Tutorial for Metrologists
on the probabilistic and statistical apparatus
underlying the GUM and related documents

— Antonio Possolo & Blaza Toman —

Statistical Engineering Division
National Institute of Standards and Technology
Gaithersburg, MD, USA

November 21, 2011

Contents

1 Preamble

Familiarity with the basic concepts and techniques from probability theory and mathematical statistics can best be gained by studying suitable textbooks and exercising those concepts and techniques by solving instructive problems. The books by DeGroot and Schervish [2011], Hoel et al. [1971a], Hoel et al. [1971b], Feller [1968], Lindley [1965a], and Lindley [1965b] are highly recommended for this purpose and appropriate for readers who will have studied mathematical calculus in university courses for science and engineering majors.

This document aims to provide an overview of some of these concepts and techniques that have proven useful in applications to the characterization, propagation, and interpretation of measurement uncertainty as described, for example, by Morgan and Henrion [1992] and Taylor and Kuyatt [1994], and in guidance documents produced by international organizations, including the *Guide to the expression of uncertainty in measurement* (GUM) [Joint Committee for Guides in Metrology, 2008a] and its supplements [Joint Committee for Guides in Metrology, 2008b]. However, the examples do not all necessarily show a direct connection to measurement science.

Our basic premises are that probability is best suited to express uncertainty quantitatively, and that Bayesian statistical methods afford the best means to exploit information about quantities of interest that originates from multiple sources, including empirical data gathered for the purpose, and preexisting expert knowledge.

Although there is nothing particularly controversial about the calculus of probability or about the mathematical methods of statistics, both the meaning of probability and the interpretation of the products of statistical inference continue to be subjects of debate.

This debate is meta-probabilistic and meta-statistical, in the same sense as metaphysics employs methods different from the methods of physics to study the world. In fact, the debate is liveliest and most scholarly among professional philosophers [Fitelson, 2007]. However, probabilists and statisticians often participate in it when they take off their professional hats and become philosophers [Neyman, 1977], as any inquisitive person is wont to do, at one time or another.

For this reason, we begin with an overview of some of the meanings that have been assigned to probability (§2) before turning to the calculus of probability (§3). In applications, the devices of this calculus are typically brought into play when considering random variables and probability distributions (§4), in particular to characterize the probability distribution of functions of random variables (§5). Statistical inference (§6) uses all of these devices to produce probabilistic statements about unknown quantity values.

2 Probability

2.1 Meaning

In *Truth and Probability* [Ramsey, 1926, 1931], Frank Ramsey takes the view that probability is “a branch of logic, the logic of partial belief and inconclusive argument”. In this vein, and more generally, probabilities serve to quantify uncertainty. For example, when one states that, with 99 % confidence, the distance between two geodesic marks is within 0.07 m of 936.84 m, one believes that the actual distance most likely lies between 936.77 m and 936.91 m, but still entertains the possibility that it may lie elsewhere. Similarly, a weather service announcement of 20 % chance of rain tomorrow for a particular region summarizes an assessment of uncertainty about what will come to pass.

Although relevant to the interpretation of measurement uncertainty, and generally to all applications of probability and statistics, the meaning of probability really is a philosophical issue [Gillies, 2000, Hájek, 2007, Mellor, 2005]. And while there is much disagreement about what probabilities mean, and how they are created to begin with (*interpretation* and *elicitation* of probability), there also is essentially universal agreement about how numerical assessments of probability should be manipulated and combined (*calculus* of probability).

2.2 Chance and Propensity

Chances arise in connection with games of chance, and with phenomena that conceivably can recur under essentially the same circumstances. Thus one speaks of the chances of a pair of Kings in a poker hand, or of the chances that the nucleus of an atom of a particular uranium isotope will emit an alpha particle within a given time interval, or of the chances that a person born in France will have blood of type AB. Chances seem to be intrinsic properties of objects or processes in specific environments, maybe propensities for something to happen: their most renowned theorists have been Hans Reichenbach [Reichenbach, 1949], Richard von Mises [von Mises, 1981], and Karl Popper [Popper, 1959].

2.3 Credence and Belief

Credences measure subjective beliefs. They are best illustrated in relation with betting on the outcomes of events one is uncertain about. For example, in this most memorable of bets offered when the thoroughbred *Eclipse* was about to run against *Gower*, *Tryal*, *Plume*, and *Chance* in the second heat of the races on May 3rd, 1769, at Epsom Downs: "*Eclipse* first, the rest nowhere", with odds of 6-to-4 [Clee, 2007].

The strength or degree of these beliefs can be assessed numerically by techniques that include the observation of betting behavior (actual or virtual), and this assessment can be gauged, and improved, by application of scoring rules [Lindley, 1985].

Dennis Lindley [Lindley, 1985] suggests that degrees of belief can be measured by comparison with a standard, similarly to how length or mass are measured. In general, subjective probabilities can be revealed by judicious application of elicitation methods [Garthwaite et al., 2005].

Consider an urn that contains 100 balls that are identical but for their colors: β are black and $100 - \beta$ are white. The urn's contents are thoroughly mixed, and the standard is the probability of the event B of drawing a black ball. Now, given an event E, for example that, somewhere in London, it will rain tomorrow, whose probability he wishes to gauge, Peter will select a value for β such that he regards gambling on B as equivalent to gambling on E (for the same prize): in these circumstances, $\beta/100$ is Peter's credence on E.

The beliefs that credences measure are subjective and personal, hence the probabilities that gauge them purport to a relationship between a particular knowing subject and the object of this subject's interest. These beliefs certainly are informed by such knowledge as one may have about a situation, but they also are tempered by one's preferences or tastes, and do not require that a link be drawn explicitly between that knowledge or sentiment and the corresponding bet. Bruno de Finetti [de Finetti, 1937, 1990], Jimmie Savage [Savage, 1972], and Dennis Lindley [Lindley, 2006] have been leading developers of the subjective, personalistic viewpoint.

2.4 Epistemic Probability

Logical (or *epistemic*, that is, involving or relating to knowledge) probabilities measure the degree to which the truth of a proposition justifies, warrants, or rationally supports the truth of another [Carnap, 1962]. For example, when a medical doctor concludes that a positive result in a tuberculin sensitivity test indicates tuberculosis with 62.5 % probability (§3.6), or when measurements made during a total eclipse of the sun overwhelmingly favor Einstein's theory of gravitation over Newton's [Dyson et al., 1920].

The fact that scientists or judges may not necessarily or explicitly use probabilities to convey their confidence in theories or in arguments [Glymour, 1980] does not reduce the value that probabilities have in models for the rational process of learning from experience, either for human subjects or for reasoning machines that are programmed to make decisions in situations of uncertainty. In this fashion, probability is an extension of deductive logic, and measures degree of confirmation: it does this objectively because it does not involve subjective personal opinion, hence is as incontrovertible as deduction by any of the forms of classical logic.

The difficulty lies in specifying a starting point, a state of *a priori* ignorance that is similarly objective and hence universally acceptable. Harold Jeffreys [Jeffreys, 1961] provided maybe the first modern, thorough account of how this may be done. He argued, and illustrated in many substantive examples, that it is fit to address the widest range of scientific problems where one wishes to exploit the information in observational data.

The interpretation of probability as an extension of logic makes it particularly well-suited to applications in measurement science, where it is desirable to be able to treat different uncertainty components, which may have been evaluated using different methods, simultaneously, using a uniform vocabulary, and a single set of technical tools. This concept underlies the treatment of measurement uncertainty in the GUM and in its supplements. Richard Cox [Cox, 1946, 1961] and Edwin Jaynes [Jaynes, 1958, 2003] have articulated cogent arguments in support of this view, and José Bernardo [Bernardo, 1979] and James Berger [Berger, 2006] have greatly expanded it.

2.5 Difficulties

Even in situations where, on first inspection, chances seem applicable, closer inspection reveals that something else really is needed.

There may be no obvious reason to doubt that the chance is ½ that a coin tossed to assign sides before a football game will land *Heads* up. However, if the coin instead is spun on its edge on a table, that chance will be closer to either 1/3 or 2/3 than to ½ [Diaconis and Ylvisaker, 1985].

And when it is the magician *Persi Warren* [DeGroot, 1986] who tosses the coin, then all bets are off because he can manage to toss it so that it always lands

Heads up on his hand after he flips it in the air: while the unwary may be willing to bet at even odds on the outcome, for Persi the probability is 1.

And there are situations that naturally lend themselves to, or that even seem to require, multiple interpretations. Take the 20 % chance of rain: does this mean that, of all days when the weather conditions have been similar to today's in the region the forecast purports to, it has rained some time during the following day with historical frequency of 20 %? Or is this the result of a probabilistic forecast that is to be interpreted epistemically? Maybe it means something else altogether, like: of all things that will fall from the sky tomorrow, 1 in 5 will be a raindrop.

2.6 Role of Context

The use of probability as an extension of logic ensures that different people who have the same information (empirical or other) about a measurand should produce the same inferences about it. Example §2.7 illustrates the fact that contextual information relating to a proposition, situation, or event, will influence probabilistic assessments to the extent that different people with different information may, while all acting rationally, produce different uncertainty assessments, and hence different probabilities, for the same proposition, situation, or event.

When probabilities express subjective beliefs, or when they express states of incomplete or imperfect knowledge, different people typically will assign different probabilities to the same statements or events. If they have to reach a consensus on a course of action that is informed by their plural, varied assessments, then they have to engage in a harmonization exercise that preserves the internal coherence of their individual positions. Both statisticians [Stone, 1961, Morris, 1977, Lindley, 1983, Clemen and Winkler, 1999] and philosophers [Bovens and Rabinowicz, 2006, Hartmann and Sprenger, 2011] have addressed this topic.

2.7 EXAMPLE: Prospecting

James and Claire, who both make investments in mining prospects, have been told that samples from a region surveyed recently have mass fractions of titanium averaging $3\,\mathrm{g\,kg^{-1}}$, give or take $1\,\mathrm{g\,kg^{-1}}$ (where "give or take" means that the true mass fraction of titanium in the region sampled is between $2\,\mathrm{g\,kg^{-1}}$ and $4\,\mathrm{g\,kg^{-1}}$ with 95 % probability). James, however, has also been told that the samples are of a sandstone with grains of ilmenite. On this basis, James may assign a much higher probability than Claire to the proposition that asserts that the region sampled includes an economically viable deposit of titanium ore.

3 Probability Calculus

3.1 Axioms

Once numeric probabilities are in hand, irrespective of how they may be interpreted, the same set of rules, or axioms, is used to combine them. We formulate these axioms in the context where probability is regarded as measuring degree of (rational) belief in the truth of propositions, given a particular body of knowledge and universe of discourse, H, that makes all the participating elements meaningful.

Let A and B denote propositions whose probabilities $\Pr(A|H)$ and $\Pr(B|H)$ express degrees of belief about their truth given (or, *conditionally* upon) the context defined by H. The notation $\Pr(B|A \text{ and } H)$ denotes the *conditional* probability of B, assuming that A is true and given the context defined by H. Note that $\Pr(B|A \text{ and } H)$ is not necessarily 0 when $\Pr(A|H) = 0$: for example, the probability is 0 that a point chosen uniformly at random over the surface of the earth will be on the equator; yet the probability is ½ that, conditionally on its being on the equator, its longitude is between 0° and 180° West of the prime meridian at Greenwich, UK.

The axioms for the calculus of probability are these:

Convexity: $\Pr(A|H)$ is a number between 0 and 1, and it is 1 if and only if H logically implies A;

Addition: $\Pr(A \text{ or } B|H) = \Pr(A|H) + \Pr(B|H) - \Pr(A \text{ and } B|H)$, where the expression "*A or B*" is true if and only if A, B, or both are true;

Multiplication: $\Pr(A \text{ and } B|H) = \Pr(B|A \text{ and } H)\Pr(A|H)$.

