# 11

# Projections

*A projection of a random variable is defined as a closest element in a given set of functions. We can use projections to derive the asymptotic distribution of a sequence of variables by comparing these to projections of a simple form. Conditional expectations are special projections. The Hájek projection is a sum of independent variables; it is the leading term in the Hoeffding decomposition.*

## 11.1 Projections

A common method to derive the limit distribution of a sequence of statistics $T_n$ is to show that it is asymptotically equivalent to a sequence $S_n$ of which the limit behavior is known. The basis of this method is Slutsky's lemma, which shows that the sequence $T_n = T_n - S_n + S_n$ converges in distribution to $S$ if both $T_n - S_n \xrightarrow{\text{P}} 0$ and $S_n \rightsquigarrow S$.

How do we find a suitable sequence $S_n$? First, the variables $S_n$ must be of a simple form, because the limit properties of the sequence $S_n$ must be known. Second, $S_n$ must be close enough. One solution is to search for the closest $S_n$ of a certain predetermined form. In this chapter, "closest" is taken as closest in square expectation.

Let $T$ and $\{S : S \in \mathcal{S}\}$ be random variables (defined on the same probability space) with finite second-moments. A random variable $\hat{S}$ is called a *projection* of $T$ onto $\mathcal{S}$ (or $L_2$-projection) if $\hat{S} \in \mathcal{S}$ and minimizes

$$S \mapsto \mathrm{E}(T - S)^2, \qquad S \in \mathcal{S}.$$

Often $\mathcal{S}$ is a linear space in the sense that $\alpha_1 S_1 + \alpha_2 S_2$ is in $\mathcal{S}$ for every $\alpha_1, \alpha_2 \in R$, whenever $S_1, S_2 \in \mathcal{S}$. In this case $\hat{S}$ is the projection of $T$ if and only if $T - \hat{S}$ is *orthogonal* to $\mathcal{S}$ for the inner product $\langle S_1, S_2 \rangle = \mathrm{E}S_1 S_2$. This is the content of the following theorem.

**11.1** ***Theorem.*** *Let $\mathcal{S}$ be a linear space of random variables with finite second moments. Then $\hat{S}$ is the projection of $T$ onto $\mathcal{S}$ if and only if $\hat{S} \in \mathcal{S}$ and*

$$\mathrm{E}(T - \hat{S})S = 0, \qquad \textit{every } S \in \mathcal{S}.$$

*Every two projections of $T$ onto $\mathcal{S}$ are almost surely equal. If the linear space $\mathcal{S}$ contains the constant variables, then $\mathrm{E}T = \mathrm{E}\hat{S}$ and $\mathrm{cov}(T - \hat{S}, S) = 0$ for every $S \in \mathcal{S}$.*

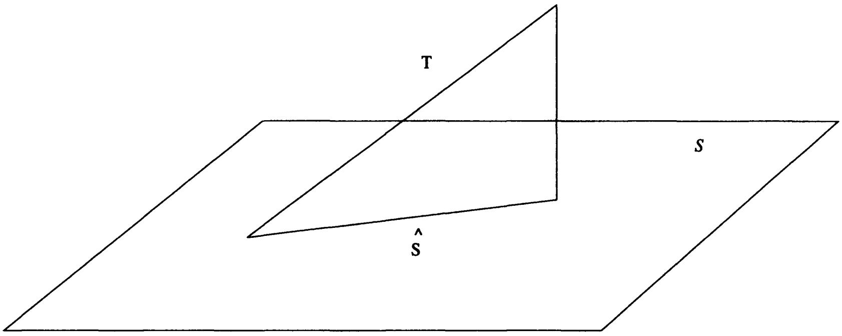

**Figure 11.1.** A variable $T$ and its projection $\hat{S}$ on a linear space.

***Proof.*** For any $S$ and $\hat{S}$ in $\mathcal{S}$,

$$E(T-S)^2 = E(T-\hat{S})^2 + 2E(T-\hat{S})(\hat{S}-S) + E(\hat{S}-S)^2.$$

If $\hat{S}$ satisfies the orthogonality condition, then the middle term is zero, and we conclude that $E(T-S)^2 \geq E(T-\hat{S})^2$, with strict inequality unless $E(\hat{S}-S)^2 = 0$. Thus, the orthogonality condition implies that $\hat{S}$ is a projection, and also that it is unique.

