad a.e.$$

Proof: Since $0 \leq X_n \leq X$ almost everywhere, the assumption that X has first moment implies that each X_n also has first moment, and so the various conditional expectations exists. Choose specific versions.

We have previously noted that the conditional expectation has a monotonicity property, which implies that $E(X_1 \mid \mathbb{D}) \leq E(X_2 \mid \mathbb{D}) \leq \ldots$ almost everywhere. Hence there is a random variable Y (potentially with values in $[0, \infty]$) such that

$$E(X_n \mid \mathbb{D}) \nearrow Y \qquad \text{a.e.}$$

By standard arguments we can assume Y to be $\mathbb{D}$-measurable. The challenge is to prove that Y has finite integral (enabling us to disregard the possibility of Y having the value

∞) and that Y satisfies the integral condition characterizing $E(X \mid \mathbb{D})$. Both properties can be obtained from the same computation. Let $D \in \mathbb{D}$ be arbitrary. Then we have that

$$1_D\, X_n \nearrow 1_D\, X \qquad \text{and} \qquad 1_D\, E(X_n \mid \mathbb{E}) \nearrow 1_D\, Y \qquad \text{a.e.}$$

Relying on the monotone convergence theorem we obtain that

$$\begin{aligned} \int_D Y\, dP &= \int_D \lim_{n\to\infty} E(X_n \mid \mathbb{D})\, dP = \int \lim_{n\to\infty} 1_D\, E(X_n \mid \mathbb{D})\, dP \\ &= \lim_{n\to\infty} \int 1_D\, E(X_n \mid \mathbb{D})\, dP = \lim_{n\to\infty} \int 1_D\, X_n\, dP \\ &= \int_D X\, dP \end{aligned}$$

Using $D = \Omega$ we obtain the desired integrability of Y. And obviously Y satisfies the integral condition characterizing $E(X \mid \mathbb{D})$. □

Corollary 7. *Let X and Y be real-valued random variables, defined on a common background space $(\Omega, \mathbb{F}, P)$, and let $\mathbb{D} \subseteq \mathbb{F}$ be a sub-σ-algebra. Suppose that X is $\mathbb{D}$-measurable. If $E\,|Y| < \infty$ and $E\,|XY| < \infty$ then it holds that*

$$E(XY \mid \mathbb{D}) = X\, E(Y \mid \mathbb{D}) \qquad \textit{a.e.}$$

Proof: The result is easily obtained if X is an indicator, say $X = 1_{D_0}$ for a suitable event $D_0 \in \mathbb{D}$. Clearly, $X\, E(Y \mid \mathbb{D})$ is $\mathbb{D}$-measurable and integrable. And for any $D \in \mathbb{D}$ it holds that

$$\begin{aligned} \int_D X\, E(Y \mid \mathbb{D})\, dP &= \int_D 1_{D_0} E(Y \mid \mathbb{D})\, dP = \int_{D \cap D_0} E(Y \mid \mathbb{D})\, dP \\ &= \int_{D \cap D_0} Y\, dP = \int_D 1_{D_0} Y\, dP = \int_D XY\, dP\,, \end{aligned}$$

using that $D \cap D_0 \in \mathbb{D}$. So the variable $X\, E(Y \mid \mathbb{D})$ satisfies the integral condition characterizing $E(XY \mid \mathbb{D})$.

If X is a simple function, say $X = \sum_{i=1}^n c_i 1_{D_i}$, the result follows from the above calculations by linearity,

$$\begin{aligned} E(XY \mid \mathbb{D}) &= E\left(\sum_{i=1}^n c_i 1_{D_i} Y \mid \mathbb{D}\right) = \sum_{i=1}^n c_i\, E\left(1_{D_i} Y \mid \mathbb{D}\right) \\ &= \sum_{i=1}^n c_i 1_{D_i}\, E(Y \mid \mathbb{D}) = X\, E(Y \mid \mathbb{D})\,. \end{aligned}$$

Before we attack the general case, consider one more special example: Assume that $X \geq 0$ and that $Y \geq 0$. By applying Corollary 5.17 of [MT] we can find simple non-negative

$\mathbb{D}$-measurable variables S_n such that $S_n \nearrow X$. Obviously, this implies that $S_n Y \nearrow XY$ so from Theorem 6 we have that

$$E(S_n Y \mid \mathbb{D}) \nearrow E(XY \mid \mathbb{D}) \qquad \text{a.e.}$$

But as S_n is simple, we also have that

$$E(S_n Y \mid \mathbb{D}) = S_n \, E(Y \mid \mathbb{D}) \to X \, E(Y \mid \mathbb{D}) \qquad \text{a.e.}$$

As the two limits must be equal almost everywhere, we have established the result in the case where both variables are non-negative.