Since the roles of A and B are interchangeable, the multiplication axiom obviously can also be written as $\Pr(A \text{ and } B|H) = \Pr(A|B \text{ and } H)\Pr(B|H)$.

Most accounts of mathematical probability theory use an additional rule (*countable additivity*) that ensures that the probability that at least one proposition is true, among countably infinitely many mutually exclusive propositions, equals the sum of their individual probabilities [Casella and Berger, 2002, Definition

1.2.4]. ("Countably infinitely many" means "as many as there are integer numbers".)

When the context that H defines is obvious, often one suppresses explicit reference to it, as in this derivation: if $\widetilde{A}$ denotes the negation of A, then Convexity and the Addition Rule imply that $1 = \Pr(A \text{ or } \widetilde{A}) = \Pr(A) + \Pr(\widetilde{A})$ because one but not both of A and $\widetilde{A}$ must be true.

3.2 Independence

The concept of *independence* pervades much of probability theory. Two propositions A and B are independent if the probability that both are true equals the product of their individual probabilities of being true. If A asserts that there is a Queen in Alexandra's poker hand, and B asserts that Beatrice's comprises red cards only, both hands having been dealt from the same deck, then A and B are independent. Intuitively, if knowledge of the truth of one proposition influences the assessment of probability of another, then they are dependent: in particular, two mutually exclusive propositions are *dependent*. If $\Pr(B \,|\, A \text{ and } H) = \Pr(B \,|\, H)$, then A and B are independent given H.

3.3 Extending the Conversation

When considering the probability of a proposition, it often proves advantageous to consider the truth or falsity of another one, somehow related to the first [Lindley, 2006, §5.6]. To assess the probability $\Pr(+)$ of a positive tuberculin skin test (§3.6), it is convenient to consider how the test performs separately in persons infected or not infected with *Mycobacterium tuberculosis*: if I denotes infection, then $\Pr(+) = \Pr(+ \,|\, I)\Pr(I) + \Pr(+ \,|\, \widetilde{I})\Pr(\widetilde{I})$, where $\Pr(+ \,|\, \widetilde{I})$ is the probability of a *false positive*, and $1 - \Pr(+ \,|\, I)$ is the probability of a *false negative*, both more accessible than $\Pr(+)$.

3.4 Coherence

If an ideal reasoning agent (human or machine) assigns probabilities to events or to the truth of propositions according to the foregoing axioms, then this agent's beliefs are said to be *coherent*. In these circumstances, if probabilities are used to inform bets concerning the truth of propositions in the universe of discourse where these probabilities are meaningful, then it is impossible (for a "bookie") to devise a collection of bets that bring an assured loss to this agent (a so-called "Dutch Book").

Now, suppose that, having ascertained the truth of a proposition A, one produces $\Pr(C\,|\,A)$ as assessment of C's truth on the evidence provided by A. Next, one determines that B, too, is true and revises this last assessment of C's truth to become $\Pr(C\,|\,B \text{ and } A)$. The process whereby probabilities are updated is *coherently extensible* if the resulting assessment is the same irrespective of whether the evidence provided by A and B is brought to bear either sequentially, as just considered, or simultaneously. The incorporation of information from multiple sources, and the corresponding propagation of uncertainty, that is carried out by application of Bayes' formula, which is described next and illustrated in examples §3.6 and §6.7, is coherently extensible.

3.5 Bayes's Formula

If exactly one (that is, one and one only) among propositions $A_1, \dots, A_n$ can be true, and B is another proposition with positive probability, then

$$\Pr(A_j|B) = \frac{\Pr(B|A_j)\Pr(A_j)}{\sum_{i=1}^{n} \Pr(B|A_i)\Pr(A_i)}, \text{ for } j = 1, \dots, n. \tag{1}$$

This follows from the axioms above because $\Pr(A_j|B) = \Pr(A_j \text{ and } B)/\Pr(B)$ (Multiplication axiom), whose numerator equals $\Pr(B|A_j)\Pr(A_j)$ (Multiplication axiom), and whose denominator equals $\Pr(B|A_1)\Pr(A_1) \dots \Pr(B|A_n)\Pr(A_n)$ ("extending the conversation", as in §3.3).

3.6 EXAMPLE: Tuberculin Test

Richard has been advised that his tuberculin skin test has returned a positive result. The tuberculin skin test has a reported false-negative rate of 25 % during the initial evaluation of persons with active tuberculosis [American Thoracic Society, 1999, Holden et al., 1971]: this means that the probability is 0.25 that the test will yield a negative $(-)$ response when administered to an infected person (I), $\Pr(-\,|\,I) = 0.25$. Therefore, the probability is only 0.75 that infection will yield a positive test result. In populations where cross-reactivity with other mycobacteria is common, the test's false-positive rate is 5%: that is, the conditional probability of a positive result $(+)$ for a person that is not infected $(\widetilde{I})$ is $\Pr(+\,|\,\widetilde{I}) = 0.05$.

Richard happens to live in an area where tuberculosis has a prevalence of 10 %. Given the positive result of the test he underwent, the probability that he is infected is

$$\begin{aligned}\Pr(I\,|\,+) &= \frac{\Pr(+\,|\,I)\Pr(I)}{\Pr(+\,|\,I)\Pr(I)+\Pr(+\,|\,\widetilde{I})\Pr(\widetilde{I})}\\ &= \frac{(0.75\times 0.10)}{(0.75\times 0.10)+(0.05\times 0.90)} = 0.625.\end{aligned}$$

Common sense suggests that the diagnostic value of the test should depend on its false-negative and false-positive rates, as well as on the prevalence of the disease: Bayes' formula states exactly how these ingredients should be combined to produce $\Pr(I\,|\,+)$, which expresses that diagnostic value quantitatively.

Richard has the tuberculin skin test repeated, and this second test also turns out positive. To incorporate this additional piece of evidence into the probability that Richard is infected, first we summarize the state of knowledge (about whether he is infected) determined by the result from the first test. This is done by defining $Q(I) = \Pr(I\,|\,+) = 0.625$ and $Q(\widetilde{I}) = 1 - Q(I) = 0.375$, and using them in the role that the overall probability of infection (10 %) or non-infection (90 %) played prior to Richard's first test, when all one knew about his condition was that he was a member of a population where the prevalence of tuberculosis was 10 %.

Again applying Bayes' theorem, and assuming that the two tests are independent, the revised probability that Richard is infected after two positive tests is

$$\begin{aligned}Q(I\,|\,+) &= \frac{\Pr(+\,|\,I)Q(I)}{\Pr(+\,|\,I)Q(I)+\Pr(+\,|\,\widetilde{I})Q(\widetilde{I})}\\ &= \frac{(0.75\times 0.625)}{(0.75\times 0.625)+(0.05\times 0.375)} = 0.962.\end{aligned}$$

If, instead, one had been initially told that Richard had had two independent, positive tuberculin skin tests, then the calculation would have been:

$$\begin{aligned}\Pr(I\,|\,++) &= \frac{\Pr(++\,|\,I)\Pr(I)}{\Pr(++\,|\,I)\Pr(I)+\Pr(++\,|\,\widetilde{I})\Pr(\widetilde{I})}\\ &= \frac{(0.75^2\times 0.10)}{(0.75^2\times 0.10)+(0.05^2\times 0.90)} = 0.962.\end{aligned}$$

This example illustrates the fact that Bayes' theorem produces the same probability irrespective of whether the information is incorporated sequentially, or all at once.

3.7 Growth of Knowledge

The example in §3.6 illustrated how the probability of Richard being infected increased (relative to the overall probability of infection in the town where he lives) as a first, and then a second tuberculin test turned out positive. However, even if he is infected, by chance alone a test may turn out negative. In a sequence of tests, therefore, the probability of his being infected may oscillate, increasing when a test turns out positive, decreasing when some subsequent test turns out negative.

Therefore, the question naturally arises of whether a person employing the Bayesian method of exploiting information, and incorporating it into the current state of knowledge, ever will, in situations of uncertainty, arrive at conclusions with overwhelming confidence. Jimmie Savage proved rigorously that the answer is "yes" with great generality: "with observation of an abundance of relevant data, the person is almost certain to become highly convinced of the truth, and [...] he himself knows this to be the case" [Savage, 1972, §3.6].

The restriction to "relevant data" is critical: in relation with the tuberculin test, if it happened that $\Pr(+\,|\,I) = \Pr(+\,|\,\widetilde{I}) = 1/2$, then the test would have no discriminatory power, and in fact would be irrelevant to learning about disease status.

4 Random Variables and Probability Distributions

4.1 Random Variables

The notion of *random variable* originates in games of chance, like roulette, whose outcomes are unpredictable. Its rigorous mathematical definition (measurable function from one probability space into another) is unlikely to be of great interest to the metrologist. Instead, one may like to keep in mind its heuristic meaning: the value of a quantity that has a probability distribution as an attribute whose role is to describe the uncertainty associated with that value.

4.2 EXAMPLE: Roulette

In the version of roulette played in Monte Carlo, the possible outcomes are numbers in the set $\mathscr{X} = \{0, 1, \dots, 36\}$ (usually one disregards other possible, but "uninteresting" outcomes, including those where the ball exits the wheel and lands elsewhere, or where it lands inside the wheel but in none of its numbered pockets). Once those 37 numbers are deemed to be equally likely, one can speak of a random variable that is equal to 0 with probability 1/37, or that is odd with probability 18/37. (Note that these statements are meaningful irrespective of whether the event in question will happen in the future, or has happened already, provided one does not know its actual outcome yet).

4.3 EXAMPLE: Light Bulb

The GE A19 Party Light 25 W incandescent light bulb has expected lifetime 2000 h: this is usually taken to mean that, if a brand new bulb is turned on and left on supplied with constant 120 V electrical current until it burns out, its actual lifetime may be described as a realized value (*realization*, or *outcome*) of a random variable with an exponential probability distribution (§4.10) whose expected value is 2000 h — this is denoted η in §4.10, and in general it needs to be estimated from experimental data.

The concept of random variable applies just as well to domains of discourse unrelated to games of chance, hence can be used to suggest uncertainty about the value of a quantity, irrespective of the source of this uncertainty, including situations where there is nothing "random" (in the sense of "chancy") in play.

4.4 Notational Convention

For the most part, upper case letters (Roman or Greek) denote generic quantity values modeled as random variables, and their lowercase counterparts denote particular values.

Upper case letters like X or $X_1, X_2, \dots$, and Y, denote generic random variables, without implying that any of the former necessarily play the role of *input quantity values* (as defined in the international vocabulary of metrology (VIM) Joint Committee for Guides in Metrology [2008c], VIM 2.50), or that the latter necessarily plays the role of *output quantity values* (VIM 2.51) in a measurement model (VIM 2.48).

The probabilities most commonly encountered in metrological practice concern sets of numbers that a quantity value may take: in this case, if X denotes a random variable whose values belong to a set $\mathcal{X}$, and A is a subset of $\mathcal{X}$, then $\Pr(X \in A)$ denotes the probability that X's value lies in A. For example, if X represents the length (expressed in meter, say) of a gauge block, then $\mathcal{X}$ would be the set of all possible values of length, and A could be the subset of such values between 0.0423 m and 0.0427 m, say.

4.5 Probability Distributions

Given a random variable X one can then define a function P_X such that $P_X(A) = \Pr(X \in A)$ for all $A \subset \mathscr{X}$ to which a probability can be assigned. This P_X is called X's *probability distribution*.

If $\mathscr{X}$ is countable (that is, either finite or infinite but with as many elements as there are positive integers), then one says that X has a *discrete* distribution, which is fully specified by the probability it assigns to each value in $\mathscr{X}$. For example, the outcome of a roulette wheel is a random variable whose probability distribution is discrete.