Conversely, for any number $\alpha$,

$$E(T-\hat{S}-\alpha S)^2 - E(T-\hat{S})^2 = -2\alpha E(T-\hat{S})S + \alpha^2 ES^2.$$

If $\hat{S}$ is a projection, then this expression is nonnegative for every $\alpha$. But the parabola $\alpha \mapsto \alpha^2 ES^2 - 2\alpha E(T-\hat{S})S$ is nonnegative if and only if the orthogonality condition $E(T-\hat{S})S = 0$ is satisfied.

If the constants are in $\mathcal{S}$, then the orthogonality condition implies $E(T-\hat{S})c = 0$, whence the last assertions of the theorem follow. ■

The theorem does not assert that projections always exist. This is not true: The infimum $\inf_S E(T-S)^2$ need not be achieved. A sufficient condition for existence is that $\mathcal{S}$ is closed for the second-moment norm, but existence is usually more easily established directly.

The orthogonality of $T-\hat{S}$ and $\hat{S}$ yields the Pythagorean rule $ET^2 = E(T-\hat{S})^2 + E\hat{S}^2$. (See Figure 11.1.) If the constants are contained in $\mathcal{S}$, then this is also true for variances instead of second moments.

Now suppose a sequence of statistics $T_n$ and linear spaces $\mathcal{S}_n$ is given. For each $n$, let $\hat{S}_n$ be the projection of $T_n$ on $\mathcal{S}_n$. Then the limiting behavior of the sequence $T_n$ follows from that of $\hat{S}_n$, and vice versa, provided the quotient $\mathrm{var}T_n/\mathrm{var}\hat{S}_n$ converges to 1.

**11.2** ***Theorem.*** *Let $\mathcal{S}_n$ be linear spaces of random variables with finite second moments that contain the constants. Let $T_n$ be random variables with projections $\hat{S}_n$ onto $\mathcal{S}_n$. If $\mathrm{var}T_n/\mathrm{var}\hat{S}_n \to 1$ then*

$$\frac{T_n - ET_n}{\mathrm{sd}\,T_n} - \frac{\hat{S}_n - E\hat{S}_n}{\mathrm{sd}\,\hat{S}_n} \xrightarrow{P} 0.$$

***Proof.*** We shall prove convergence in second mean, which is stronger. The expectation of the difference is zero. Its variance is equal to

$$2 - 2\frac{\operatorname{cov}(T_n, \hat{S}_n)}{\operatorname{sd} T_n \operatorname{sd} \hat{S}_n}.$$

By the orthogonality of $T_n - \hat{S}_n$ and $\hat{S}_n$, it follows that $\mathrm{E}T_n\hat{S}_n = \mathrm{E}\hat{S}_n^2$. Because the constants are in $\mathcal{S}_n$, this implies that $\operatorname{cov}(T_n, \hat{S}_n) = \operatorname{var}\hat{S}_n$, and the theorem follows. ■

The condition $\operatorname{var}T_n/\operatorname{var}\hat{S}_n \to 1$ in the theorem implies that the projections $\hat{S}_n$ are asymptotically of the same size as the original $T_n$. This explains that "nothing is lost" in the limit, and that the difference between $T_n$ and its projection converges to zero. In the preceding theorem it is essential that the $\hat{S}_n$ are the projections of the variables $T_n$, because the condition $\operatorname{var}T_n/\operatorname{var}S_n \to 1$ for general sequences $S_n$ and $T_n$ does not imply anything.

## 11.2 Conditional Expectation

The expectation $\mathrm{E}X$ of a random variable $X$ minimizes the quadratic form $a \mapsto \mathrm{E}(X - a)^2$ over the real numbers $a$. This may be expressed as follows: $\mathrm{E}X$ is the best prediction of $X$, given a quadratic loss function, and in the absence of additional information.