What remains is to reduce the general situation to the non-negative case just treated. This can be done by a bit of bookkeeping. Let X and Y be arbitrary, only subject to the conditions of the theorem. We can decompose each of them into positive and negative parts, $X = X^+ - X^-$ and $Y = Y^+ - Y^-$. Clearly X^+ and X^- are non-negative and $\mathbb{D}$-measurable, and Y^+ and Y^- are nonnegative and integrable. Furthermore, the crossproducts X^+Y^+, X^+Y^-, X^-Y^+ and X^-Y^- are all integrable, for instance

$$E\,|X^+Y^+| \leq E\,|XY| < \infty\,.$$

Hence

$$\begin{aligned}
E(XY \mid \mathbb{D}) &= E\Big((X^+ - X^-)(Y^+ - Y^-) \mid \mathbb{D}\Big) \\
&= E(X^+Y^+ \mid \mathbb{D}) - E(X^+Y^- \mid \mathbb{D}) - E(X^-Y^+ \mid \mathbb{D}) + E(X^-Y^- \mid \mathbb{D}) \\
&= X^+ \, E(Y^+ \mid \mathbb{D}) - X^+ \, E(Y^- \mid \mathbb{D}) - X^- \, E(Y^+ \mid \mathbb{D}) + X^- \, E(Y^- \mid \mathbb{D}) \\
&= (X^+ - X^-) \, E(Y^+ - Y^- \mid \mathbb{D}) \\
&= X \, E(Y \mid \mathbb{D}) \qquad\qquad\qquad\qquad \text{a.e}
\end{aligned}$$

and the result is established. □

To illustrate the use of the computational rules for the conditional expectation, as established above, we show a different representation of the conditional variance akin to the well known formula for the unconditional variance.

Corollary 8. *For X a real-valued random variable with second moment and $\mathbb{D} \subseteq \mathbb{F}$ a sub-σ-algebra it holds that*

$$V(X \mid \mathbb{D}) = E\Big(X^2 \mid \mathbb{D}\Big) - \Big(E(X \mid \mathbb{D})\Big)^2.$$

Proof: Observe that

$$(X - E(X \mid \mathbb{D}))^2 = X^2 - 2XE(X \mid \mathbb{D}) + \Big(E(X \mid \mathbb{D})\Big)^2.$$

By Corollary 7

$$E\Big(XE(X \mid \mathbb{D}) \mid \mathbb{D}\Big) = \Big(E(X \mid \mathbb{D})\Big)^2,$$

and by linearity of the conditional expectation, Lemma 7, we get that

$$\begin{aligned} V(X \mid \mathbb{D}) &= E\Big(X^2 \mid \mathbb{D}\Big) - 2E\Big(XE(X \mid \mathbb{D}) \mid \mathbb{D}\Big) + E\Big(\Big(E(X \mid \mathbb{D})\Big)^2 \mid \mathbb{D}\Big) \\ &= E\Big(X^2 \mid \mathbb{D}\Big) - 2\big(E(X \mid \mathbb{D})\big)^2 + \Big(E(X \mid \mathbb{D})\Big)^2 \\ &= E\Big(X^2 \mid \mathbb{D}\Big) - \Big(E(X \mid \mathbb{D})\Big)^2 \qquad \text{a.e.} \end{aligned}$$

□

1.1 Problems

1.1. Let X be a real-valued random variable defined on a background space $(\Omega, \mathbb{F}, P)$ and let $\mathbb{D} \subseteq \mathbb{F}$ be a sub-σ-algebra. Show that if $X = c$ a.e. then it holds that $E(X \mid \mathbb{D}) = c$ a.e.

1.2. Show the variance formula (6) when X has second moment.

1.3. Let $(\Omega, \mathbb{F}, P)$ be a probability space. Let $\mathbb{D}$ be the σ-algebra generated by a finite partition $\mathbb{H} = (H_1, \dots, H_n)$ of Ω (where every $\mathbb{H}$-atom is $\mathbb{F}$-measurable).

(a) Show that any $\mathbb{D}$-measurable real-valued random variable Y can written in the form

$$Y = \sum_{i=1}^{k} c_i 1_{H_i} \tag{9}$$

for suitable constants $c_1, \dots c_n \in \mathbb{R}$.

