2
Statistical inference

Chapter 1 introduced the notion of an expectation as my 'best guess' for the value of some operationally-defined random quantity. In order to qualify as expectations (i.e. to be coherent) my collection of guesses must have the axiomatic properties of lower boundedness (which applies one random quantity at a time) and additivity (which applies for collections of random quantities); see Def. 1.4.

Def. 1.3 provided an informal definition of statistical inference, covering 'coherence' and 'extension', and appended the comment that many of my expectations might be 'undecided'. This short chapter provides a complete description of statistical inference and explains how 'undecided' expectations arise. The key result is the Fundamental Theorem of Prevision (FTP, Thm 1.6). Sec. 2.1 presents the FTP again, but in a different form, establishing the conditions under which a set of expectations is coherent. Sec. 2.2 shows how a set of expectations may be extended. Sec. 2.3 discusses the various ways in which beliefs can be quantified, and Sec. 2.4 provides a general representation of data.

2.1 *The FTP again, coherence*

Recall the FTP in Thm 1.6. Here is another statement of it. As before, denote the joint realm of $\boldsymbol{X} := (X_1, \dots, X_m)$ as

$$\mathcal{X} := \left\{ \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(s)} \right\} \qquad \boldsymbol{x}^{(j)} \in \mathbb{R}^m.$$

Now let $\{g_1, \dots, g_k\}$ be a specified set of functions of $\boldsymbol{x}$, and construct the $(k \times s)$ matrix

$$\{G\}_{ij} := g_i(\boldsymbol{x}^{(j)}).$$

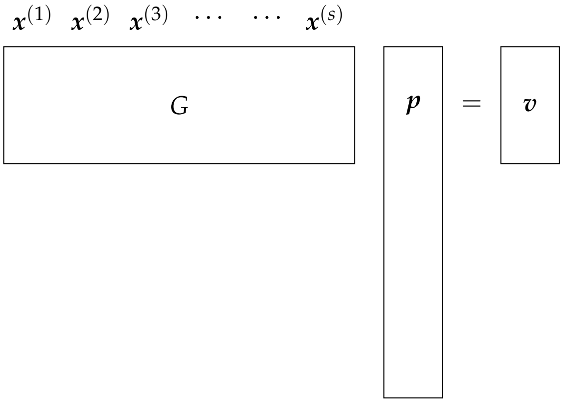
Let $\boldsymbol{v} := (v_1, \dots, v_k)$. According to the FTP,

$$\begin{aligned} \mathrm{E}\{g_1(\boldsymbol{X})\} &\leftarrow v_1 \\ &\vdots \\ \mathrm{E}\{g_k(\boldsymbol{X})\} &\leftarrow v_k \end{aligned} \tag{2.1}$$

is a coherent set of expectations if and only if

$$G\boldsymbol{p} = \boldsymbol{v} \quad \text{for some } \boldsymbol{p} \in \mathbb{S}^{s-1}. \tag{2.2}$$

This matrix equation can be visualised as



illustrating that G is typically 'short and fat', i.e. I have fewer beliefs about $\boldsymbol{X}$ than there are elements in $\mathcal{X}$. So that if there is one $\boldsymbol{p}$ that satisfies (2.2), then there likely to be an infinity of them.[1]

Another way to write (2.2) is

$$G_{(1)} \cdot p_1 + \cdots G_{(s)} \cdot p_s = \boldsymbol{v}$$

where $G_{(j)}$ is the jth column of G.[2] Because $\boldsymbol{p} \in \mathbb{S}^{s-1}$, another way to state the FTP is that $\boldsymbol{v}$ must lie inside the *convex hull* of the columns of G. This convex hull is a subset of $\mathbb{R}^k$; the case where $k \leftarrow 2$ is shown in Figure 2.1.