The *conditional expectation* $\mathrm{E}(X \mid Y)$ of a random variable $X$ given a random vector $Y$ is defined as the best "prediction" of $X$ given knowledge of $Y$. Formally, $\mathrm{E}(X \mid Y)$ is a measurable function $g_0(Y)$ of $Y$ that minimizes

$$\mathrm{E}\big(X - g(Y)\big)^2$$

over all measurable functions $g$. In the terminology of the preceding section, $\mathrm{E}(X \mid Y)$ is the projection of $X$ onto the linear space of all measurable functions of $Y$. It follows that the conditional expectation is the unique measurable function $\mathrm{E}(X \mid Y)$ of $Y$ that satisfies the orthogonality relation

$$\mathrm{E}\big(X - \mathrm{E}(X \mid Y)\big)g(Y) = 0, \qquad \text{every } g.$$

If $\mathrm{E}(X \mid Y) = g_0(Y)$, then it is customary to write $\mathrm{E}(X \mid Y = y)$ for $g_0(y)$. This is interpreted as the expected value of $X$ given that $Y = y$ is observed. By Theorem 11.1 the projection is unique only up to changes on sets of probability zero. This means that the function $g_0(y)$ is unique up to sets $B$ of values $y$ such that $\mathrm{P}(Y \in B) = 0$. (These could be very big sets.)

The following examples give some properties and also describe the relationship with conditional densities.

**11.3** ***Example.*** The orthogonality relationship with $g \equiv 1$ yields the formula $\mathrm{E}X = \mathrm{E}\mathrm{E}(X \mid Y)$. Thus, "the expectation of a conditional expectation is the expectation." □

**11.4** ***Example.*** If $X = f(Y)$ for a measurable function $f$, then $\mathrm{E}(X \mid Y) = X$. This follows immediately from the definition, in which the minimum can be reduced to zero. The interpretation is that $X$ is perfectly predictable given knowledge of $Y$. □

**11.5** ***Example.*** Suppose that $(X, Y)$ has a joint probability density $f(x, y)$ with respect to a $\sigma$-finite product measure $\mu \times \nu$, and let $f(x \mid y) = f(x, y)/f_Y(y)$ be the conditional density of $X$ given $Y = y$. Then

$$\mathrm{E}(X \mid Y) = \int x f(x \mid Y)\, d\mu(x).$$

(This is well defined only if $f_Y(Y) > 0$.) Thus the conditional expectation as defined above concurs with our intuition.

The formula can be established by writing

$$\mathrm{E}\bigl(X - g(Y)\bigr)^2 = \int \Bigl[\int \bigl(x - g(y)\bigr)^2 f(x \mid y)\, d\mu(x)\Bigr] f_Y(y)\, d\nu(y).$$

To minimize this expression over $g$, it suffices to minimize the inner integral (between square brackets) by choosing the value of $g(y)$ for every $y$ separately. For each $y$, the integral $\int (x - a)^2 f(x \mid y)\, d\mu(x)$ is minimized for $a$ equal to the mean of the density $x \mapsto f(x \mid y)$. □

**11.6** ***Example.*** If $X$ and $Y$ are independent, then $\mathrm{E}(X \mid Y) = \mathrm{E}X$. Thus, the extra knowledge of an unrelated variable $Y$ does not change the expectation of $X$.

The relationship follows from the fact that independent random variables are uncorrelated: Because $\mathrm{E}(X - \mathrm{E}X)g(Y) = 0$ for all $g$, the orthogonality relationship holds for $g_0(Y) = \mathrm{E}X$. □

**11.7** ***Example.*** If $f$ is measurable, then $\mathrm{E}\bigl(f(Y)X \mid Y\bigr) = f(Y)\mathrm{E}(X \mid Y)$ for any $X$ and $Y$. The interpretation is that, given $Y$, the factor $f(Y)$ behaves like a constant and can be "taken out" of the conditional expectation.