(b) Let X be a real-valued random variable with $E\,|X| < \infty$. Show that when $E(X \mid \mathbb{D})$ is written in the form (9) it holds that

$$c_i = \frac{1}{P(H_i)} \int_{H_i} X \, dP$$

for all atoms such that $P(H_i) > 0$. Can anything be said about c_i if $P(H_i) = 0$?

(c) Can the above construction be carried over to the case where $\mathbb{H}$ is a countable partition?

(d) Let Y be a $\mathbb{Z}$-valued random variable on $(\Omega, \mathbb{F}, P)$. Find $E(X \mid Y)$ by computing the value on each atom $(Y = k)$.

1.4. Let X be a positive real valued random variable and let $\lfloor X \rfloor$ denote the integer part of X. That is, if $X = N + \delta$ with $N \in \mathbb{N}_0$ and $\delta \in [0, 1)$ then $\lfloor X \rfloor = N$.

(a) Assume that the distribution of X has density f w.r.t. the Lebesgue measure on $(0,\infty)$. Define for $n \in \mathbb{N}_0$

$$p(n) = \int_n^{n+1} f(x)\,dx \quad \text{and} \quad \xi(n) = \int_n^{n+1} xf(x)\,dx.$$

Show that

$$E(X \mid \lfloor X \rfloor) = \frac{\xi(\lfloor X \rfloor)}{p(\lfloor X \rfloor)} \qquad \text{a.e.}$$

(b) Assume that X is uniformly distributed on $(0,10)$. Show that

$$E(X \mid \lfloor X \rfloor) = \lfloor X \rfloor + \frac{1}{2} \qquad \text{a.e.}$$

(c) Assume that X is exponentially distributed on $(0,\infty)$. Show that

$$E(X \mid \lfloor X \rfloor) = \lfloor X \rfloor + \frac{e-2}{e-1} \qquad \text{a.e.}$$

Compare with the result for the uniform distribution and explain the difference.

1.5. Let X be a real-valued random variable on a probability space $(\Omega, \mathbb{F}, P)$. Assume that $E\,|X| < \infty$. For $x \in \mathbb{R}$ define $\mathbb{D}_x$ as the collection of $\mathbb{F}$-measurable sets of the form $(X \in B)$ or $(X \in B) \cup (X > x)$ where $B \subseteq (-\infty, x]$ is a Borel set.

(a) Show that $\mathbb{D}_x$ is a σ-algebra.

(b) Assume that the distribution of X has density f w.r.t. the Lebesgue measure, and assume that $f(x) > 0$ for all $x > 0$. Show that for $x > 0$

$$E(X \mid \mathbb{D}_x) = X1_{(X \le x)} + \xi(x)1_{(X>x)} \qquad \text{a.e.}$$

where

$$\xi(x) = \frac{1}{P(X > x)} \int_x^\infty yf(y)\,dy$$

(c) Show that if X is Weibull distributed with shape parameter $c > 0$ then

$$\xi(x) = x + e^{x^c} \int_x^\infty e^{-y^c}\,dy.$$

1.6. Let X and Y be random variables defined on a common background space $(\Omega, \mathbb{F}, P)$. Suppose that X is real-valued and that $E\,|X| < \infty$, but Y can have values in an arbitrary measurable space. Show that if X and Y are independent, then

$$E(X \mid Y) = E\,X \qquad \text{a.e.}$$

1.7. Let X be a real-valued, integrable random variable defined on a background space $(\Omega, \mathbb{F}, P)$. Let Y and Z be two random variables on the same background space, with values in arbitrary measurable spaces. Assume that (X, Y) is independent of Z. Show that

$$E\Big(X \mid (Y,Z)\Big) = E(X \mid Y) \qquad \text{a.e.}$$

Hint: the events of the form $(Y \in B, Z \in C)$ is a generator, stable under intersection, for the conditioning σ-algebra on the left.

1.8. Let $X_1, \dots, X_n$ be independent and identically distributed real-valued random variables. Assume that $E\,|X_i| < \infty$. Let $S = \sum_{i=1}^n X_i$.

(a) Show by symmetry that

$$\int_{(S \in B)} X_1 \, dP = \int_{(S \in B)} X_2 \, dP$$

for any set $B \in \mathbb{B}$.

(b) Show that $E(X_1 \mid S) = E(X_2 \mid S)$ almost everywhere.

(c) Conclude that $E(X_1 \mid S) = \frac{1}{n}\, S$.