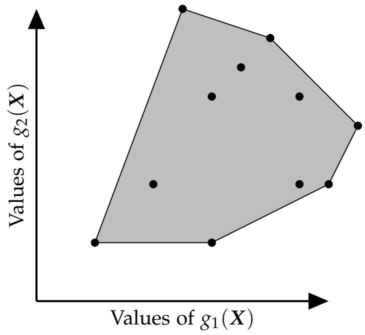

Values of $g_2(\boldsymbol{X})$
Values of $g_1(\boldsymbol{X})$

Figure 2.1: The shaded polygon shows the convex hull of the columns of G when $k \leftarrow 2$. Each point is one column.

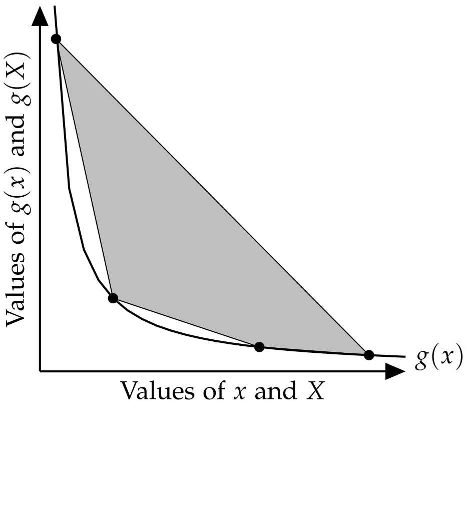
As one immediate consequence of this way of stating the FTP, we have the following very powerful result. Recollect that a function $g : \mathbb{R}^m \to \mathbb{R}$ is convex on the set $A \subset \mathbb{R}^m$ exactly when

$$\alpha g(\boldsymbol{x}) + (1-\alpha) g(\boldsymbol{x}') \geq g\big(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{x}'\big)$$

for all $\boldsymbol{x}, \boldsymbol{x}' \in A$ and all $0 \leq \alpha \leq 1$.

Theorem 2.1 (Jensen's inequality)**.**
If g is a convex function, then $\mathrm{E}\{g(\boldsymbol{X})\} \geq g(\mathrm{E}\{\boldsymbol{X}\})$.

Proof. Here is a proof-by-picture in the case where X is a scalar; the generalisation to a vector $\boldsymbol{X}$ is immediate. The function g is a particular convex function but it represents an arbitrary convex function.



[1] See, e.g., Strang (2006) to review basic properties of matrices and matrix equalities.

[2] This matrix notation is from Mardia *et al.* (1979, ch. 1).

The dots indicate the realm of $\big(X, g(X)\big)$ and the shaded polygon shows the set of values of $\big(\mathrm{E}\{X\}, \mathrm{E}\{g(X)\}\big)$ which are coherent, according to the FTP. Because g is convex, no point in the polygon is below the function $g(x)$. □

2.2 *Extension*

Now suppose I am interested in some new random quantities, $h_1(\boldsymbol{X}), \dots, h_n(\boldsymbol{X})$. I am not been able to specify my expectations of these random quantities directly, but I can use those expectations which I am able to specify, see (2.1), to constrain my expectations of these new random quantities, again using the FTP.

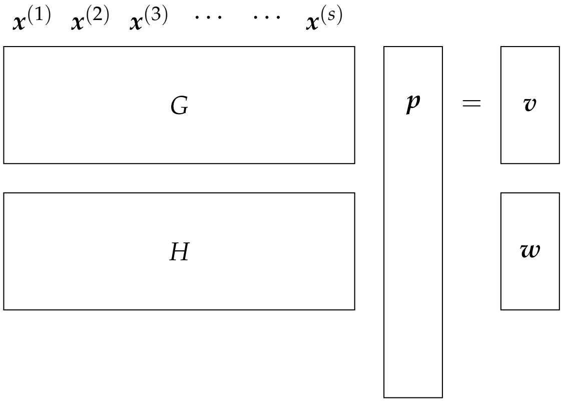
As before, construct the matrix

$$\{H\}_{ij} := h_i(\boldsymbol{x}^{(j)}).$$

According to the FTP, $\boldsymbol{w} := (w_1, \dots, w_n)$ is a coherent expectation for these new random quantities if and only if