Formally, the rule can be established by checking the orthogonality relationship. For every measurable function $g$,

$$\mathrm{E}\bigl(f(Y)X - f(Y)\mathrm{E}(X \mid Y)\bigr)\, g(Y) = \mathrm{E}\bigl(X - \mathrm{E}(X \mid Y)\bigr)\, f(Y)g(Y) = 0,$$

because $X - \mathrm{E}(X \mid Y)$ is orthogonal to all measurable functions of $Y$, including those of the form $f(Y)g(Y)$. Because $f(Y)\mathrm{E}(X \mid Y)$ is a measurable function of $Y$, it must be equal to $\mathrm{E}\bigl(f(Y)X \mid Y\bigr)$. □

**11.8** ***Example.*** If $X$ and $Y$ are independent, then $\mathrm{E}\bigl(f(X, Y) \mid Y = y\bigr) = \mathrm{E}f(X, y)$ for every measurable $f$. This rule may be remembered as follows: The known value $y$ is substituted for $Y$; next, because $Y$ carries no information concerning $X$, the unconditional expectation is taken with respect to $X$.

The rule follows from the equality

$$\mathrm{E}\bigl(f(X, Y) - g(Y)\bigr)^2 = \int\!\!\int \bigl(f(x, y) - g(y)\bigr)^2\, dP_X(x)\, dP_Y(y).$$

Once again, this is minimized over $g$ by choosing for each $y$ separately the value $g(y)$ to minimize the inner integral. □

**11.9** ***Example.*** For any random vectors $X$, $Y$ and $Z$,

$$\mathrm{E}\bigl(\mathrm{E}(X \mid Y, Z) \mid Y\bigr) = \mathrm{E}(X \mid Y).$$

This expresses that a projection can be carried out in steps: The projection onto a smaller set can be obtained by projecting the projection onto a bigger set a second time.

Formally, the relationship can be proved by verifying the orthogonality relationship $\mathrm{E}\big(\mathrm{E}(X \mid Y, Z) - \mathrm{E}(X \mid Y)\big)g(Y) = 0$ for all measurable functions $g$. By Example 11.7, the left side of this equation is equivalent to $\mathrm{E}\mathrm{E}\big(Xg(Y) \mid Y, Z\big) - \mathrm{E}\mathrm{E}\big(g(Y)X \mid Y\big) = 0$, which is true because conditional expectations retain expectations. □

## 11.3 Projection onto Sums

Let $X_1, \dots, X_n$ be independent random vectors, and let $\mathcal{S}$ be the set of all variables of the form

$$\sum_{i=1}^n g_i(X_i),$$

for arbitrary measurable functions $g_i$ with $\mathrm{E}g_i^2(X_i) < \infty$. This class is of interest, because the convergence in distribution of the sums can be derived from the central limit theorem. The projection of a variable onto this class is known as its *Hájek projection.*

**11.10** ***Lemma.*** *Let $X_1, \dots, X_n$ be independent random vectors. Then the projection of an arbitrary random variable $T$ with finite second moment onto the class $\mathcal{S}$ is given by*

$$\hat{S} = \sum_{i=1}^n \mathrm{E}(T \mid X_i) - (n-1)\mathrm{E}T.$$

***Proof.*** The random variable on the right side is certainly an element of $\mathcal{S}$. Therefore, the assertion can be verified by checking the orthogonality relation. Because the variables $X_i$ are independent, the conditional expectation $\mathrm{E}\big(\mathrm{E}(T \mid X_i) \mid X_j\big)$ is equal to the expectation $\mathrm{E}\mathrm{E}(T \mid X_i) = \mathrm{E}T$ for every $i \neq j$. Consequently, $\mathrm{E}(\hat{S} \mid X_j) = \mathrm{E}(T \mid X_j)$ for every $j$, whence

$$\mathrm{E}(T - \hat{S})g_j(X_j) = \mathrm{E}\mathrm{E}(T - \hat{S} \mid X_j)g_j(X_j) = \mathrm{E}0g_j(X_j) = 0.$$