1.9. Let X and Y be two real-valued random variables defined on a common background space $(\Omega, \mathbb{F}, P)$. Assume that $E\,|X| < \infty$ and that the joint distribution of (X, Y) has density $f(x, y)$ with respect to m_2 on $\mathbb{R}^2$. Recall that this implies that Y has density with respect to m given by

$$g(y) = \int_{-\infty}^{\infty} f(x, y)\, dx\,.$$

Define

$$\phi(y) = \int_{-\infty}^{\infty} x\, \frac{f(x, y)}{g(y)}\, dx \qquad \text{for } y \in \mathbb{R}\,.$$

Show that this definition makes sense for Lebesgue-almost all values of y, and that ϕ can be considered to be a $\mathbb{B}$-measurable function. Prove that

$$E(X \mid Y) = \phi(Y) \qquad \text{a.e.}$$

1.10. Let X and Y be random variables, defined on a common background space $(\Omega, \mathbb{F}, P)$ and with values in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Assume that

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \xi \\ \mu \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right),$$

where Σ_{22} is invertible. Show that

$$E(X \mid Y) = \xi + \Sigma_{12}\, \Sigma_{22}^{-1}\, (Y - \mu) \qquad \text{a.e.}$$

Hint: Introduce the auxiliary variable $Z = X - \Sigma_{12}\Sigma_{22}^{-1} Y$. Find the joint distribution of (Z, Y) and compute $E(Z \mid Y)$.

2 The Central Limit Theorem

2.1 Introduction

The mathematical theorem that has become known as the *Central Limit Theorem*, or just CLT for short, is unquestionably the most important of all results in probability theory. It says that the distribution of the sum, or average, of independent and identically distributed real valued random variables with second moment is well approximated by a normal distribution. We will refer to it as Laplace's CLT, since Laplace was the first to prove general results in this form. It is not really fair to refer to CLT as one theorem, because there are many versions that differ by the assumptions made, and the form in which the conclusions are given. The assumption that the variables are identically distributed can be relaxed without much more work. The independence assumption can also be relaxed, but this is considerably more difficult.

The proof presented in this note is one of the most direct based on little more than the various introductory concepts and results in probability theory, most notably the concept of independence, and then some analytic approximation techniques like the second order Taylor expansion with remainder.

2.2 The Central Limit Theorem

Recall that the standard normal distribution on $(\mathbb{R}, \mathbb{B})$ has distribution function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} \, \mathrm{d}y.$$

Theorem 9 (Laplace's CLT)**.** *If $X_1, X_2, \ldots$ are independent identically distributed real valued random variables with mean ξ and variance σ^2 then*

$$P\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \xi}{\sigma} \leq x \right) \to \Phi(x) \tag{10}$$

for $n \to \infty$.

Observe that with

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

then $E\overline{X}_n = \xi$, and by independence

$$V\overline{X}_n = \frac{1}{n^2} \sum_{i=1}^{n} VX_i = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Moreover,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i-\xi}{\sigma}=\frac{\sqrt{n}}{\sigma}(\overline{X}_n-\xi),$$

thus the Central Limit Theorem, as formulated in Theorem 9, is also a statement about the asymptotic distribution of the standardized average, and could have been formulated as

$$P\left(\overline{X}_n\leq\xi+x\frac{\sigma}{\sqrt{n}}\right)\to\Phi(x) \tag{11}$$

for $n\to\infty$. We usually say that $\overline{X}_n$ is asympticaly normally distributed with mean ξ and variance $\frac{\sigma^2}{n}$, and write

$$\overline{X}_n\overset{\text{as}}{\sim}\mathcal{N}\left(\xi,\frac{\sigma^2}{n}\right).$$

2.3 Applying the Central Limit Theorem

The practical importance of the CLT is due to the fact that it gives computable approximations to probabilities of interest concerning $\overline{X}_n$. It follows from (11) that

$$P\left(\left|\overline{X}_n-\xi\right|\leq x\frac{\sigma}{\sqrt{n}}\right)=P\left(\overline{X}_n\leq\xi+x\frac{\sigma}{\sqrt{n}}\right)-P\left(\overline{X}_n<\xi-x\frac{\sigma}{\sqrt{n}}\right)\to\Phi(x)-\Phi(-x).$$