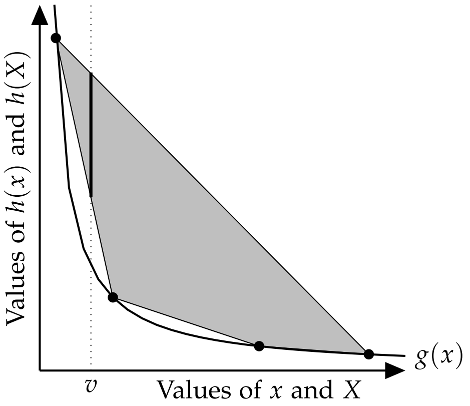
$$\begin{bmatrix} G \\ H \end{bmatrix} \boldsymbol{p} = \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix} \quad \text{for some } \boldsymbol{p} \in \mathbb{S}^{s-1}. \tag{2.3}$$

Or, visually,



If there are many values of $\boldsymbol{p}$ which satisfy (2.2), then there are many values of $\boldsymbol{w}$. Hence a short fat G implies that expectations for other random quantities such as $h_i(\boldsymbol{X})$ will be undecided.

The extension result can be illustrated using the Jensen's inequality picture (from the proof of Thm 2.1).



This is the same picture as before, except now I have marked the value $\mathrm{E}(X) \leftarrow v$. This value is coherent because it lies inside the convex hull of the realm of X. Only the values on the thick black line are coherent values of $\mathrm{E}\{h(X)\}$ corresponding to this value of $\mathrm{E}(X)$.

Here is one very important property of the set of coherent extensions. Informally, it states that the coherent set for any undecided expectation is an interval; i.e. completely defined by lower and upper bounds. Recollect that a set S is convex exactly when $s, s' \in S$ implies that $\alpha s + (1-\alpha)s' \in S$ for all $0 \leq \alpha \leq 1$. The unit simplex $\mathbb{S}^{s-1}$ is convex.

Theorem 2.2. *Let $\mathcal{H}$ be the set of coherent expectations for $h_1(\boldsymbol{X}), \dots, h_n(\boldsymbol{X})$. The set $\mathcal{H}$ is convex.*

Proof. Empty sets are convex, so let $\mathcal{H}$ be non-empty, and let $\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{H}$. We have $\boldsymbol{w} \in \mathcal{H}$ if and only if there exists a $\boldsymbol{p} \in \mathbb{S}^{s-1}$ satisfying (2.3).

Now consider the new point

$$\boldsymbol{w}'' := \alpha \boldsymbol{w} + (1-\alpha)\boldsymbol{w}'$$

for some $\alpha \in [0,1]$. Then

$$\begin{aligned}
\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{w}'' \end{bmatrix} &= \alpha \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix} + (1-\alpha) \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{w}' \end{bmatrix} \\
&= \alpha \begin{bmatrix} G \\ H \end{bmatrix} \boldsymbol{p} + (1-\alpha) \begin{bmatrix} G \\ H \end{bmatrix} \boldsymbol{p}' \quad \text{for some } \boldsymbol{p}, \boldsymbol{p}' \in \mathbb{S}^{(s-1)} \\
&= \begin{bmatrix} G \\ H \end{bmatrix} \big(\alpha \boldsymbol{p} + (1-\alpha)\boldsymbol{p}'\big) \\
&= \begin{bmatrix} G \\ H \end{bmatrix} \boldsymbol{p}'' \qquad \text{for some } \boldsymbol{p}'' \in \mathbb{S}^{s-1}, \text{ because } \mathbb{S}^{s-1} \text{ is convex}
\end{aligned}$$

showing that $\boldsymbol{w}'' \in \mathcal{H}$. □

Thus for any specified set of expectations, I can bound any other expectation $\mathrm{E}\{h(\boldsymbol{X})\}$ with lower and upper values. These bounds describe the amount of 'undecided' in my inference about $h(\boldsymbol{X})$.