If $\mathscr{X}$ is uncountable (that is, it has as many elements as there are real numbers) and $\Pr(X = x) = 0$ for all $x \in \mathscr{X}$, then one says that X has a *continuous* distribution. For example, the lifetime of an incandescent light bulb that does light up and then is constantly left on until it burns out is a random variable with a continuous distribution.

A distribution may be neither discrete nor continuous, but of a mixed type instead: for example, when a random variable is equal to 0 with probability $\epsilon > 0$, and has an exponential distribution (see §4.10) with probability $1 - \epsilon$. Since a brand new light bulb has a positive probability of burning out the instant it is turned on, its lifetime may more realistically be modeled as a random variable that has an "atom" of probability at 0, and is exponential with the complementary probability.

4.6 Probability Distribution Function

The probability distribution of a random variable X whose possible values are real numbers, can be succinctly described by its *probability distribution function*, which is the function P_X such that $P_X(x) = \Pr(X \leqslant x)$ for every real number x.

Note that the symbol we use here to denote the probability distribution function, is the same that we used in §4.5 to denote the probability distribution itself. Any confusion this may cause will be promptly resolved by examining the argument of P_X: if it is a set, then we mean the distribution itself, while if it is a number or a vector with numerical components, then we mean the probability distribution function.

For example, if X is real-valued and x is a particular real number, then in $P_X(x) = P_X((-\infty, x])$ the P_X on the left hand side refers to the probability distribution function, while the P_X on the right hand side refers to the distribution itself because $(-\infty, x]$ denotes the set of all real numbers no greater than x. Since the distribution function determines the distribution, the confusion is harmless.

4.7 Probability Density Function

If X has a discrete distribution (§4.5), then its probability density (also known as its *probability mass function*) is the function p_X such that $p_X(x) = \Pr(X = x)$ for $x \in \mathscr{X}$.

If $\mathscr{X}$ is uncountable and X's distribution is continuous and sufficiently smooth (in the sense described next), then the corresponding *probability density function* (PDF) is defined similarly to a material object's mass density, as follows.

Consider the simplest case, where $\mathscr{X}$ is an interval of real numbers, and suppose that x is one point in the interior of this interval. Now suppose that $\delta_1 > \delta_2 > \dots$ is an infinite sequence of positive numbers decreasing to zero. If X's probability distribution is sufficiently smooth, then the limit $p_X(x) = \lim_{n\to\infty} \big(P_X(x + \delta_n) - P_X(x - \delta_n)\big)/(2\delta_n)$ exists. The function p_X so defined is X's probability density function. If the distribution function is differentiable, then the probability density is the derivative of the probability distribution function, $p_X = P'_X$.

Both the probability distribution function and the probability density function have multivariate counterparts.

4.8 Expected Value, Variance, and Standard Deviation

The *expectation* (*expected value*, or *mean value*) of a (scalar or vector valued) function φ of a random variable X that takes values in a set $\mathscr{X}$ is $\mathbb{E}(\varphi(X)) = \int_{\mathscr{X}} \varphi(x) p_X(x) \mathrm{d}x$ if X has a continuous probability distribution with density p_X, or $\mathbb{E}(\varphi(X)) = \sum_{x\in\mathscr{X}} \varphi(x) p_X(x)$ if X has a discrete distribution. Note that $\mathbb{E}(\varphi(X))$ can be computed without determining the probability distribution of the random variable $\varphi(X)$ explicitly.

$\mathbb{E}(X)$ indicates X's location, or the center of its probability distribution: therefore it is a most succinct summary of this distribution, and it is the best estimate of X's value in the sense that it has the smallest mean squared error.

The *median* is another indication of location for a scalar random variable: it is any value ξ such that $\Pr(X \leqslant \xi) \geqslant 1/2$ and $\Pr(X \geqslant \xi) \geqslant 1/2$, and it need not be unique. The median is the best estimate of X's value in the sense that it has the smallest absolute deviation.

Neither the mean nor the median need be "representative" values of the distribution. For example, when X denotes a proportion whose most common values are close to 0 or to 1 and its mean is close to ½, then values close to the mean are very unlikely. $\mathbb{E}(X)$ need not exist (in the sense that the defining integral or sum may fail to converge).

$\mathbb{E}(X^k)$, where k is a positive integer, is called X's kth *moment*. The *variance* of X is $\sigma^2 = \mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$, or, equivalently, the difference between its second moment and the square of its first moment. The positive square root of the variance, σ, is the *standard deviation* of X.

4.9 EXAMPLE: Poisson Distribution

The only values that a Poisson distributed random variable X can take are the non-negative integers: 0, 1, 2, ..., and the probability that its value is x is $p_X(x) = \lambda^x e^{-\lambda}/x!$, where λ is some given positive number, and $x! = x(x-1)\dots 1$. This model distributes its unit of probability into infinitely many lumps, one at each non-negative integer, so that $p_X(x)$ decreases rapidly with increasing x, and $p_X(0) + p_X(1) + p_X(2) + \cdots = 1$. Both the expected value and the variance equal λ. The number of alpha particles emitted by a sample containing the radionuclide ^{210}Po, during a period of t seconds that is a small fraction of this isotope's half-life (138 days), is a value of a Poisson random variable with mean proportional to t.

4.10 EXAMPLE: Exponential Distribution

Suppose that X represents the lifetime (thousands of hours) of an incandescent light bulb, such that, for $0 < a < b$, $\Pr(a < X < b) = \exp(-a/\eta) - \exp(-b/\eta)$, for some given number $\eta > 0$: note that, as a decreases toward 0, and b increases without limit, $\Pr(a < X < b)$ approaches 1. Focus on a particular number $x > 0$, and consider the ratio $\Pr(x - \delta < X < x + \delta)/(2\delta) = \exp(-x/\eta)[\exp(\delta/\eta) - \exp(-\delta/\eta)]/(2\delta)$ for some $\delta > 0$. As δ decreases to 0 this ratio approaches $(1/\eta)\exp(-x/\eta)$. Therefore, the function p_X such that $p_X(x) = (1/\eta)\exp(-x/\eta)$ is the probability density of the exponential distribution. In this case, the probability distribution function is P_X such that $P_X(x) = \Pr(X \leqslant x) = 1 - \exp(-x/\eta)$. X's mean value is $\mathbb{E}(X) = \eta$, and its variance is $\mathbb{V}(X) = \eta^2$. Figure 1 illustrates both the distribution function and the density for the case where $\eta = 2000\,\text{h}$.

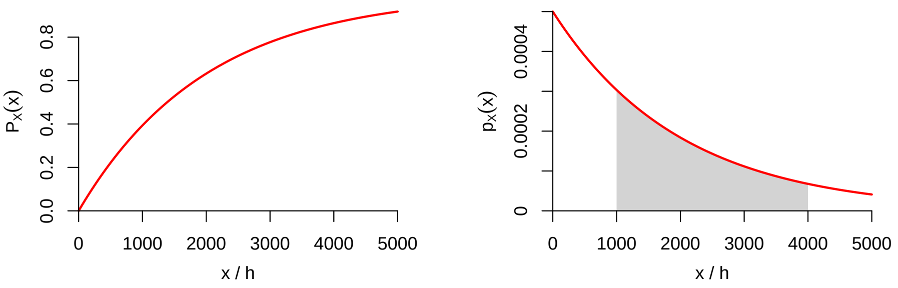

Figure 1: **Exponential Distribution with Mean** 2000 h**.** The left panel shows a portion of the graph of the probability distribution function. The right panel shows the corresponding portion of the graph of the probability density function. The area of the shaded region equals the probability of a value between 1000 h and 4000 h: this is $P_X(4000) - P_X(1000) = \exp(-1/2) - \exp(-4/2) \approx 0.47$. Note that the vertical scales in the left and right panels are different.

4.11 Joint, Marginal, and Conditional Distributions

Suppose that X represents a bivariate quantity value, for example, the Cartesian coordinates (U, V) of a point inside a circle of unit radius centered at $(0, 0)$. In this case the range $\mathscr{X}$ of $X = (U, V)$ is this unit circle. The *joint* probability distribution of U and V describes a state of knowledge about the location of X: for example, that more likely than not X is less than ½ away from the center of the circle: statements of this kind involve U and V together (that is, jointly).

The *marginal* probability distributions of U and V are the probability distributions that characterize the state of knowledge about each of them separately from the other: for example, that more likely than not $-½ < U < ½$, irrespective of V.

Clearly, the marginal distributions have to be consistent with the joint distribution, and while it is true that the joint distribution determines the marginal distributions, the reverse is not true, in that typically there are many joint distributions consistent with given marginal distributions [Possolo, 2010].

Now, suppose one knows that $U = 2/3$. This implies that $-\sqrt{5}/3 < V < \sqrt{5}/3$, hence that $X = (U, V)$ is somewhere on a particular chord $\mathscr{C}$ of the unit circle. The conditional probability distribution of V given that $U = 2/3$ is a (univariate) probability distribution over this chord.

4.12 EXAMPLE: Shark's Fin

The random variables X and Y take values in the interval $(1,2)$ and have joint probability density function $p_{X,Y}$ such that $p_{X,Y}(x,y) = (x+y)/3$ for $1 \leqslant x, y \leqslant 2$ and is zero otherwise (Figure 2): since $p_{X,Y} \geqslant 0$ and $\int_1^2 \int_1^2 p_{X,Y}(x,y) \mathrm{d}x\mathrm{d}y = 1$, $p_{X,Y}$ is a *bona fide* (bivariate) probability density.

The density of the marginal distribution of X is $p_x(x) = \int_1^2 p_{X,Y}(x,y)\mathrm{d}y = x/3 + 1/2$, and similarly for Y. And the density of the conditional distribution of Y given $X = x$ is $p_{Y|X}(y|x) = p_{X,Y}(x,y)/p_X(x) = (x+y)/(x+3/2)$.

To determine the probability density of $R = Y/X$, also depicted in Figure 2, note that $1/2 \leqslant R \leqslant 2$. First, consider the case $1/2 < r \leqslant 1$: $\Pr(R \leqslant r) = (2r-1)^2/(2r)$ and $p_R(r) = (4r^2-1)/(2r^2)$. Next, consider the case $1 < r \leqslant 2$: $\Pr(R \leqslant r) = 1-(2-r)^2/(2r)$ and $p_R(r) = (4-r^2)/(2r^2)$.

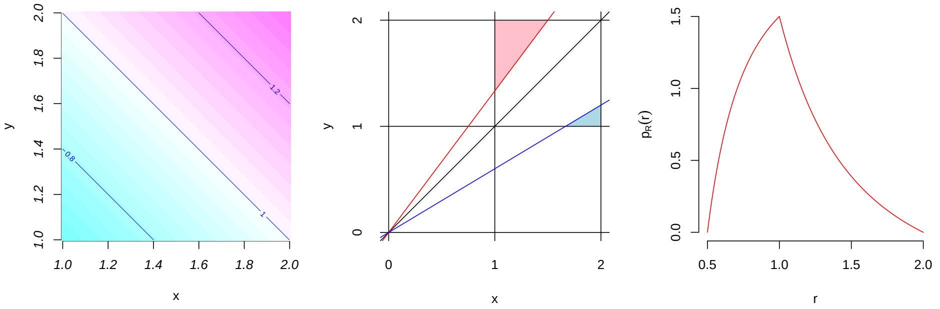

Figure 2: **Shark's Fin.** The left panel shows the probability density of the joint distribution of X and Y defined in §4.12. In the middle panel, the dashed (blue) line has slope $1/2 < r < 1$, and the points inside the small (blue) triangle have coordinates that satisfy the conditions $1 < x, y < 2$ and $y/x < r < 1$. The dotted (red) line has slope $1 < r < 2$, and the points inside the large (red) triangle have coordinates that satisfy the conditions $1 < x, y < 2$ and $y/x > r > 1$. The right panel shows the probability density function of $R = Y/X$.