A bound on $|\overline{X}_n-\xi|$ that holds with high probability can be computed by equating

$$\Phi(x)-\Phi(-x)=1-2\Phi(-x)$$

equal to the desired probability and then solve for x. A popular choice is $1-2\Phi(-x)=0.95$, or equivalently,

$$x=-\Phi^{-1}(0.025)=1.96.$$

If we want the bound to hold with probability 0.99 instead, the solution is

$$x=-\Phi^{-1}(0.005)=2.58.$$

Taking $x=1.96$ the conclusion from the CLT is that the absolute deviation of $\overline{X}_n$ from ξ is bounded by

$$1.96\frac{\sigma}{\sqrt{n}}$$

with approximate probability 0.95. The CLT implies that the approximation of the probability by 0.95 will become increasingly accurate as $n\to\infty$. Compare this with the bound from Chebychev's inequality, which can be restated as

$$P\left(\left|\overline{X}_n-\xi\right|\leq x\frac{\sigma}{\sqrt{n}}\right)\geq1-\frac{1}{x^2}.$$

Equating the lower probability bound equal to 0.95 gives $x = \sqrt{20} = 4.47$. The conclusion using Chebychev's inequality is, in principle, a little stronger – it assures that the absolute deviation is bounded by

$$4.47\frac{\sigma}{\sqrt{n}}$$

with probability greater than or equal to 0.95. This is a guarantied bound on the probability. The probability can be larger, but not smaller. The CLT provides an approximation, which can be smaller or larger than 0.95. Since $\Phi(-4.47) = 3.91 \times 10^{-6}$ the CLT suggests that the probability statement made by Chebychev's inequality is actually very conservative.

An obstacle to the practical use of the bounds provided by the CLT is that the standard deviation σ enters. It is rarely known. Just as $\overline{X}_n$ is an estimate of the (unknown) ξ from $X_1, \ldots, X_n$, we can also compute an estimate of σ. The usual estimate is

$$\hat{\sigma}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}.$$

Division by $n-1$ (and not n) assures that $E\hat{\sigma}_n^2 = \sigma^2$. A computable bound on $|\overline{X}_n - \xi|$ is then obtained by plugging in $\hat{\sigma}_n$, that is, the estimated bound becomes

$$1.96\frac{\hat{\sigma}_n}{\sqrt{n}}.$$

It is possible to show that

$$\frac{\hat{\sigma}_n}{\sigma} \xrightarrow{P} 1$$

for $n \to \infty$, and with a little more work it follows that also with the plug-in estimate of σ it holds that

$$P\left(\left|\overline{X}_n - \xi\right| \leq x\frac{\hat{\sigma}_n}{\sqrt{n}}\right) \to 1 - 2\Phi(-x) \tag{12}$$

for $n \to \infty$. See Problem 2.1 for an argument based on a finite fourth moment assumption.

The common way to report the bound is in terms of the interval

$$\left[\overline{X}_n - 1.96\frac{\hat{\sigma}_n}{\sqrt{n}}, \overline{X}_n + 1.96\frac{\hat{\sigma}_n}{\sqrt{n}}\right]$$

or equivalently

$$\overline{X}_n \pm 1.96\frac{\hat{\sigma}_n}{\sqrt{n}}.$$

This is a random interval, and the probability statement about this interval is that it will cover ξ with approximate probability 0.95. In the statistical terminology, it is a *confidence interval* for ξ, which has approximate 95% coverage.

2.4 Lindeberg's proof of the CLT

In everything that follows, $X_1, X_2, \ldots$ are independent and identically distributed real valued random variables with mean ξ and variance σ^2. We introduce the variables

$$Y_i = \frac{X_i - \xi}{\sqrt{n}\sigma}, \quad i = 1, \ldots, n,$$

which are then also independent and identically distributed but with mean 0 and variance n^{-1}. We let

$$S_n = \sum_{i=1}^{n} Y_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \xi}{\sigma},$$

such that $ES_n = 0$ and $VS_n = 1$.

The main idea in the proof is to replace the variables Y_i by normally distributed variables with the same mean and variance. To this end, let $Z_1, Z_2, \ldots$ be independent identically distributed with $Z_i \sim \mathcal{N}(0, n^{-1})$. Let

$$W_n = \sum_{i=1}^{n} Z_i \sim \mathcal{N}(0,1).$$

The variables $Z_1, Z_2, \ldots$ are, moreover, assumed independent of $Y_1, Y_2, \ldots$.

