Marginal Probability of Discrete Variables

From the joint probability of several variables, we can obtain the *marginal probability* for a subset of them by summing over the rest of the variables. For example, given $P(x, y)$, the marginal probability $P(x)$ is obtained summing over the variable we want to "eliminate," namely, Y:

$$P(x) = \sum_{y} P(x, y).$$

In the same way, the marginal probability $P(y)$ is obtained summing over X:

$$P(y) = \sum_{x} P(x, y).$$

We can represent the joint and marginal probabilities of this example as given in Table 1.

These results can be generalized for more than two variables. For instance, from the joint probability $P(x, y, z)$, we can derive six marginal probabilities:

$$P(x,y) = \sum_{z} P(x,y,z)$$
$$P(x,y) = \sum_{z} P(x,y,z)$$
$$P(x,y) = \sum_{z} P(x,y,z)$$
$$P(x) = \sum_{y}\sum_{z} P(x,y,z)$$
$$P(y) = \sum_{x}\sum_{z} P(x,y,z)$$
$$P(z) = \sum_{x}\sum_{y} P(x,y,z)$$

Joint and Marginal Probability Distributions of Continuous Variables

Axiomatically, the definition of the joint probability of several continuous variables is based on a *cumulative distribution function*, $F(x, y, z, K)$, whose properties are analogous to the case of a single variable, and the *probability density function* derives from it. However, in practice, the probability P is never defined from a cumulative function, F, but from a density function, f, assuming that f belongs to one of the families of multivariate continuous probability distributions, such as the multivariate Gaussian, also called multivariate normal. Other families of continuous probability density distributions can be found in statistical textbooks and on the Internet.

Table 1 Joint and marginal probabilities for a hypothetical population

P(x, y)	*male*	*female*	*P(y)*
young	.183	.167	.350
adult	.210	.222	.432
elderly	.093	.125	.218
P(x)	.486	.514	1.000

The derivation of the marginal density function from a joint probability density is analogous to the case of discrete variables, just replacing the sum with an integral. For example, in the case of a joint density $f(x, y)$ defined over two variables X and Y, the marginal density distributions are

$$f(x) = \int_{-\infty}^{+\infty} f(x,y)dy,$$

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

The generalization of these equations for a higher number of variables is obvious.

Francisco J. Díez

See also Bayes's Theorem; Conditional Independence; Conditional Probability; Diagnostic Tests; Odds and Odds Ratio, Risk Ratio; Probability, Verbal Expressions of; Subjective Probability; Violations of Probability Theory

Further Readings

de Finetti, B. (1974, 1975). *Theory of probability* (2 vols.). New York: Wiley.

Jeffrey, R. C. (1992). *Probability and the art of judgment*. New York: Cambridge University Press.

Kolmogorov, A. N. (1950). *Foundations of the theory of probability*. New York: Chelsea.

Probability, Verbal Expressions of

Verbal expressions of probability are those used for communicating degrees of uncertainty, such as *likely, very likely, possible*, and *uncertain*, in contrast with numerical expressions, such as "The probability of *X* is 70%" or "The odds are 3 to 1."

Correspondence Between Verbal and Numerical Expressions

Since the 1960s, several authors have studied the equivalence between verbal and numerical expressions of probability. The typical experiment consists in selecting a set of linguistic expressions, asking a group of subjects to translate each one into a percentage or a number between 0 and 1, and building a table or a graph that summarizes the results. Most of such studies include expressions that do not represent probabilities in a strict sense.

Variability of Numerical Assignments

Several experiments have indicated low within-subject variability, that is, estimates given by the same subject on different occasions are very similar, while all the studies have shown a high degree of between-subject variability. A related empirical finding is that in general, people underestimate how much individuals vary in their interpretation of probability terms.

The studies have also shown a reasonable degree of between-experiment consistency. In some cases, the numerical values obtained in a study differed from those obtained in others, but the ranking of expressions was essentially the same.

Several experiments that compared the numerical values assigned by different groups of subjects have shown that between-subject variability is smaller among people with similar backgrounds. For instance, a study by Nakao and Axelrod showed that consensus was significantly higher among physicians than among laymen for around half of the expressions of frequency examined and among native-English-speaking physicians than among those with other native languages, but it was not higher among board-certified physicians than among the others.