This shows that $T - \hat{S}$ is orthogonal to $\mathcal{S}$. ■

Consider the special case that $X_1, \dots, X_n$ are not only independent but also identically distributed, and that $T = T(X_1, \dots, X_n)$ is a permutation-symmetric, measurable function of the $X_i$. Then

$$\mathrm{E}(T \mid X_i = x) = \mathrm{E}T(x, X_2, \dots, X_n).$$

Because this does not depend on $i$, the projection $\hat{S}$ is also the projection of $T$ onto the smaller set of variables of the form $\sum_{i=1}^n g(X_i)$, where $g$ is an arbitrary measurable function.

## *11.4 Hoeffding Decomposition

The Hájek projection gives a best approximation by a sum of functions of one $X_i$ at a time. The approximation can be improved by using sums of functions of two, or more, variables. This leads to the *Hoeffding decomposition.*

Because a projection onto a sum of orthogonal spaces is the sum of the projections onto the individual spaces, it is convenient to decompose the proposed projection space into a sum of orthogonal spaces. Given independent variables $X_1, \dots, X_n$ and a subset $A \subset \{1, \dots, n\}$, let $H_A$ denote the set of all square-integrable random variables of the type

$$g_A(X_i : i \in A),$$

for measurable functions $g_A$ of $|A|$ arguments such that

$$\mathrm{E}\big(g_A(X_i : i \in A) \mid X_j : j \in B\big) = 0, \qquad \text{every } B : |B| < |A|.$$

(Define $\mathrm{E}(T \mid \emptyset) = \mathrm{E}T$.) By the independence of $X_1, \dots, X_n$ the condition in the last display is automatically valid for any $B \subset \{1, 2, \dots, n\}$ that does not contain $A$. Consequently, the spaces $H_A$, when $A$ ranges over all subsets of $\{1, \dots, n\}$, are pairwise orthogonal. Stated in its present form, the condition reflects the intention to build approximations of increasing complexity by projecting a given variable in turn onto the spaces

$$[1], \qquad \Big[\sum_i g_{\{i\}}(X_i)\Big], \qquad \Big[\sum\sum_{i<j} g_{\{i,j\}}(X_i, X_j)\Big], \qquad \dots,$$

where $g_{\{i\}} \in H_{\{i\}}$, $g_{\{i,j\}} \in H_{\{i,j\}}$, and so forth. Each new space is chosen orthogonal to the preceding spaces.

Let $P_A T$ denote the projection of $T$ onto $H_A$. Then, by the orthogonality of the $H_A$, the projection onto the sum of the first $r$ spaces is the sum $\sum_{|A| \le r} P_A T$ of the projections onto the individual spaces. The projection onto the sum of the first two spaces is the Hájek projection. More generally, the projections of zero, first, and second order can be seen to be

$$\begin{aligned}
P_\emptyset T &= \mathrm{E}T,\\
P_{\{i\}} T &= \mathrm{E}(T \mid X_i) - \mathrm{E}T,\\
P_{\{i,j\}} T &= \mathrm{E}(T \mid X_i, X_j) - \mathrm{E}(T \mid X_i) - \mathrm{E}(T \mid X_j) + \mathrm{E}T.
\end{aligned}$$

Now the general formula given by the following lemma should not be surprising.

**11.11** ***Lemma.*** *Let $X_1, \dots, X_n$ be independent random variables, and let $T$ be an arbitrary random variable with $\mathrm{E}T^2 < \infty$. Then the projection of $T$ onto $H_A$ is given by*

$$P_A T = \sum_{B \subset A} (-1)^{|A| - |B|} \mathrm{E}(T \mid X_i : i \in B).$$

*If $T \perp H_B$ for every subset $B \subset A$ of a given set $A$, then $\mathrm{E}(T \mid X_i : i \in A) = 0$. Consequently, the sum of the spaces $H_B$ with $B \subset A$ contains all square-integrable functions of $(X_i : i \in A)$.*