4.13 Independent Random Variables

The (scalar or vectorial) random variables X and Y are independent if and only if $\Pr(X \in A \text{ and } Y \in B) = \Pr(X \in A)\Pr(Y \in B)$ for all subsets A and B in their respective ranges (to which probabilities can be coherently assigned.) Suppose X and Y have joint probability distribution with probability density function $p_{X,Y}$, and marginal density functions p_X and p_Y: the random variables are independent if and only if $p_{X,Y} = p_X p_Y$.

4.14 EXAMPLE: Unit Circle

Suppose that the probability distribution of a point is uniform inside the circle of unit radius centered at the origin $(0,0)$ of the Euclidean plane. This means that the probability that a point with Cartesian coordinates (X,Y) should lie in a subset S of this circle is proportional to S's area, but is otherwise independent of S's shape or location within the circle. The probability density function of the joint distribution of X and Y is the function $p_{X,Y}$ such that $p_{X,Y}(x,y) = 1/\pi$ if $x^2 + y^2 < 1$, and $p_{X,Y}(x,y) = 0$ otherwise. The random variables X and Y are *dependent* (§4.13): for example, if one is told that $X > ½$, then one can surely conclude that $-\sqrt{3}/2 < Y < \sqrt{3}/2$. The marginal distribution of X has density p_X such that $p_X(x) = (2/\pi)\int_0^{\sqrt{1-x^2}} \mathrm{d}y = (2/\pi)\sqrt{1-x^2}$ for $-1 < x < 1$. X has expected value 0 and standard deviation ½. Owing to symmetry, X and Y have identical marginal distributions.

4.15 Correlations and Copulas

If two or more of the random variables are dependent, then modeling their individual probability distributions will not suffice to specify their joint behavior: their joint probability distribution is needed.

One commonly used metric of dependence between two random variables X and Y is Pearson's product-moment *correlation coefficient*, defined as $\rho(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]/\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}$. However, it is possible for the variables to be dependent and still have $\rho(X,Y) = 0$.

When the only information in hand are the expected values, standard deviations, and correlations, and still one needs a joint distribution consistent with this information, then the usual course of action is to assign distributions to them individually, and then manufacturing a joint distribution using a *copula* [Possolo, 2010] — there is, however, a multitude of different copulas that can be used for this purpose, and the choice that must be made generally is influential.

5 Functions of Random Variables

5.1 Overview

If a random variable Y is a function of other random variables, $Y = \varphi(X_1, \dots, X_n)$, then φ and the joint probability distribution of $X_1, \dots, X_n$ determine the probability distribution of Y.

If only the means, standard deviations, and correlations of $X_1, \dots, X_n$ are known, then it still is possible to derive approximations to the mean and standard deviation of Y, by application of the *Delta Method*.

If the joint probability distribution of $X_1, \dots, X_n$ is known, then it may be possible to determine the probability distribution of Y analytically, using the *change of variables formula*.

In general, it is possible to obtain a sample from Y's distribution by taking a sample from the joint distribution of $X_1, \dots, X_n$ and applying φ to each element of this sample (§5.8). The results may then be summarized in several different ways: one of them is an estimate of the probability density of Y [Silverman, 1986], a procedure that is implemented in function `density` of the R environment for statistical programming and graphics [R Development Core Team, 2010].

5.2 Delta Method

If X is a random variable with mean μ and variance σ^2, φ is a differentiable real-valued function of a real variable whose first derivative does not vanish at μ, and $Y = \varphi(X)$, then $\mathbb{E}(Y) \approx \varphi(\mu)$, and $\mathbb{V}(Y) \approx [\varphi'(\mu)]^2\sigma^2$. (This results from the so-called Taylor approximation that replaces φ by a straight line tangent to its graph at μ.)

If $X = (V_1 + \dots + V_m)/m$ is an average of independent, identically distributed random variables with finite variance, then $\sqrt{m}(\varphi(V_m) - \varphi(\mu))$ also is approximately Gaussian with mean 0 and standard deviation $|\varphi'(\mu)|\sigma$, where $|\varphi'(\mu)|$ denotes the absolute value of the first derivative of φ evaluated at μ. The quality of the approximation improves with increasing m.

5.3 Delta Method — Degeneracy

When $\varphi'(\mu) = 0$ and $\varphi''(\mu)$ exists and is not zero, $X = (V_1 + \dots + V_m)/m$ is an average of independent, identically distributed random variables with finite variance, and m is large, then the probability distribution of $m(\varphi(X) - \varphi(\mu))$ is approximately like that of $\sigma^2\varphi''(\mu)Z^2/2$, where Z denotes a Gaussian (or, normal) random variable with mean 0 and standard deviation 1. Since the variance of Z^2 is 2, the standard deviation of $\varphi(V_m)$ is approximately $\sigma^2|\varphi''(\mu)|/\sqrt{2}$, rather different from what applies in the conditions of §5.2.

5.4 EXAMPLE: Radiant Power

Consider a surface whose reflectance is Lambertian: that is, light falling on it is scattered in such a way that the surface's brightness apparent to an observer is the same regardless of the observer's angle of view. The radiant power W emitted by such surface that is measured by a sensor aimed at angle A to the surface's normal is proportional to $\cos(A)$, hence one writes $W = \kappa\cos(A)$ [Cannon, 1998, Köhler, 1998].

If knowledge about the value of A is modeled by a Gaussian distribution with mean $\alpha > 0$ and standard measurement uncertainty $u(A)$ (both expressed in radians), then §5.2 (with $m = 1$) suggests that knowledge of W should be described approximately by a Gaussian distribution with mean $\kappa\cos(\alpha)$ and standard measurement uncertainty $\kappa u(A)\sin(\alpha)$.

If, for example, $\alpha = \pi/3\,\text{rad}$, $\kappa = 2\,\text{W}$, and $u(A) = \pi/100\,\text{rad}$, then the approximation to W's distribution that the Delta Method suggests, of a Gaussian distribution with mean $\cos(\pi/3) = 0.5\,\text{W}$ and standard deviation $2(\pi/100)\sin(\pi/3) = 0.0544\,\text{W}$, is remarkably accurate (Figure 3).

When the detector is aimed squarely at the target, that is $\alpha = 0$, this approximation no longer works because the first derivative of the cosine vanishes at 0, which is the degenerate case that §5.3 contemplates. In this case, W's standard measurement uncertainty is approximately $\kappa u^2(A)/\sqrt{2}$. For $\kappa = 2\,\text{W}$ and $u(A) = \pi/100\,\text{rad}$, this equals $0.0014\,\text{W}$, which is accurate to the two significant digits shown. However, Figure 3 shows that, in this case, the Delta Method produces a poor approximation.

When $\alpha = 0$, the probability density of W is markedly asymmetrical, and the meaning of W's standard deviation is rather different from its meaning when $\alpha > 0$. Indeed, when $\alpha = 0$ the probability that W should lie within one standard deviation of its expected value is 88 % approximately.

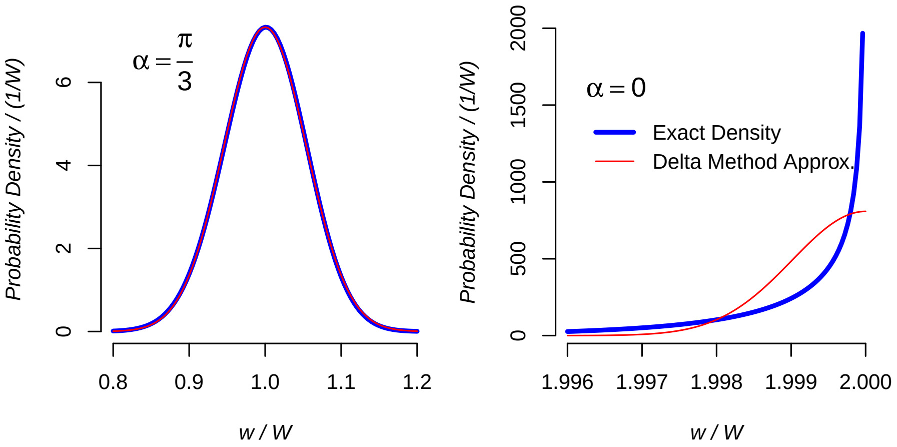

Figure 3: **Radiant Power.** Probability density of the radiant power $W = \kappa\cos(A)$ emitted by a Lambertian surface that is measured by a sensor aimed at an angle A to the surface's normal, when A has a Gaussian distribution with mean $\alpha = \pi/3$ rad (left panel) or $\alpha = 0$ rad (right panel), and standard deviation $\sigma = \pi/100$ rad. In both cases, the thick blue line is the exact density, and the thin red line is the Delta Method approximation.

5.5 Delta Method — Multivariate

The Delta Method can be extended to apply to a function of several random variables. Suppose that $X_1 = (V_{1,1}+\cdots+V_{m_1,1}), \dots, X_n = (V_{1,n}+\cdots+V_{m_n,n})$ are averages of sets of random variables whose variances are finite. The variables in each set are independent and identically distributed, those in set $1 \leqslant j \leqslant n$ having mean μ_j and variance σ_j^2. However, variables in different sets may be dependent, hence the $\{X_j\}$ may be dependent, too. Let Σ denote the $n \times n$ symmetrical matrix whose element $\sigma_{j_1 j_2} = \mathbb{E}\big((V_{m,j_1} - \mu_{j_1})(V_{i,j_2} - \mu_{j_2})\big)$ is the covariance between X_{j_1} and X_{j_2}, for $1 \leqslant j_1, j_2 \leqslant n$.

Now, consider the random variable $Y = \varphi(X_1, \dots, X_n)$, where φ denotes a real-valued function of n variables whose first partial derivatives are continuous and none vanishes at $\mu_1, \dots, \mu_n$. If $\tau^2 = \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \sigma_{j_1 j_2} (\partial\varphi/\partial\mu_{j_1})(\boldsymbol{\mu}) (\partial\varphi/\partial\mu_{j_2})(\boldsymbol{\mu})$ is finite, then $\sqrt{m}(\varphi(Y) - \varphi(\mu_1, \dots, \mu_n))$ also is approximately Gaussian with mean 0 and variance τ^2.

If $X_1, \dots, X_n$ are uncorrelated and have means $\mu_1, \dots, \mu_n$ and standard deviations $\sigma_1, \dots, \sigma_n$, then the Delta Method approximation reduces to a well-known formula first presented by Gauss [Gauss, 1823, §18, *Problema*]: $\mathbb{V}(Y) \approx c_1^2\sigma_1^2 + \cdots + c_n^2\sigma_n^2$, where the *sensitivity coefficient* $c_j = \partial\varphi(x_1,\dots,x_n)/\partial x_j$ is the value at $(x_1,\dots,x_n)$ of the jth partial derivative of φ with respect to x_j.

5.6 EXAMPLE: Beer-Lambert-Bouguer Law

If a beam of monochromatic light of power I_0 (W) travels a path of length L (m) through a solution containing a solute whose molar absorptivity for that light is E ($\mathrm{L\,mol^{-1}\,m^{-1}}$), and whose molar concentration is C ($\mathrm{mol\,L^{-1}}$), then the beam's power is reduced to I (W) such that $I = I_0 10^{ELC}$. Application of Gauss's formula (§5.5) to $C = \log_{10}(I_0/I)/(EL)$ yields:

$$\mathbb{V}(C) \approx \frac{\dfrac{\sigma_{I_0}^2}{I_0^2} + \dfrac{\sigma_I^2}{I^2}}{(EL\log 10)^2} + \left(\frac{\sigma_E^2}{E^2} + \frac{\sigma_L^2}{L^2}\right)\log_{10}^2(I/I_0)$$

5.7 EXAMPLE: Darcy's Law

Darcy's law relates the dynamic viscosity H of a fluid to the volumetric rate of discharge Q (volume per unit of time) when the fluid flows through a permeable cylindrical medium of cross-section A and intrinsic permeability K under a pressure drop of Δ along a length L, as follows: $H = KA\Delta/(QL)$.