The proof of Theorem 9 is broken into two parts. First we give an intermediate result that states that

$$Eh(S_n) - Eh(W_n) \to 0$$

for all C^2-functions $h : \mathbb{R} \to \mathbb{R}$ with bounded and uniformly continuous second derivative. Theorem 9 would follow directly from this result with $h(y) = 1_{(-\infty,x]}(y)$ if this function were regular enough. Clearly, it is not, but it can easily be approximated by C^2-functions with a piecewise linear second derivative in such a way that we can transfer the convergence to indicator functions, and in this way get the desired convergence of distribution functions as stated in Theorem 9.

Lemma 10. *If h is C^2 with h'' uniformly continuous and bounded then*

$$Eh(S_n) - Eh(W_n) \to 0 \tag{13}$$

for $n \to \infty$.

Proof: By assumption, $|h''(x)| \leq M < \infty$ and from this it follows, by the mean value theorem, that $|h'(x)| \leq |h'(0)| + M|x|$. Likewise, $|h(x)| \leq h(0) + |h'(0)||x| + M|x|^2$, which shows that $h(S_n)$ as well as $h(W_n)$ have finite expectation. Moreover, by uniform continuity

$$\delta(\varepsilon) = \sup_{|x-y|<\varepsilon} |h''(x) - h''(y)|$$

fulfills that $\delta(\varepsilon) \to 0$ for $\varepsilon \to 0$, and by the bound on h'' it holds that $|\delta(\varepsilon)| \leq 2M$. Introduce the variables

$$S_i = \sum_{j=1}^{i} Y_j + \sum_{j=i+1}^{n} Z_j \quad \text{and} \quad U_i = \sum_{j=1}^{i-1} Y_j + \sum_{j=i+1}^{n} Z_j,$$

such that

$$S_i = U_i + Y_i \quad \text{and} \quad S_{i-1} = U_i + Z_i.$$

Then $S_0 = W_n$ and

$$h(S_n) - h(W_n) = \sum_{i=1}^{n} h(S_i) - h(S_{i-1}).$$

The idea in the proof is to bound the individual terms in the sum above. By a second order Taylor expansion of h with remainder it holds for all u and y that

$$h(u+y) = h(u) + h'(u)y + \frac{1}{2}h''(u+vy)y^2$$

for some $v \in [0,1]$. This gives that

$$\begin{aligned} h(S_i) - h(S_{i-1}) &= h(U_i + Y_i) - h(U_i + Z_i) \\ &= h'(U_i)(Y_i - Z_i) + \frac{1}{2}Y_i^2 h''(U_i + vY_i) - \frac{1}{2}Z_i^2 h''(U_i + v'Z_i) \end{aligned}$$

for $v, v' \in [0,1]$. Since h' is bounded by an affine function, $h'(U_i)$ has first moment, and since U_i and $Y_i - Z_i$ are independent, we get that $h'(U_i)$ and $Y_i - Z_i$ are independent. Hence

$$Eh'(U_i)(Y_i - Z_i) = Eh'(U_i)E(Y_i - Z_i) = 0.$$

From this it follows that

$$\begin{aligned} 2|Eh(S_i) - Eh(S_{i-1})| &= |EY_i^2 h''(U_i + vY_i) - EZ_i^2 h''(U_i + v'Z_i)| \\ &\leq EY_i^2 |h''(U_i + vY_i) - h''(U_i)| \\ &\quad + EZ_i^2 |h''(U_i + v'Z_i) - h''(U_i)| \\ &\quad + |EY_i^2 h''(U_i) - EZ_i^2 h''(U_i)| \\ &\leq EY_i^2 \delta(|Y_i|) + EZ_i^2 \delta(|Z_i|) \\ &\quad + |EY_i^2 h''(U_i) - EZ_i^2 h''(U_i)|. \end{aligned}$$

Since h'' is bounded, $h''(U_i)$ has first moment, and $h''(U_i)$ is independent of Y_i^2 as well as Z_i^2. Hence

$$EY_i^2 h''(U_i) = EY_i^2 Eh''(U_i) = EZ_i^2 Eh''(U_i) = EZ_i^2 h''(U_i),$$

and the last term above is 0. That is, we have

$$2|Eh(S_i) - Eh(S_{i-1})| \leq \underbrace{EY_i^2\delta(|Y_i|)}_{=EY_1^2\delta(|Y_1|)} + \underbrace{EZ_i^2\delta(|Z_i|)}_{=EZ_1^2\delta(|Z_1|)}.$$