Now consider the case of finding the lower bound on $\mathrm{E}\{h(\boldsymbol{X})\}$ for some specified function h, based on beliefs $[G, \boldsymbol{v}]$. We must solve

$$\min_{\boldsymbol{p} \in \mathbb{R}^s} \boldsymbol{h}^T \boldsymbol{p} \quad \text{subject to} \quad \begin{cases} G\boldsymbol{p} = \boldsymbol{v} \\ \sum_j p_j = 1 \\ p_j \geq 0 \quad j = 1, \dots, s \end{cases}$$

where $\boldsymbol{h} := (h^{(1)}, \dots, h^{(s)})$ and $h^{(j)} := h(\boldsymbol{x}^{(j)})$. This is a *linear programming (LP)* problem. LP represents one of the pinnacles of computer-based optimisation, discussed in Nocedal and Wright (2006, chs 13 and 14). As Nocedal and Wright explain, linear programming has been revolutionised by the recent development of interior point methods, which are replacing the simplex methods that were originally developed in the 1940s.

* * *

Unfortunately, however, even modern linear programming methods will grind to a halt if s, the size of the joint realm of $\boldsymbol{X}$, is too large. And because s is exponential in the number of random quantities, it only takes a few tens of random quantities before this happens. This is a tragedy for statistical inference as I have presented it here, because our inability to do the computations forces us down another route which provides a very different framework for specifying beliefs, one in which almost all of our limitations as uncertainty assessors is suppressed. This alternative framework is discussed in detail in Chapter 4.

But I believe it is valuable to explore how we *ought* to do statistical inference, and then to encounter the practical difficulties, in order to understand better why in practice we do statistical inference the way we do. I hazard that most people who work with uncertainty are not aware that there is a rich calculus of expectation that allows me to specify just as many beliefs as I feel able, and represents the results in terms of 'undecided' intervals for those expectations that I am unable to specify. It is true that in many applications these unaware people are not disadvantaged, because the implementation of such a calculus is computationally impractical. But even then it is important to know that there is a substantial gulf between what one ought to do, and what one ends up doing.

Therefore I will continue to explore what we ought to do: which is what we can do in some situations, and what we can expect to do as computing power continues to grow.

2.3 *Representing beliefs*

Suppose I am satisfied that my beliefs $[G, \boldsymbol{v}]$ are coherent, and I am now considering their extension to some new random quantity $h(\boldsymbol{X})$. The best possible outcome is to find that my set of coherent values for $\mathrm{E}\{h(\boldsymbol{X})\}$ is constrained to a single point; in other words, my expectation of $h(\boldsymbol{X})$ is completely constrained by my expectations for $g_1(\boldsymbol{X}), \dots, g_k(\boldsymbol{X})$. This can arise in the obvious way: for example, where $g_1(\boldsymbol{x}) \leftarrow x_1$, $g_2(\boldsymbol{x}) \leftarrow x_2$, and $h(\boldsymbol{x}) \leftarrow x_1 + x_2$. But it can also arise in much less obvious ways, involving the interplay of the more subtle constraints that are represented by the theorems of the expectations calculus. Because these theorems follow directly from the axioms, they are automatically enacted in the FTP. Thus $\mathcal{H}$ in Thm 2.2 must respect Schwartz's inequality, Jensen's inequality, Markov's inequality, and so on. Expectations for a rich set of g_i's will have many more implications for the nature of $\mathcal{H}$ than I can easily envisage, and computation is the only method I have to infer them all. Computation was briefly discussed at the end of Sec. 2.2.

In general, however, we must accept that many of my expectations will not be constrained to a point, i.e. I will remain undecided about my $\mathrm{E}\{h(\boldsymbol{X})\}$. Thm 2.2 states that my set of coherent expectations for $\mathrm{E}\{h(\boldsymbol{X})\}$ can be represented by an interval, and defined in terms of lower and upper bounds. It is important to present this clearly. For example, to state "My expectation for $X_1 + 2\log X_2$ is *undecided* but lies in the interval $[3.2, 5.5]$." This is because there are advocates of a more general calculus of expection, who propose that my beliefs about the $g_i(\boldsymbol{X})$'s may themselves be expressed in terms of intervals (see, e.g., Walley, 1991; Troffaes and de Cooman, 2014). So I would like the word 'undecided' to indicate a technical meaning associated with a purely mechanical derivation from a coherent set of specified expectations.