To compute an approximation to the standard deviation of H, one may use the formula from §5.5 directly, or first take the logarithm of both sides, which linearizes the relationship, $\log(H) = \log(K) + \log(A) + \log(\Delta) - \log(Q) - \log(L)$. Applied to these logarithms, the formula from §5.5 is exact. The approximation is then done for each term separately, using the univariate Delta Method.

Since $\mathbb{V}(\log(\eta)) \approx \mathbb{V}(\eta)/\eta^2$, and similarly for the other logarithmic terms, $\mathbb{V}(\eta)/\eta^2 \approx \mathbb{V}(\kappa)/\kappa^2 \cdots + \mathbb{V}(L)/L^2$. In other words, the square of the coefficient of variation of η is approximately equal to the sum of the squares of the variation coefficients of the other variables.

5.8 Monte Carlo Method

The Monte Carlo method offers several important advantages over the Delta Method described in §5.2 and §5.5: (i) it can produce as many correct significant digits in its results as may be required; (ii) it does not involve the computation of derivatives, either analytically or numerically; (iii) it is applicable in many situations where the Delta Method is not; (iv) it provides a picture of the whole probability distribution of a function of several random variables, not just an approximation to it, or to its mean and standard deviation.

The Monte Carlo method in general dates back to the middle of the twentieth century [Metropolis and Ulam, 1949, Metropolis et al., 1953]. A variant used in mathematical statistics is known as the *parametric bootstrap* [Efron and Tibshirani, 1993]. This involves using random draws from a (possibly multivariate) probability distribution whose parameters have been replaced by estimates thereof (for example, means of posterior probability distributions, §6) to ascertain the probability distribution of a function of one (or more) random variables. Morgan and Henrion [1992] and Joint Committee for Guides in Metrology [2008b] describe how it may be employed to evaluate measurement uncertainty, and provide illustrative examples.

The procedure comprises the following steps:

MC1 Define the joint probability distribution of $X_1, \dots, X_n$.

MC2 Choose a suitably large positive integer K and draw a sample of size K from this joint distribution to obtain $(x_{11}, \dots, x_{n1})$, …, $(x_{1K}, \dots, x_{nK})$. (If X_1, …, X_n happen to be independent, then this amounts to drawing a sample of size K from the distribution of each of them separately.)

MC3 Compute $y_1 = \varphi(x_{11}, \dots, x_{n1})$, …, $y_K = \varphi(x_{1K}, \dots, x_{nK})$, which are a sample from Y's distribution.

MC4 Summarize this sample in one or more of these different ways:

MC4.a — Probability Density The most inclusive summarization is in the form of an estimate of Y's probability density function: this may be either a simple histogram, or a kernel density estimate [Silverman, 1986].

MC4.b — Mean and Standard Deviation The mean and standard deviation of Y are estimated by the mean and the standard deviation of $\{y_1, \dots, y_K\}$. (To ascertain the number of significant digits in this mean and standard deviation, hence to decide whether K is large enough for the intended purpose, or should be increased, one may employ either the adaptive procedure explained in the Supplement 1 to the GUM [Joint Committee for Guides in Metrology, 2008b, 7.9], or resort to the non-parametric statistical bootstrap or to other resampling methods [Davison and Hinkley, 1997].)

MC4.c — Probability Interval If $y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(K)}$ denote the result of ordering $y_1, \dots, y_K$ from smallest to largest, then the interval $(y_{(K\alpha/2)}, y_{(K(1-\alpha/2))})$ includes Y's true value with probability $1 - \alpha$. (Since $K\alpha/2$ and $K(1-\alpha/2)$ need not be integers, the end-points of this coverage interval may be calculated by interpolation of adjacent $y_{(i)}$s.)

5.9 EXAMPLE: Volume of Cylinder

The radius R and the height H of a cylinder are values of independent random variables with exponential probability distributions with mean 1 m. To characterize the probability distribution of its volume $V = \pi R^2 H$, draw a sample of size 10^7 from the joint distribution of R and H, and compute the volume corresponding to each pair of sampled values (r, h) to obtain a sample of the same size from the distribution of V. The average and standard deviation of these values are $6.3\,\text{m}^3$ and $21\,\text{m}^3$, and they are estimates (whose two most significant digits are exact) of the mean and standard deviation of V. Figure 4 depicts an estimate of the corresponding probability density.

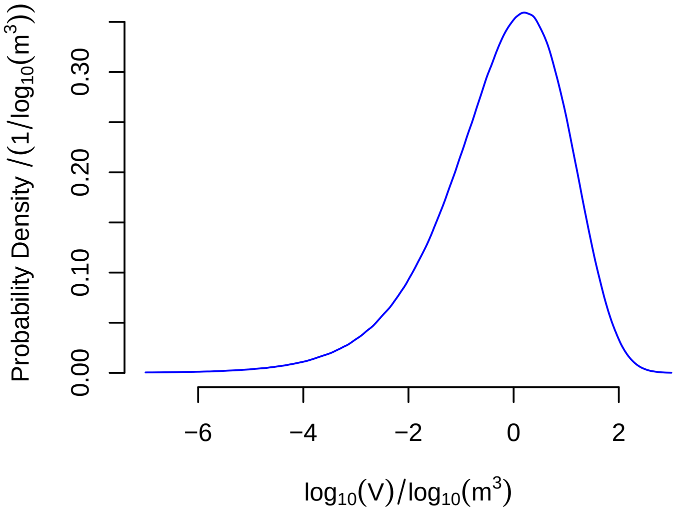

Figure 4: **Cylinder Volume.** Kernel density estimate [Silverman, 1986] of the probability density of the volume of a cylinder whose radius and height are realized values of independent, exponentially distributed random variables with mean 1 m.

5.10 Change-of-Variable Formula — Univariate

Suppose that X is a random variable with a continuous distribution and values in $\mathscr{X}$, with probability distribution function P_X and probability density function p_X, and consider the random variable $Y = \varphi(X)$ where φ denotes a real-valued function of a real variable. Let $\mathscr{Y}$ denote the set where Y takes its values, and let P_Y and p_Y denote Y's probability distribution function and probability density function, respectively. In these circumstances [Casella and Berger, 2002, Chapter 2].

- If φ is increasing on $\mathscr{X}$ and ψ denotes its inverse, then $P_Y(y) = \Pr(Y \leqslant y) = \Pr(X \leqslant \psi(y)) = P_X[\psi(y)]$ for $y \in \mathscr{Y}$; and if φ is decreasing, then $P_Y(y) = 1 - P_X[\psi(y)]$.
- If φ is either increasing or decreasing on $\mathscr{X}$ (but not both), and its inverse ψ has a continuous first derivative $\dot{\psi}$, then $p_Y(y) = p_X[\psi(y)]|\dot{\psi}(y)|$, for $y \in \mathscr{Y}$, where $|\dot{\psi}(y)|$ denotes the absolute value of the derivative of ψ at y.

5.11 EXAMPLE: Oscillating Mirror

A horizontal beam of light emerges from a tiny hole in a wall and travels along a 1 m long path at right angles to the wall, towards a flat mirror that oscillates freely around a vertical axis. When the mirror's surface normal makes an angle A with the beam, its reflection hits the wall at distance $D = \tan(A)$ from the hole (positive to the right of the hole and negative to the left). If A is uniformly (or, rectangularly) distributed between $-\pi/2\,\text{rad}$ and $\pi/2\,\text{rad}$, then $P_D(d) = \Pr(D \leqslant d) = \Pr(A \leqslant \arctan(d)) = (\arctan(d) + \pi/2)/\pi$, and D's probability density is p_D such that $p_D(d) = 1/[\pi(1 + d^2)]$ for $-\infty < d < \infty$. As it turns out, both the mean and the standard deviation of D are infinite [Feller, 1971, Page 51].

5.12 Change-of-Variable Formula — Multivariate

Suppose that $X_1, \dots, X_n$ are random variables and consider $Y_j = \varphi_j(X_1, \dots, X_n)$ for $j = 1, \dots, n$, and where $\varphi_1, \dots, \varphi_n$ are real-valued functions of n real variables each.

Suppose also that (i) the vector $\boldsymbol{X} = (X_1, \dots, X_n)$ takes values in an open subset $\mathscr{X}$ of n-dimensional Euclidean space, and has a continuous joint probability distribution with probability density function $p_{\boldsymbol{X}}$; (ii) the vector-valued function $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_n)$ is invertible, and the inverse $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$ has a Jacobian determinant $J_{\boldsymbol{\psi}}$ that does not vanish on $\mathscr{Y}$.

CV1 Solve the n equations $y_1 = \varphi_1(x_1, \ldots, x_n)$, $\ldots$, $y_n = \varphi_n(x_1, \ldots, x_n)$, for $x_1, \ldots, x_n$, to obtain the inverse transformation such that $x_1 = \psi_1(y_1, \ldots, y_n)$, $\ldots$, $x_n = \psi_n(y_1, \ldots, y_n)$.

CV2 Find $\dot{\psi}_{ij}$, the partial derivative of ψ_i with respect to its jth argument, for $i, j = 1, \ldots, n$, and compute the Jacobian determinant of the inverse transformation at $\boldsymbol{y} = (y_1, \ldots, y_n)$:

$$J_{\boldsymbol{\psi}}(\boldsymbol{y}) = \det \begin{bmatrix} \dot{\psi}_{11}(\boldsymbol{y}) & \dot{\psi}_{12}(\boldsymbol{y}) & \ldots & \dot{\psi}_{1n}(\boldsymbol{y}) \\ \dot{\psi}_{21}(\boldsymbol{y}) & \dot{\psi}_{22}(\boldsymbol{y}) & \ldots & \dot{\psi}_{2n}(\boldsymbol{y}) \\ \vdots & \vdots & \ddots & \vdots \\ \dot{\psi}_{n1}(\boldsymbol{y}) & \dot{\psi}_{n2}(\boldsymbol{y}) & \ldots & \dot{\psi}_{nn}(\boldsymbol{y}) \end{bmatrix}$$

CV3 The density the joint probability distribution of the random vector $\boldsymbol{Y}$ is $p_{\boldsymbol{Y}}$ such that

$$p_{\boldsymbol{Y}}(\boldsymbol{y}) = p_{\boldsymbol{X}}\left[\psi(\boldsymbol{y})\right] \left|J_{\boldsymbol{\psi}}(\boldsymbol{y})\right|. \qquad (2)$$

Note that $J_{\boldsymbol{\psi}}(\boldsymbol{y})$ is a scalar, and $\left|J_{\boldsymbol{\psi}}(\boldsymbol{y})\right|$ denotes its absolute value.

CV4 The probability density of Y_1 is $p_{Y_1}(y_1) = \int \ldots \int g(y_1, y_2, \ldots, y_n) \mathrm{d}y_2 \ldots \mathrm{d}y_n$, where the $n-1$ integrals are over the ranges of $Y_2, \ldots, Y_n$.