***Proof.*** Abbreviate $\mathrm{E}(T \mid X_i : i \in A)$ to $\mathrm{E}(T \mid A)$ and $g_A(X_i : i \in A)$ to $g_A$. By the independence of $X_1, \dots, X_n$ it follows that $\mathrm{E}\big(\mathrm{E}(T \mid A) \mid B\big) = \mathrm{E}(T \mid A \cap B)$ for every subsets $A$

and $B$ of $\{1,\dots,n\}$. Thus, for $P_AT$ as defined in the lemma and a set $C$ strictly contained in $A$,

$$\begin{aligned}\mathrm{E}(P_AT\mid C)&=\sum_{B\subset A}(-1)^{|A|-|B|}\mathrm{E}(T\mid B\cap C)\\&=\sum_{D\subset C}\sum_{j=0}^{|A|-|C|}(-1)^{|A|-|D|-j}\binom{|A|-|C|}{j}\mathrm{E}(T\mid D).\end{aligned}$$

By the binomial formula, the inner sum is zero for every $D$. Thus the left side is zero. In view of the form of $P_AT$, it was not a loss of generality to assume that $C\subset A$. Hence $P_AT$ is contained in $H_A$.

Next we verify the orthogonality relationship. For any measurable function $g_A$,

$$\mathrm{E}(T-P_AT)g_A=\mathrm{E}\bigl(T-\mathrm{E}(T\mid A)\bigr)g_A-\sum_{\substack{B\subset A\\ B\neq A}}(-1)^{|A|-|B|}\mathrm{E}\mathrm{E}(T\mid B)\mathrm{E}(g_A\mid B).$$

This is zero for any $g_A\in H_A$. This concludes the proof that $P_AT$ is as given.

We prove the second assertion of the lemma by induction on $r=|A|$. If $T\perp H_\emptyset$, then $\mathrm{E}(T\mid\emptyset)=\mathrm{E}T=0$. Thus the assertion is true for $r=0$. Suppose that it is true for $0,\dots,r-1$, and consider a set $A$ of $r$ elements. If $T\perp H_B$ for every $B\subset A$, then certainly $T\perp H_C$ for every $C\subset B$. Consequently, the induction hypothesis shows that $\mathrm{E}(T\mid B)=0$ for every $B\subset A$ of $r-1$ or fewer elements. The formula for $P_AT$ now shows that $P_AT=\mathrm{E}(T\mid A)$. By assumption the left side is zero. This concludes the induction argument.

The final assertion of the lemma follows if the variable $T_A:=T-\sum_{B\subset A}P_BT$ is zero for every $T$ that depends on $(X_i:i\in A)$ only. But in this case $T_A$ depends on $(X_i:i\in A)$ only and hence equals $\mathrm{E}(T_A\mid A)$, which is zero, because $T_A\perp H_B$ for every $B\subset A$. ■

If $T=T(X_1,\dots,X_n)$ is permutation-symmetric and $X_1,\dots,X_n$ are independent and identically distributed, then the Hoeffding decomposition of $T$ can be simplified to

$$T=\sum_{r=0}^{n}\sum_{|A|=r}g_r(X_i:i\in A),$$

for

$$g_r(x_1,\dots,x_r)=\sum_{B\subset\{1,\dots,r\}}(-1)^{r-|B|}\mathrm{E}T(x_i\in B,X_i\notin B).$$

The inner sum in the representation of $T$ is for each $r$ a $U$-statistic of order $r$ (as discussed in the Chapter 12), with degenerate kernel. All terms in the sum are orthogonal, whence the variance of $T$ can be found as $\mathrm{var}T=\sum_{r=1}^n\binom{n}{r}\mathrm{E}g_r^2(X_1,\dots,X_r)$.