A wide range of beliefs can be encoded as expectations, and we should look beyond obvious beliefs such as $\mathrm{E}(X_1) \leftarrow v_1$. As discussed in Sec. 1.4, probabilities are also expectations, so each probability I specify constitutes a row of $[G, \boldsymbol{v}]$. For example, suppose that $q(\boldsymbol{x})$ is a first-order sentence, so that $Q := q(\boldsymbol{X})$ is a random proposition. If I think that Q has probability p_q then this is represented by a row of $[G, \boldsymbol{v}]$ with

$$G_{ij} \leftarrow \mathbb{1}_{q(\boldsymbol{x}^{(j)})} \quad \text{and} \quad v_i \leftarrow p_q.$$

Certainty is a special case: a random proposition to which I assign probability 1. If I am certain that Q is true, i.e. $p_q \leftarrow 1$, then this has the effect of zeroing those p_j for which $q(\boldsymbol{x}^{(j)})$ is false:

$$\text{row } i\text{: } \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ \vdots \\ p_s \end{pmatrix} = 1$$

and since $p_j \geq 0$ and $\sum_j p_j = 1$, $q(\boldsymbol{x}^{(j)})$ false implies that $p_j = 0$. So the same effect could be achieved by removing from $\mathfrak{X}$ all of the elements for which $q(\boldsymbol{x}^{(j)})$ is false.

Sec. 2.4 will discuss data (there is nothing special about data!). But it is worth mentioning measurement error separately. Measurement error typically involves two random quantities: the underlying quantity of interest, and the measurement of it, which may not be the same. Let X be the underlying quantity of interest, and Y be the measurement. Then the measurement $Y \to y$ will be included as a row in G stating that $\Pr(Y \doteq y) = 1$. Other rows in G record my beliefs about the accuracy of the measurement. For example, if I believe that the measurement error $Y - X$ has an expectation of zero and a variance of σ^2, then I include rows for $\mathrm{E}\{(Y - X)\} \leftarrow 0$ and $\mathrm{E}\{(X - Y)^2\} \leftarrow \sigma^2$.

I can extend this approach to include my uncertainty about the size of the measurement error. Let V be my variance for measurement error, which I could operationalise in terms of the performance of the instrument over a large number of uses. So this is another random quantity. I include in my beliefs with a row for my expectation, $\mathrm{E}\{V\} \leftarrow \sigma^2$, where I specify σ^2, and a row for its interpretation, $\mathrm{E}\{(X - Y)^2 - V\} \leftarrow 0$. If the same instrument were used in many measurements then V would appear in many rows of G. In this way, these measurements enable me to make inferences about V, as well as about the underlying quantities of interest.

Suppose I wanted to go further, and make the measurement error Normally distributed. I could do this by specifying a large number of moments of $Y - X$, where the kth moment is defined as $\mathrm{E}\{(Y - X)^k\}$ for $k = 1, 2, \ldots$.[3] But a small number of moments would be a better reflection of my limited beliefs. So I might also include rows for the third and fourth moments, namely $\mathrm{E}\{(Y - X)^3\} \leftarrow 0$ and $\mathrm{E}\{(Y - X)^4 - 3V^2\} \leftarrow 0$. (These are the particular values for the Normal distribution, computed from the Normal moment generating function.) If my beliefs in the 'Normality' of the measurement error were stronger I could add more moments, the fifth, the sixth, and so on.

This gives a hint of something that will be discussed further in Chapter 4: a belief about the distribution of a random quantity is a very strong belief, involving a potentially huge number of rows of G.