5.13 EXAMPLE: Linear Combinations of Gaussian Random Variables

Suppose that U and V are independent, Gaussian random variables with mean 0 and variance 1, and let $S = aU + bV$, and $T = bU - aV$, for given real numbers a and b. The inverse transformation maps (s, t) onto $\big((as+bt)/(a^2+b^2), (bs-at)/(a^2+b^2)\big)$, and has Jacobian determinant $\det \left[\begin{smallmatrix} a & b \\ b & -a \end{smallmatrix}\right] / (a^2+b^2)$ whose absolute value is $1/(a^2+b^2)$. Since the density of the joint probability distribution of U and V is $p_{U,V}(u, v) = \exp(-(u^2+v^2)/2)/(2\pi)$, application of the multivariate change-of-variable formula yields

$$p_{S,T}(s,t) = \frac{\exp\left\{-\dfrac{s^2}{2(a^2+b^2)}\right\}}{\sqrt{2\pi(a^2+b^2)}} \frac{\exp\left\{-\dfrac{t^2}{2(a^2+b^2)}\right\}}{\sqrt{2\pi(a^2+b^2)}}.$$

But this means that S and T also are independent and Gaussian with mean 0 and variance a^2+b^2. On the one hand, this result is surprising because S and T both are functions of the same random variables. On the other hand it is hardly surprising because the transformation amounts to a rotation of the coordinate axes, followed by a global dilation. Since the joint distribution of U and V is circularly symmetric relative to $(0,0)$, so will the joint distribution of S and T be, which implies independence and the same functional form for the density, up to a difference in scale.

5.14 EXAMPLE: Ratio of Exponential Lifetimes

To compute the probability density of the ratio $R = X/Y$ of two independent and exponentially distributed random variables X and Y with mean $1/\lambda$, define the function $\boldsymbol{\varphi}$ such that $\boldsymbol{\varphi}(x,y) = (x/y, y)$, whose inverse is $\boldsymbol{\psi}(r,s) = (rs, s)$, with Jacobian determinant $J_{\boldsymbol{\psi}}(r,s) = \det\left[\begin{smallmatrix} s & r \\ 0 & 1 \end{smallmatrix}\right] = s > 0$. The multivariate change-of-variable formula then yields $p_{R,S}(r,s) = s\lambda^2 \exp\left[-\lambda(1+r)s\right]$ for the density of the joint distribution of $R = X/Y$ and $S = Y$. The (marginal) density of R is $p_R(r) = \int_0^\infty p_{R,S}(r,s)\mathrm{d}s = 1/(1+r)^2$, for $r > 0$, being 0 otherwise.

p_R indeed is a probability density function because it is a non-negative function and $\int_0^\infty \mathrm{d}r/(1+r)^2 = 1 - \lim_{r\to\infty} 1/(1+r) = 1$. However, since $\int_0^\infty r\mathrm{d}r/(1+r)^2 = \lim_{r\to\infty} \log(1+r) = \infty$, neither the mean nor the variance of R is finite. The Delta Method, however, would have suggested that $\mathbb{E}(X/Y) \approx 1$ and that the coefficient of variation (ratio of the standard deviation to the mean) of X/Y is $\sqrt{2}$ approximately.

5.15 EXAMPLE: Polar Coordinates

The inverse of the transformation that maps the polar coordinates (r,α) of a point in the Euclidean plane to its Cartesian coordinates (x,y) is ψ such that $\psi(r,a) = (r\cos a, r\sin a)$ for $r > 0$ and $0 < a < 2\pi$, with Jacobian determinant $J_\psi(r,\alpha) = r$. If X and Y are independent Gaussian random variables with mean 0 and variance 1, then the probability density of (R,A) is $p_{R,A}(r,\alpha) = r\exp(-r^2/2)/(2\pi)$. Since $\int_0^\infty r\exp(-r^2/2) = 1$, it follows that R and A are independent, the former having a Rayleigh distribution with mean $\sqrt{\pi/2}$, the latter a uniform distribution between 0 and 2π.

6 Statistical Inference

The statistical inferences we are primarily interested in are probabilistic statements about the unknown value of a quantity, produced by application of a statistical method. In the example of §6.6, one of the inferences is this statement: *the probability is 95 % that the difference between the numbers of hours gained with the two soporifics is between* 0.7 h *and* 2.5 h.

Another common inference is an estimate of the value of a quantity, which must be qualified with an assessment of the associated uncertainty. In the example of §6.8, a typical inference of this kind would be this: *the difference in mean levels of thyroxine in the serum of two groups of children diagnosed with hypothyroidism is estimated as* 14 nmol/L *with standard uncertainty* 18 nmol/L (Figure 6).

In our treatment of this example, the inference is based entirely on a small set of empirical data, and on a particular choice of statistical model used to describe the dispersion of the data, and to characterize the fact that no knowledge other than the data was brought into play.

Statistical methods different from the one we used could have been employed: some of these would produce the same result (in particular those illustrated when this dataset was first described [Student, 1908, Fisher, 1973]), while others would have produced different results.

Even when the result is the same, it may be variously interpreted:

- For some that statement means that if the same sampling and study method is used repeatedly, and each time the resulting dataset is modeled and analyzed in the same way to produce an interval like the one above, then about 95 % of the resulting intervals will include the true difference sought — with no guarantee or implication that the interval that was obtained is one of these;
- For others (among whom we stand) that statement expresses the degree of belief one is entitled to have about the true difference lying between 0.7 h and 2.5 h specifically, in light of all the relevant information in hand.

6.1 Bayesian Inference

Bayesian inference [Bernardo and Smith, 2000, Lindley, 2006, Robert, 2007] is a class of statistical procedures that serve to blend preexisting information about the value of a quantity with fresh information in empirical data.

The defining traits of a Bayesian procedure are these:

(i) All quantity values that are the objects of interest but are accessible to direct observation (*non-observables*) are modeled as values of non-observable random variables whose (prior, or *a priori*) distributions encode and convey states of incomplete knowledge about those values;

(ii) The empirical data (*observables*) are modeled as realized values of random variables whose probability distributions depend on those objects of interest;

(iii) Preexisting information about those objects of interest is updated in light of the fresh empirical data by application of Bayes rule, and the results are encapsulated in a (posterior, or *a posteriori*) probability distribution;

(iv) Selected aspects of this distribution are then abstracted from it and used to characterize the objects of interest and to describe the state of knowledge about them.

6.2 Prior Distribution

Let θ denote the value of the quantity of interest, which we model as realized value of a random variable Θ with probability density function p_Θ that encodes the state of knowledge about θ prior to obtaining fresh data, and which must be defined even if there is no prior knowledge.

Defining such p_Θ often is a challenging task. If in fact there exists substantial prior knowledge about θ, then it needs to be elicited from experts in the matter and encapsulated in the form of a particular probability density: Garthwaite et al. [2005] review how this may be done. For example, when measuring the mass fraction of titanium in a mineral specimen, then knowledge of the species (ilmenite, titanite, rutile, etc.) of the specimen is highly informative about that mass fraction. Familiarity with the process of analytical chemistry employed to make the measurement may indicate the dispersion of values to be expected.

In some cases, essentially no prior knowledge exists about θ, or none is deemed reliable enough to be taken into account. In such cases, a so-called non-informative prior distribution needs to be produced and assigned to Θ that reflects this state of affairs: if θ is univariate (that is, a single number), then the rules developed by Jeffreys [1961] often prove satisfactory; if θ is multivariate (that is, a numerical vector), then the so-called *reference* prior distributions are recommended [Bernardo and Smith, 2007] (these reduce to Jeffreys's in the univariate case).

These rules often produce a p_Θ that is *improper*, in the sense that $\int_{\mathscr{H}} p(\theta)\mathrm{d}\theta$ diverges to infinity, where $\mathscr{H}$ denotes the range of Θ. (If Θ should have a discrete distribution then this integral is replaced by a sum.) Fortunately, once used in Bayes Rule (§3.5 and §6.4), improper priors often lead to proper posterior probability distributions.

6.3 Likelihood Function

The empirical data x (which may be a single number, a numerical vector, or a data structure of still greater complexity) are modeled as realized values of a random variable X whose probability density describes the corresponding dispersion of values.

This density must depend on θ, which is another way of saying that the data are informative about θ (otherwise there would be nothing to be gained by observing them). In fact, this is the density of the conditional probability distribution of X given that $\Theta = \theta$. Choosing a specific functional form for it generally is a non-trivial exercise: it involves defining a statistical model that correctly captures the dispersion of values likely to be obtained in the experiment that produces them.

Once the data x are in hand, $p_{X|\Theta}(x|\theta)$ becomes a function of θ alone, being largest for values of θ that make the data appear most likely. As such, it still is non-negative, but its integral (or sum, if X's distribution should be discrete) over the range of Θ, need not be 1.

6.4 Posterior Distribution

Suppose that both X given that $\Theta = \theta$ and Θ have continuous distributions with densities $p_{X|\Theta}$ and p_Θ. In these circumstances, Bayes rule becomes

$$p_{\Theta|X}(\theta) = \frac{p_{X|\Theta}(x|\theta)p_\Theta(\theta)}{\int_{\mathscr{H}} p_{X|\Theta}(x|s)p_\Theta(s)\mathrm{d}s}. \tag{3}$$

The function $p_{\Theta|X}$, which is defined over the range of Θ for each fixed value x, is the density of the posterior distribution of the value of the quantity of interest Θ given the data.

In some cases this can be computed in closed form, in many others it cannot. In all cases it is possible to obtain a sample from this posterior distribution by application of a procedure known as Markov Chain Monte Carlo (MCMC)

[Gelman et al., 2003]. This sample can then be summarized as described in §5.8.

6.5 EXAMPLE: Viral Load

Once infected by influenza A virus, an epithelial cell of the upper respiratory tract releases θ virions on average, which may then go on to infect other cells. This number θ depends on the volume of the cell, and we will treat it as realized value of a non-observable random variable with an exponential distribution whose expected value, $1/\gamma$ for some $0 < \gamma < 1$, is known. Given θ, the actual number of virions that are released is x, and this is like a realized value of a Poisson random variable with mean θ.

Suppose that the prior density is $p_\Theta(\theta) = \gamma \exp(-\gamma\theta)$, and the likelihood function is $L_x(\theta) = p_{X|\Theta}(x|\theta) = \theta^x \exp(-\theta)/x!$, for $\theta > 0$. The posterior distribution of Θ given x belongs to the gamma family, and has expected value $(x+1)/(\gamma+1)$, variance $(x+1)/(\gamma+1)^2$, and density

$$p_{\Theta|X}(\theta|x) = \frac{\frac{\theta^x e^{-\theta}}{x!}\gamma e^{-\gamma\theta}}{\int_0^\infty \frac{s^x e^{-s}}{x!}\gamma e^{-\gamma s}\mathrm{d}s} = \frac{(\gamma+1)^{x+1}\theta^x e^{-\theta(\gamma+1)}}{x!}.$$

6.6 EXAMPLE: Sleep Hours

The differences between the numbers of additional hours of sleep that ten patients gained when using two soporific drugs, described in examples given by Student [1908] and [Fisher, 1973, §24], were 1.2 h, 2.4 h, 1.3 h, 1.3 h, 0.0 h, 1.0 h, 1.8 h, 0.8 h, 4.6 h, and 1.4 h.

Suppose that, given μ and σ, these are realized values of independent Gaussian random variables with mean μ and variance σ^2. Let $\overline{x} = 1.58\,\mathrm{h}$ denote their average, and $s^2 = 1.51\,\mathrm{h}^2$ denote the sum of their squared deviations from $\overline{x}$ divided by 9. In these circumstances, the likelihood function is $L_{\overline{x},s^2}(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\{-[n(\overline{x}-\mu)^2 + (n-1)s^2]/(2\sigma^2)\}$.