This finally gives the bound

$$2|Eh(S_n) - Eh(W_n)| \leq \sum_{i=1}^{n} 2|Eh(S_i) - Eh(S_{i-1})| \leq nEY_1^2\delta(|Y_1|) + nEZ_1^2\delta(|Z_1|),$$

and the proof is completed by showing that the two terms on the right hand side converge to 0 for $n \to \infty$. They are dealt with in the same way, so lets consider the first. We have that

$$nY_1^2\delta(|Y_1|) = \frac{(X_1 - \xi)^2}{\sigma^2}\delta\left(\frac{|X_1 - \xi|}{\sqrt{n}\sigma}\right) \to 0$$

for $n \to \infty$, since

$$\frac{|X_1 - \xi|}{\sqrt{n}\sigma} \to 0$$

for $n \to \infty$. Moreover, this convergence is dominated by $2M\frac{(X_1-\xi)^2}{\sigma^2}$, which is integrable. It follows by dominated convergence that

$$nEY_1^2\delta(|Y_1|) \to 0$$

for $n \to \infty$, and this completes the proof. □

Note how the proof above uses independence in two crucial places to get expectation 0 of products. We could not control the error bounds if these terms did not completely disappear, and for this purpose the independence assumption plays an important role.

Note also that the distribution of W_n does not depend upon n. Thus $Eh(W_n)$ is a constant. The conclusion of Lemma 10 could thus just as well have been phrased as

$$Eh(S_n) \to Eh(W) \tag{14}$$

for $W \sim \mathcal{N}(0, 1)$.

To prove Theorem 9 from Lemma 10 we need to construct approximations of indicator functions by a C^2-function with bounded and uniformly continuous second derivative. To this end we start with the continuous piecewise linear function

$$g_\varepsilon(y) = \begin{cases} 0 & \text{if } y \leq 0 \\ -y & \text{if } 0 < y \leq \varepsilon/4 \\ -\varepsilon/2 + y & \text{if } \varepsilon/4 < y \leq 3\varepsilon/4 \\ \varepsilon - y & \text{if } 3\varepsilon/4 < y \leq \varepsilon \\ 0 & \text{if } \varepsilon < y \end{cases}$$

Then we let $G_\varepsilon(y) = \int_{-\infty}^y g_\varepsilon(z)\mathrm{d}z$. We note that $G_\varepsilon(y) \leq 0$ and $G_\varepsilon(y) < 0$ if and only if $y \in (0, \varepsilon)$. We introduce $b(\varepsilon) = \int_{-\infty}^{\infty} |G_\varepsilon(y)|\mathrm{d}y$ and

$$h_\varepsilon(y) = \int_{-\infty}^{y} \frac{G_\varepsilon(z)}{b(\varepsilon)}\mathrm{d}z + 1.$$

Then $h_\varepsilon(y) = 1$ for $y \leq 0$, h_ε decreases from 1 to 0 on the interval $(0, \varepsilon)$, and $h_\varepsilon(y) = 0$ for $y \geq \varepsilon$. Moreover, it is constructed so that it is C^2 with the second derivative, g_ε, being bounded and uniformly continuous. The function h_ε is a regular approximation to the indicator function $1_{(-\infty,0]}$. Using this function we can now give a proof of Theorem 9.

Proof of Theorem 9: We let F_n denote the distribution function for the distribution of S_n. That is,

$$F_n(x) = P(S_n \leq x) = E1_{(-\infty,x]}(S_n).$$

Fixing $x \in \mathbb{R}$ we have that

$$h_\varepsilon(y - x + \varepsilon) \leq 1_{(-\infty,x]}(y) \leq h_\varepsilon(y - x).$$

Plugging S_n in for y, taking expectations and limits it follows from Lemma 10 that

$$Eh_\varepsilon(W - x + \varepsilon) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq Eh_\varepsilon(W - x)$$

where $W \sim \mathcal{N}(0, 1)$. Now, as

$$1_{(-\infty,x-\varepsilon]}(y) \leq h_\varepsilon(y - x + \varepsilon) \quad \text{and} \quad h_\varepsilon(y - x) \leq 1_{(-\infty,x+\varepsilon]}(y)$$

it follows that

$$\Phi(x - \varepsilon) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq \Phi(x + \varepsilon).$$

Letting $\varepsilon \to 0$, and using that Φ is continuous in x, we conclude that $F_n(x)$ converges and that

$$\lim_{n\to\infty} F_n(x) = \Phi(x).$$

□

We conclude this note with a few comments on how the CLT can be proved by other methods, and how it fits into a more general framework of convergence in distribution. We generally say that S_n converges in distribution to $W \sim \mathcal{N}(0, 1)$ if (14) holds for all bounded continuous functions h, and we write

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \xi}{\sigma} \xrightarrow{\mathcal{D}} W$$

or

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i-\xi}{\sigma}\xrightarrow{\mathcal{D}}\mathcal{N}(0,1).$$

The proof of Theorem 9 shows that convergence in distribution to the normally distributed random variable W implies pointwise convergence of the corresponding distribution functions. We should note that the argument relies in an essential way on the fact that the limit distribution function Φ is continuous everywhere.