2.4 *Data*

Statisticians seem to fetishize data; undoubtedly this is because data seem far more secure than beliefs.[4] This security is only partly reassuring. Unless we are lucky enough to measure exactly what we need to infer, we will always need beliefs to link our data with our inference. In the simplest case, these would be beliefs about measurement error, but it is usually more complicated than this. Usually, there is a large gap between the data we actually have, and the inference we need to do. Almost always, we can identify data that we would like to have, but do not: because it is too expensive, or unethical, or unobtainable (data from the past or the future, for example). Sometimes it is helpful to reflect on what data we would really like; othertimes it is just frustrating! But we always have to

[3] It is a theorem that a distribution can be specified by its moments; see Grimmett and Stirzaker (2001, ch. 5).

[4] I have a mild preference for 'data' as a plural noun.

return to reality, to the data we actually have, and the beliefs we need to link them to our inference.

There is—in principle—nothing special about the treatment of data in statistical inference. A datum $Y \to y$ is simply the probability assignment $\Pr(Y \doteq y) \leftarrow 1$. However, it is helpful to have a slightly expanded notion of data, to reflect the kinds of data typically encountered. Here is a general representation for data. We start, as usual, with a set of random quantities, $\boldsymbol{X} := (X_1, \ldots, X_m)$. Every datum is a value for a specified function of $\boldsymbol{X}$, so I write

$$Y_i = g_i(\boldsymbol{X}) \qquad i = 1, \ldots, n$$

to represent the *observables*, where the g_i are specified. These observables are the random quantities which are measured. The actual measurements are the *observations*

$$\boldsymbol{y}^{\text{obs}} := (y_1^{\text{obs}}, \ldots, y_n^{\text{obs}}).$$

The n observations are represented in my beliefs as the truth of the random proposition

$$q(\boldsymbol{X}) := \bigwedge_{i=1}^{n} q_i(\boldsymbol{X}) \qquad \text{where } q_i(\boldsymbol{x}) := \big(g_i(\boldsymbol{x}) \doteq y_i^{\text{obs}}\big).$$

In other word, q is the conjunction of n random propositions, where each random proposition represents one datum. As described in Sec. 2.3, including the row $\Pr\{q(\boldsymbol{X})\} \leftarrow 1$ in my beliefs is equivalent to removing all of the elements of the realm of $\mathcal{X}$ for which $\mathbb{1}_{q(\boldsymbol{x}^{(j)})} = 0$. That is, all the elements of $\mathcal{X}$ that are not consistent with the observations.

Here are some common situations; I will ignore measurement error, for simplicity (see Sec. 2.3). The simplest situation is where each observable corresponds to a specific component of $\boldsymbol{X}$. We can reorder $\boldsymbol{X}$ so that the observables are the first n components, i.e. $g_i(\boldsymbol{x}) \leftarrow x_i$. I term this the *simple observational model (SOM)*. It often applies when the observables are a sample from a population.

Another very common situation is where the observables are linear combinations of $\boldsymbol{X}$, often averages or sums. For example, a satellite instrument might have a spatial footprint of hundreds of kilometres. If each X_i represents a ten-kilometre pixel it is common to represent the observable as measuring the average value of all of the pixels in the footprint (see Zammit-Mangion *et al.*, 2014, for lots of modelling of this type).

Observables which are nonlinear functions are also possible. Sometimes measurements are truncated: the instrument is only capable of measuring a quantity up to a specified threshold value. In this case $Y_i = g_i(\boldsymbol{X}) = \min\{X_i, u\}$, where u is the upper threshold. Truncation also happens in studies which run for a fixed duration. In a medical trial, for example, some people will survive until the end of the trial: their age at death is not known (unlike those who died), although it is bounded below by their age at the end of the trial. Such measurements are known as 'right censored'.

In other cases the instrument may have a bias which depends on the value of the quantity being measured. Consider, for example, the difficulty of measuring wind speed with a mechanical anenometer (there are about ten different types), or precipitation with a gauge that has to work for snow, rain, and drizzle.

In all of these situations, the function g_i is specified. In a more complicated analysis g_i can itself be made uncertain, by including additional random quantities which represent arguments to g_i; e.g. $Y_i = g_i(\boldsymbol{X}, U) = \min\{X_i, U\}$ for an instrument with an uncertain upper threshold U. As with an uncertain measurement error variance, the value of U can be accurately inferred if the same instrument is used to make many measurements.