Assume, in addition, that μ and σ are realized values of non-observable random variables M and Σ that are independent *a priori* and such that M and $\log\Sigma$ are uniformly distributed between $-\infty$ and $+\infty$ (both improper prior distributions). Then, given $\overline{x}$ and s, $(\mu - \overline{x})/(s/\sqrt{n})$ is like a realized value of a random variable with a Student's t distribution with $n-1 = 9$ degrees of freedom, and $(n-1)s^2/\sigma^2$ is like a realized value of a random variable with a chi-squared distribution with $n-1$ degrees of freedom [Box and Tiao, 1973, Theorem 2.2.1].

Therefore, the expected value of the posterior distribution of the mean difference of hours of sleep gained is $\widetilde{\mu} = 1.58\,\text{h}$, and the standard deviation is $(s/\sqrt{n})\left(\sqrt{(n-1)/(n-3)}\right) = 0.441\,\text{h}$. A 95 % probability interval for μ ranges from 0.7 h to 2.5 h, and a similar one for σ ranges from 0.8 h to 2.2 h.

Suppose that $M_{\overline{x},s}$ and $\Sigma_{\overline{x},s}$ are the counterparts of M and Σ once the information in the data has been taken into account: that is, their probability distribution is the joint (or, bivariate) posterior probability distribution given the data. Even though M and Σ were assumed to be independent *a priori*, $M_{\overline{x},s}$ and $\Sigma_{\overline{x},s}$ turn out to be dependent *a posteriori* (that is, given the data), but their correlation is zero [Lindley, 1965b, §5.4].

6.7 EXAMPLE: Hurricanes

A major hurricane has category 3, 4, or 5 on the Saffir-Simpson Hurricane Scale [Simpson, 1974]: its central pressure is no more than 945 mbar (94 500 Pa), it has winds of at least 111 mile/hour ($49.6\,\text{m}\,\text{s}^{-1}$), generates sea surges of 9 feet (2.7 m) or greater, and has the potential to cause extensive damage.

The numbers of major hurricanes that struck the U.S. mainland directly, in each decade starting with 1851–1860 and ending with 1991–2010, are: 6, 1, 7, 5, 8, 4, 7, 5, 8, 10, 8, 6, 4, 4, 5, 7 [Blake et al., 2011]. Let $n = 16$ denote the number of decades, $x_1, \dots, x_n$ denote the corresponding counts, and $s = x_1 + \cdots + x_n$. Suppose that one wishes to predict y, the number of such hurricanes in the decade 2011–2020.

Assume that the mean number of such hurricanes per decade will have remained constant between 1851 and 2010 (certainly a questionable assumption), with unknown value λ, and that, conditionally on this value, $x_1, \dots, x_n$, and y are realized values of independent Poisson random variables $X_1, \dots, X_n$ (observable), Y (non-observable), all with mean value λ: their common probability density is $p_{X|\Lambda}(k|\lambda) = \lambda^k \exp(-\lambda)/k!$ for $k = 0, 1, 2, \dots$. This model is commonly used for phenomena that result from the cumulative effect of many improbable events [Feller, 1968, XI.6b].

Even though the goal is to predict Y, the fact that there is no *a priori* knowledge about λ other than that it must be positive, requires that this be modeled as the (non-observable) value of a random variable Λ whose probability distribution must reflect this ignorance. (According to the Bayesian paradigm, *all* states of knowledge, even complete ignorance, have to be modeled using probability distributions.)

If the prior distribution chosen for Λ is the *reference* prior distribution [Berger, 2006, Bernardo, 1979, Bernardo and Smith, 2000], then the value of its probability density p_Λ at λ should be proportional to $1/\sqrt{\lambda}$ [Bernardo and Smith, 2000, A.2], an improper prior probability density. However, the corresponding posterior distribution for Λ is proper, in fact it is a gamma distribution with expected value $(s+½)/n$ and probability density function $p_{\Lambda|X_1,\dots,X_n}$ such that

$$p_{\Lambda|X_1,\dots,X_n}(\lambda|x_1,\dots,x_n) = \frac{\dfrac{\lambda^s \exp(-\lambda n)}{x_1!\dots x_n!}\dfrac{1}{\sqrt{\lambda}}}{\displaystyle\int_0^{+\infty} \frac{l^s \exp(-ln)}{x_1!\dots x_n!}\frac{1}{\sqrt{l}}\mathrm{d}l} = \frac{n^{s+½}}{\Gamma(s+½)}\lambda^{s-½}\exp(-\lambda n). \tag{4}$$

However, what is needed for the aforementioned prediction is the conditional distribution of Y given the observed counts: the so-called *predictive* distribution [Schervish, 1995, Page 18]. If π denotes the corresponding density, then

$$\begin{aligned}\pi(y\,|\,x_1,\dots,x_n) &= \int_0^{+\infty} g(y\,|\,\lambda)p_{\Lambda|X_1,\dots,X_n}(\lambda\,|\,x_1,\dots,x_n)\mathrm{d}\lambda \\ &= \frac{n^{s+½}}{\Gamma(s+½)}\frac{\Gamma(y+s+½)}{y!(n+1)^{y+s+½}}. \end{aligned} \tag{5}$$

This defines a discrete probability distribution on the non-negative integers, often called a Poisson-gamma mixture distribution [Bernardo and Smith, 2000, §3.2.2]. For our data, since π achieves a maximum at $y = 5$ (Figure 5), this is the (*a posteriori*) most likely number Y of major hurricanes that will hit the U.S. mainland in 2011–2020. The mean of the posterior distribution is 6. Since the probability is 0.956 that Y's value lies between 2 and 11 (inclusive), the interval whose end-points are 2 and 11 is a 95.6 % coverage interval for Y.

6.8 EXAMPLE: Hypothyroidism

[Altman, 1991, Table 9.6] lists measurement results from Hulse et al. [1979], for the concentration of thyroxine in the serum of sixteen children diagnosed with hypothyroidism, of which nine had slight or no symptoms, and the other seven had marked symptoms. The values measured for the former, all in units of nmol/L, were 34, 45, 49, 55, 58, 59, 60, 62, and 86; and for the latter they were 5, 8, 18, 24, 60, 84, and 96. The averages are $\overline{x} = 56.4$ and $\overline{y} = 42.1$, and the standard deviations are $s = 14.2$ and $t = 37.5$.

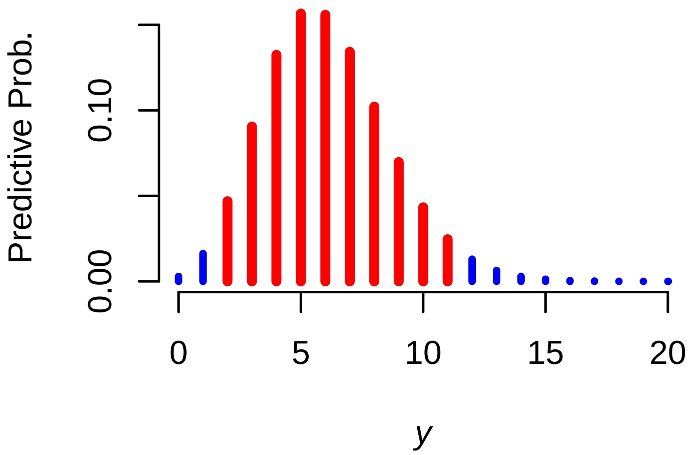

Figure 5: **Hurricanes.** Predictive probabilities for the number of major hurricanes that will hit the U.S. mainland in 2001–2010. The vertical axis indicates values of $\pi(y \mid x_1, \dots, x_n)$ from equation (5). The most likely number is 5, the expected number is 6, and the probability is 0.956 that the number will be between 2 and 11 (inclusive) (red bars).

Our goal is to produce a probability interval for the difference between the corresponding means, μ and ν, say, when nothing is assumed known *a priori* either about these means or about the corresponding standard deviations, σ and τ, which may be different.

Given the values of these four parameters, suppose that the values measured in the $m = 9$ children with slight or no symptoms are observed values of independent Gaussian random variables $U_1, \dots, U_m$ with common mean μ and standard deviation σ, and that those measured in the $n = 7$ children with marked symptoms are observed values of independent Gaussian random variables $V_1, \dots, V_n$, also independent of the $\{U_i\}$, with common mean ν and standard deviation τ.

The problem of constructing a probability interval for $\mu - \nu$ under these circumstances is known as the *Behrens-Fisher* problem [Ghosh and Kim, 2001]. For the Bayesian solution, we regard μ, ν, σ and τ as realized values of non-observable random variables M, N, Σ, and T, assumed independent *a priori* and such that M, N, $\log \Sigma$, and $\log T$ all are uniformly distributed over the real numbers (hence have improper prior distributions). The corresponding posterior distributions all are proper provided $m \geqslant 2$ and $n \geqslant 2$. However, the density of the posterior probability distribution of $M - N$ given the data cannot be computed in closed form.

This problem in Bayesian inference, and other problems much more demanding than this, can be solved using the MCMC sampling technique mentioned in §6.4, for which there exist several generic software implementations: we obtained the results presented below using function `metrop` of the R package `mcmc` [Geyer, 2010]. Typically, all that is needed is the logarithm of the numerator of Bayes formula (3). Leaving out constants that do not involve μ, ν, σ or τ, this is

$$-(m+1)\log(\sigma)-(n+1)\log(\tau)-\frac{m(\mu-\overline{x})^2+(m-1)s^2}{2\sigma^2}-\frac{n(\nu-\overline{y})^2+(n-1)t^2}{2\tau^2}.$$

MCMC produces a sample of suitably large size K from the joint posterior distribution of M, N, Σ, and T, given the data, say $(\mu_1, \nu_1, \sigma_1, \tau_1), \dots, (\mu_K, \nu_K, \sigma_K, \tau_K)$. The 95 % probability interval for the difference in mean levels of thyroxine in the serum of the two groups, which extends from -22 nmol/L to 51 nmol/L, and Figure 6, are based on a sample of size $K = 4.5 \times 10^6$. The probability density in this figure, and that probability interval, were computed as described in MC4.c and MC4.d of §5.8, only applied to the differences $\{\mu_k - \nu_k\}$.

In this particular case it is possible to ascertain the correctness of the results owing to an interesting, albeit surprising result: the posterior means are independent *a posteriori*, and have probability distributions that are re-scaled, shifted versions of Student's t distributions with $m-1$ and $n-1$ degrees of freedom [Box and Tiao, 1973, 2.5.2]. Therefore, by application of the Monte Carlo method of §5.8, one may obtain a sample from the posterior distribution of the difference $M - T$ independently of the MCMC procedure: the results are depicted in Figure 6 (where they are labeled "Jeffreys (Exact)"), and are essentially indistinguishable from the results of MCMC.

This same figure shows yet another posterior density that differs hardly at all from the posterior density corresponding to Jeffreys's prior: this alternative result corresponds to the "matching" prior distribution (also improper) derived by Ghosh and Kim [2001], whose density is proportional to $(\sigma^2/m+\tau^2/n)/(\sigma\tau)^3$. This illustrates a generally good practice: that the sensitivity of the results of Bayesian analysis should be evaluated by comparing how they vary when different but comparably acceptable priors are used.