We haven't actually established convergence in distribution completely. Lemma 10 is almost a proof of convergence in distribution of S_n towards W except that h is assumed a little more regular than just continuous. In the general theory on convergence of distribution it is usually shown explicitly that proving a result like Lemma 10 is enough to ensure that the convergence holds for all bounded continuous h, but we will not pursue this here. The conclusion, that we have established, on the convergence of the distribution functions is what matters for practical applications.

An alternative to the substitution argument given above is based on a more analytic method. It can first of all be shown that the convergence of the expectations (14) for all continuous h follows if we can just check the convergence for some very special C^∞ functions h, namely $h(x) = \cos(\theta x)$ and $h(x) = \sin(\theta x)$ for $\theta \in \mathbb{R}$. The *characteristic functions* of S_n is

$$\phi_n(\theta) = E\cos(\theta S_n) + iE\sin(\theta S_n) = Ee^{i\theta S_n}.$$

The CLT is then proved by showing that

$$\phi_n(\theta) \to e^{-\theta^2/2} = Ee^{i\theta W}$$

for all $\theta \in \mathbb{R}$, and this is enough to establish convergence in distribution of S_n toward $W \sim \mathcal{N}(0,1)$. This approach binds the theory of convergence in distribution together with characteristic functions (or Fourier transforms) of probability measures.

A third, and very different, approach was first given by Stanford statistician Charles Stein in his seminal 1972 paper *A bound for the error in the normal approximation to the distribution of a sum of dependent variables.* Stein's method is centered around Stein's differential equation, which is the first order linear inhomogeneous differential equation

$$f'(s) - sf(s) = h(s) - Eh(W)$$

for $W \sim \mathcal{N}(0,1)$ and h bounded and continuous. From this equation we get that

$$|Eh(S_n) - Eh(W)| = |Ef'(S_n) - ES_n f(S_n)|,$$

and a detailed analysis of the regularity of the solution f', combined with clever arguments involving $Ef'(S_n) - ES_n f(S_n)$, give Laplace's CLT and much more. As the title of Stein's original paper indicates, this approach is useful also when handling sums of dependent random variables.

2.5 Problems

2.1. Assume that $X_1, X_2, \ldots$ have finite fourth moment. Use the Law of Large Numbers to show that

$$\hat{\sigma}^2 \xrightarrow{P} \sigma^2$$

for $n \to \infty$. Introduce the interval $I(\varepsilon) = \left[\sqrt{1-\varepsilon}, \sqrt{1+\varepsilon}\right]$ for $\varepsilon \in (0,1)$, and conclude that for all $\varepsilon \in (0,1)$

$$P\left(\frac{\hat{\sigma}}{\sigma} \notin I(\varepsilon)\right) \to 0$$

for $n \to \infty$. Argue that

$$P\left(\left|\overline{X}_n - \xi\right| \leq x\frac{\hat{\sigma}_n}{\sqrt{n}}\right) = P\left(\left|\overline{X}_n - \xi\right| \leq x\frac{\hat{\sigma}_n}{\sqrt{n}}, \frac{\hat{\sigma}}{\sigma} \notin I(\varepsilon)\right) + P\left(\left|\overline{X}_n - \xi\right| \leq x\frac{\hat{\sigma}_n}{\sqrt{n}}, \frac{\hat{\sigma}}{\sigma} \in I(\varepsilon)\right),$$

and show that the first term tends to 0 for $n \to \infty$, while the second term implies that

$$\begin{aligned} 1 - 2\Phi(-x\sqrt{1-\varepsilon}) &\leq \liminf_{n\to\infty} P\left(\left|\overline{X}_n - \xi\right| \leq x\frac{\hat{\sigma}_n}{\sqrt{n}}\right) \\ &\leq \limsup_{n\to\infty} P\left(\left|\overline{X}_n - \xi\right| \leq x\frac{\hat{\sigma}_n}{\sqrt{n}}\right) \leq 1 - 2\Phi(-x\sqrt{1+\varepsilon}). \end{aligned}$$

Conclude that (12) holds.