## Notes

Orthogonal projections in Hilbert spaces (complete inner product spaces) are a classical subject in functional analysis. We have limited our discussion to the Hilbert space $L_2(\Omega,\mathcal{U},\mathrm{P})$ of all square-integrable random variables on a probability space. Another popular method to introduce conditional expectation is based on the Radon-Nikodym theorem. Then $\mathrm{E}(X \mid Y)$ is naturally defined for every integrable $X$. Hájek stated his projection lemma in [68] when proving the asymptotic normality of rank statistics under alternatives. Hoeffding [75] had already used it implicitly when proving the asymptotic normality of $U$-statistics. The "Hoeffding" decomposition appears to have received its name (for instance in [151]) in honor of Hoeffding's 1948 paper, but we have not been able to find it there. It is not always easy to compute a projection or its variance, and, if applied to a sequence of statistics, a projection may take the form $\sum g_n(X_i)$ for a function $g_n$ depending on $n$ even though a simpler approximation of the form $\sum g(X_i)$ with $g$ fixed is possible.

## PROBLEMS

**1.** Show that "projecting decreases second moment": If $\hat{S}$ is the projection of $T$ onto a linear space, then $\mathrm{E}\hat{S}^2 \leq \mathrm{E}T^2$. If $\mathcal{S}$ contains the constants, then also $\mathrm{var}\hat{S} \leq \mathrm{var}T$.

**2.** Another idea of projection is based on minimizing variance instead of second moment. Show that $\mathrm{var}(T - S)$ is minimized over a linear space $\mathcal{S}$ by $\hat{S}$ if and only if $\mathrm{cov}(T - \hat{S}, S) = 0$ for every $S \in \mathcal{S}$.

**3.** If $X \geq Y$ almost surely, then $\mathrm{E}(X \mid Z) \geq \mathrm{E}(Y \mid Z)$.

**4.** For an arbitrary random variable $X \geq 0$ (not necessarily square-integrable), define a conditional expectation $\mathrm{E}(X \mid Y)$ by $\lim_{M\to\infty} \mathrm{E}(X \wedge M \mid Y)$.

(i) Show that this is well defined (the limit exists almost surely).

(ii) Show that this coincides with the earlier definition if $\mathrm{E}X^2 < \infty$.

(iii) If $\mathrm{E}X < \infty$ show that $\mathrm{E}\big(X - \mathrm{E}(X \mid Y)\big)g(Y) = 0$ for every bounded, measurable function $g$.

(iv) Show that $\mathrm{E}(X \mid Y)$ is the almost surely unique measurable function of $Y$ that satisfies the orthogonality relationship of (iii).

How would you define $\mathrm{E}(X \mid Y)$ for a random variable with $\mathrm{E}|X| < \infty$?

**5.** Show that a projection $\hat{S}$ of a variable $T$ onto a convex set $\mathcal{S}$ is almost surely unique.

**6.** Find the conditional expectation $\mathrm{E}(X \mid Y)$ if $(X, Y)$ possesses a bivariate normal distribution.

**7.** Find the conditional expectation $\mathrm{E}(X_1 \mid X_{(n)})$ if $X_1, \ldots, X_n$ are a random sample of standard uniform variables.

**8.** Find the conditional expectation $\mathrm{E}(X_1 \mid \overline{X}_n)$ if $X_1, \ldots, X_n$ are i.i.d.

**9.** Show that for any random variables $S$ and $T$ (i) $\mathrm{sd}(S + T) \leq \mathrm{sd}\, S + \mathrm{sd}\, T$, and (ii) $|\,\mathrm{sd}\, S - \mathrm{sd}\, T\,| \leq \mathrm{sd}(S - T)$.

**10.** If $S_n$ and $T_n$ are arbitrary sequences of random variables such that $\mathrm{var}(S_n - T_n)/\mathrm{var}T_n \to 0$, then

$$\frac{S_n - \mathrm{E}S_n}{\mathrm{sd}\, S_n} - \frac{T_n - \mathrm{E}T_n}{\mathrm{sd}\, T_n} \xrightarrow{\mathrm{P}} 0.$$

Moreover, $\mathrm{var}S_n/\mathrm{var}T_n \to 1$. Show this.

**11.** Show that $P_A h(X_j : X_j \in B) = 0$ for every set $B$ that does not contain $A$.