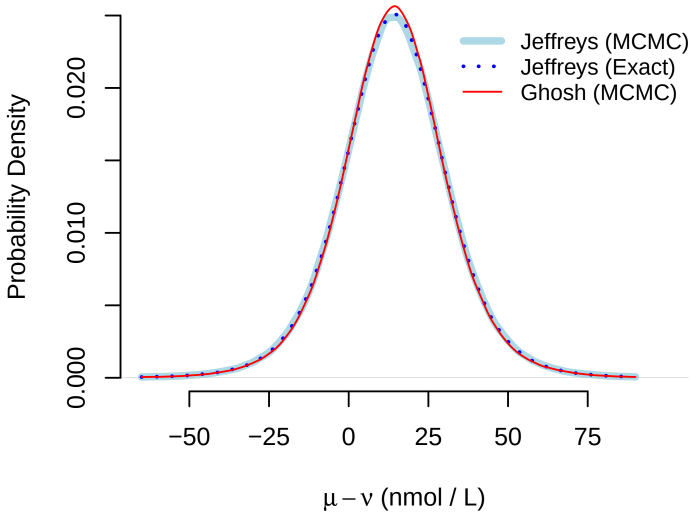

Figure 6: **Hypothyroidism.** Probability density of the posterior distribution of the difference in mean levels of thyroxine in the serum of two groups of children diagnosed with hypothyroidism: computed via MCMC using either Jeffreys's prior or the "matching" prior of Ghosh and Kim [2001], alongside the exact version corresponding to Jeffreys's prior. The posterior mean and standard deviation are 14 nmol/L and 18 nmol/L.

7 Acknowledgments

Antonio Possolo thanks colleagues and former colleagues from the Working Group 1 of the Joint Committee for Guides in Metrology, who offered valuable comments and suggestions in multiple discussions of an early draft of a document that he had submitted to the consideration of this Working Group and that included material now in the present document: Walter Bich, Maurice Cox, René Dybkær, Charles Ehrlich, Clemens Elster, Tyler Estler, Brynn Hibbert, Hidetaka Imai, Willem Kool, Lars Nielsen, Leslie Pendrill, Lorenzo Peretto, Steve Sidney, Adriaan van der Veen, Graham White, and Wolfgang Wöger. Antonio Possolo is particularly grateful to Tyler Estler for suggestions regarding §2.3 and §3.1, and to Graham White for suggestions and corrections that greatly improved §3.6, on the tuberculin test. Both authors thank their common NIST colleagues Andrew Rukhin and Jack Wang for suggesting many corrections and improvements. Mary Dal-Favero and Alan Heckert, both from NIST, kindly fa-

cilitated the deployment of this material on the World Wide Web.

References

D. G. Altman. *Practical Statistics for Medical Research*. Chapman & Hall/CRC, Boca Raton, FL, 1991. Reprinted 1997.

American Thoracic Society. Diagnostic standards and classification of tuberculosis in adults and children. *American Journal of Respiratory and Critical Care Medicine*, 161:1376–1395, 1999.

J. Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3): 385–402, 2006. URL `http://ba.stat.cmu.edu/`.

J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley & Sons, New York, 2000.

J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley & Sons, Chichester, England, 2nd edition, 2007.

J. M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society*, 41:113–128, 1979.

E. S. Blake, C. W. Landsea, and E. J. Gibney. The deadliest, costliest, and most intense United States tropical cyclones from 1851 to 2010 (and other frequently requested hurricane facts). Technical Report Technical Memorandum NWS NHC-6, NOAA, National Weather Service, National Hurricane Center, Miami, Florida, August 2011.

L. Bovens and W. Rabinowicz. Democratic answers to complex questions — an epistemic perspective. *Synthese*, 150:131–153, 2006.

G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, Massachusetts, 1973.

T. W. Cannon. Light and radiation. In C. DeCusatis, editor, *Handbook of Applied Photometry*, chapter 1, pages 1–32. Springer Verlag, New York, New York, 1998.

R. Carnap. *Logical Foundations of Probability*. University of Chicago Press, Chicago, Illinois, 2nd edition, 1962.

G. Casella and R. L. Berger. *Statistical Inference*. Duxbury, Pacific Grove, California, 2nd edition, 2002.

N. Clee. Who's the daddy of them all? In *Observer Sport Monthly*. Guardian News and Media Limited, Manchester, UK, Sunday March 4, 2007.

R. T. Clemen and R. L. Winkler. Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19:187–203, 1999.

R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14:1–13, 1946.

R. T. Cox. *The Algebra of Probable Inference*. The Johns Hopkins Press, Baltimore, Maryland, 1961.

A. C. Davison and D. Hinkley. *Bootstrap Methods and their Applications*. Cambridge University Press, New York, NY, 1997.

B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Anales de l'Institut Henri Poincaré*, 7:1–68, 1937.

B. de Finetti. *Theory of Probability: A critical introductory treatment*. John Wiley & Sons, Chichester, 1990. Two volumes, translated from the Italian and with a preface by Antonio Machì and Adrian Smith, with a foreword by D. V. Lindley, Reprint of the 1975 translation.

M. H. DeGroot. A conversation with Persi Diaconis. *Statistical Science*, 1(3): 319–334, August 1986.

M. H. DeGroot and M. J. Schervish. *Probability and Statistics*. Addison-Wesley, 4th edition, 2011.

P. Diaconis and D. Ylvisaker. Quantifying prior opinion. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. Smith, editors, *Bayesian Statistics*, volume 2, pages 163–175. North-Holland, Amsterdam, 1985.

F. W. Dyson, A. S. Eddington, and C. Davidson. A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society of London, Series A, Containing Papers of a Mathematical or Physical Character*, 220:291–333, 1920.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, London, UK, 1993.

W. Feller. *An Introduction to Probability Theory and Its Applications*, volume I. John Wiley & Sons, New York, 3rd edition, 1968. Revised Printing.

W. Feller. *An Introduction to Probability Theory and Its Applications*, volume II. John Wiley & Sons, New York, 2nd edition, 1971.

R. A. Fisher. *Statistical Methods for Research Workers*. Hafner Publishing Company, New York, NY, 14th edition, 1973.

B. Fitelson. Likelihoodism, bayesianism, and relational confirmation. *Synthese*, 156(3):473–489, 2007.

P. H. Garthwaite, J. B. Kadane, and A. O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100: 680–701, June 2005.

C. F. Gauss. Theoria combinationis observationum erroribus minimis obnoxiae. In *Werke, Band IV*. Könighlichen Gesellschaft der Wissenschaften, Göttingen, 1823.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall / CRC, 2nd edition, 2003.

C. J. Geyer. *mcmc: Markov Chain Monte Carlo*, 2010. URL `http://CRAN.R-project.org/package=mcmc`. R package version 0.8.

M. Ghosh and Y.-H. Kim. The Behrens-Fisher problem revisited: A Bayes-Frequentist synthesis. *The Canadian Journal of Statistics*, 29(1):5–17, March 2001.

D. Gillies. *Philosophical Theories of Probability*. Routledge, London, UK, 2000.

C. Glymour. *Theory and evidence*. Princeton University Press, Princeton, New Jersey, 1980.

A. Hájek. Interpretations of probability. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, Stanford, California, 2007. URL `http://plato.stanford.edu/archives/win2007/entries/probability-interpret/`.

S. Hartmann and J. Sprenger. Judgment aggregation and the problem of tracking the truth. *Synthese*, pages 1–13, 2011. URL `http://dx.doi.org/10.1007/s11229-011-0031-5`.

P.G. Hoel, S. C. Port, and C. J. Stone. *Introduction to Probability Theory*. Houghton Mifflin, 1971a.

P.G. Hoel, S. C. Port, and C. J. Stone. *Introduction to Statistical Theory*. Houghton Mifflin, 1971b.

M. Holden, M. R. Dubin, and P. H. Diamond. Frequency of negative intermediate-strength tuberculin sensitivity in patients with active tuberculosis. *New England Journal of Medicine*, 285:1506–1509, 1971.

J. A. Hulse, D. Jackson, D. B. Grant, P. G. H. Byfield, and R. Hoffenberg. Different measurements of thyroid function in hypothyroid infants diagnosed by screening. *Acta Pædiatrica*, 68:21–25, 1979.

E. T. Jaynes. *Probability Theory in Science and Engineering*. Colloquium Lectures in Pure and Applied Science, No. 4. Socony Mobil Oil Company, Dallas, Texas, 1958.

E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK, 2003. G. L. Bretthorst, *Editor*.

H. Jeffreys. *Theory of Probability*. Oxford University Press, London, 3rd edition, 1961. Corrected Impression, 1967.

Joint Committee for Guides in Metrology. *Evaluation of measurement data — Guide to the expression of uncertainty in measurement*. International Bureau of Weights and Measures (BIPM), Sèvres, France, September 2008a. URL `http://www.bipm.org/en/publications/guides/gum.html`. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 100:2008, GUM 1995 with minor corrections.

Joint Committee for Guides in Metrology. *Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008b. URL `http://www.bipm.org/en/publications/guides/gum.html`. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 101:2008.

Joint Committee for Guides in Metrology. *International vocabulary of metrology — Basic and general concepts and associated terms (VIM)*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008c. URL `http://www.bipm.org/en/publications/guides/vim.html`. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 200:2008.

R. Köhler. Photometric and radiometric quantities. In C. DeCusatis, editor, *Handbook of Applied Photometry*, chapter 2, pages 33–54. Springer Verlag, New York, New York, 1998.

D. Lindley. *Understanding Uncertainty*. John Wiley & Sons, Hoboken, New Jersey, 2006.

D. V. Lindley. *Introduction to Probability and Statistics from a Bayesian Viewpoint — Part 1, Probability*. Cambridge University Press, Cambridge, UK, 1965a.

D. V. Lindley. *Introduction to Probability and Statistics from a Bayesian Viewpoint — Part 2, Inference*. Cambridge University Press, Cambridge, UK, 1965b.

D. V. Lindley. Reconciliation of probability distributions. *Operations Research*, 31(5):866–880, September-October 1983.

D. V. Lindley. *Making Decisions*. John Wiley & Sons, London, 2nd edition, 1985.

D. H. Mellor. *Probability: A Philosophical Introduction*. Routledge, New York, 2005.

N. Metropolis and S. Ulam. The Monte Carlo Method. *Journal of the American Statistical Association*, 44:335–341, September 1949.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.

M. G. Morgan and M. Henrion. *Uncertainty — A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, New York, NY, first paperback edition, 1992. 10th printing, 2007.

P. A. Morris. Combining expert judgments: A bayesian approach. *Management Science*, 23(7):679–693, March 1977.

J. Neyman. Frequentist probability and frequentist statistics. *Synthese*, 36(1): 97–131, 1977.

K. R. Popper. The propensity interpretation of probability. *British Journal of the Philosophy of Science*, 10:25–42, 1959.

A. Possolo. Copulas for uncertainty analysis. *Metrologia*, 47:262–271, 2010.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

F. P. Ramsey. Truth and probability. In R.B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, chapter VII, pages 156–198. Harcourt, Brace and Company, New York, 1999 electronic edition, 1926, 1931. URL `http://homepage.newschool.edu/het/texts/ramsey/ramsess.pdf`.

H. Reichenbach. *The Theory of Probability: An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability*. University of California Press, Berkeley, California, 1949. English translation of the 1935 German edition.

C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, New York, NY, second edition, 2007.

L. J. Savage. *The Foundations of Statistics*. Dover Publications, New York, New York, 1972.

M. J. Schervish. *Theory of Statistics*. Springer Series in Statistics. Springer Verlag, New York, NY, 1995.

B. W. Silverman. *Density Estimation*. Chapman and Hall, London, 1986.

R. H. Simpson. The hurricane disaster potential scale. *Weatherwise*, 27:169–186, 1974.

M. Stone. The opinion pool. *The Annals of Mathematical Statistics*, 32:1339–1342, December 1961.

Student. The probable error of a mean. *Biometrika*, 6(1):1–25, March 1908.

B. N. Taylor and C. E. Kuyatt. *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. National Institute of Standards and Technology, Gaithersburg, MD, 1994. URL `http://physics.nist.gov/Pubs/guidelines/TN1297/tn1297s.pdf`. NIST Technical Note 1297.

R. von Mises. *Probability, Statistics and Truth*. Dover Publications, New York, 2nd revised edition, 1981. ISBN 0486242145. Translation of the 3rd German edition.