The Role of Modifiers

The meaning of a verbal expression of probability can be modified by the use of adverbs (*very likely*), affixes (*im-probable*, *un-likely*), or lexical negations (*not likely*). Empirical studies have led to the following ranking of adverbs: *very* > *quite* > no modifier > *rather* > *fairly* > *somewhat*, which means that *very* is the adverb that shifts most the meaning of a probability expression toward extreme values. *Very likely* denotes a higher probability than *quite likely*, which in turn denotes a higher probability than *likely* (no modifier). On the contrary, *rather, fairly*, and *somewhat* shift the meaning of the expression toward .5.

The Influence of Context

While most of the experiments asked the subjects to translate isolated linguistic expressions, other researchers have studied those expressions in context. A surprising finding is that in most of the cases, between-subject variability is higher when probabilistic expressions are given in context. Another finding, not surprisingly, is the *base-rate effect*, which means that, in general, the higher the prior probability of an event, the higher the numerical values assigned. For instance, the term *likely* in "It is likely that it will snow in December" is assigned higher values than in "It is likely that it will snow in October."

Qualitative expressions associated with more severe outcomes (e.g., "likely death" vs. "likely injury") tend to be assigned lower numbers. This might also be explained by the base-rate effect, because in general more severe outcomes have lower base rates. There is also empirical evidence that expressions associated with positive outcomes tend to be assigned higher numbers than those associated with negative outcomes.

More interestingly, Mazur and Merz proved that personal characteristics, such as age, healthcare experience, and perceived health status, influence patients' interpretations of verbal probability terms.

Preferences for Numerical or Verbal Probabilities

Reasons for Preferring Verbal Probabilities

One of the reasons for using verbal probabilities is that they are more natural than numbers: Spontaneously, people express probabilities with linguistic terms, whereas it requires an additional cognitive effort to give numeric estimates.

Additionally, verbal probabilities can reflect the speaker's lack of knowledge: Very often expressions of probability do not stem from systematic data, but they are estimates made by human beings based on the cases stored in their memory, on what they have read or heard, and so on. In this case, people do not dare convey a numerical probability, for two reasons. The first is that an empirical study might later prove the assertion to be wrong. For instance, if an expert says in a book that the prevalence of a disease is 1%—or even "around 1%," which is an imprecise probability—and a posterior study shows that it is .4% or 3%, the expert's reputation will be compromised. On the contrary, if he said that "the prevalence is relatively high," he does not commit to a particular figure, and so his assertion cannot be refuted. The second reason is that subjective estimates expressed as numerical probabilities may mislead the listener to believe that the speaker knows the true probability with precision.

In addition to cases in which the probability has an objective value but is unknown, there are other cases in which it does not make sense to assume that there exists a measurable probability. For instance, a doctor may feel unable to answer with a precise value a question about the probability of a patient's survival, because a question such as this, referring to a *single-event probability*, does not have an objective meaning. In this case, it is much easier to respond with a linguistic probability.

Another reason for using verbal expressions is that in addition to conveying a probability estimate, they can also express *directionality*. A phrase having positive direction, such as "*X* is possible," implicitly points at the reasons for the occurrence of *X*, while a phrase of negative direction, such as "*X* is uncertain" or "*X* is doubtful," implicitly underlines the causes that may prevent *X*. Therefore, verbal expressions may be preferred when the speaker, in addition to conveying a vague probability, wishes to make the listener pay more attention to the reasons in favor of or against the occurrence of an event.

Empirical Evidence

Many experiments have been carried out to study human preferences about probability expressions. The most consistent finding is that while more people prefer to receive information about probabilities numerically, they prefer to express such information verbally. This is called the *preference paradox*.

In addition to the direction of communication (giving vs. receiving information), other factors have been shown to influence human preferences. One of them is the *nature of the event*: When expressing the probability of repeated events with aleatory uncertainty, most individuals prefer to use numerical estimates, which allow them to distinguish between levels of uncertainty with higher precision, but the same individuals tend to use more imprecise methods when communicating single-event probabilities. Another factor is the *strength of the available evidence*: People tend to use more precise expressions of probability when the information is firmer and more reliable.

In some of the studies, the people giving information were doctors and those receiving it were patients. Other studies have set a scenario in which subjects were randomly assigned to the group of advisors or to the group of decision makers, whose choice is based on the information received from the advisors.

A different problem, related to the construction of decision support systems, is the elicitation of the parameters of a probabilistic model, such as a Bayesian network or an influence diagram. An empirical study carried out by Witteman and colleagues, in which general practitioners had to assess several conditional probabilities, concluded that the less experienced doctors preferred a purely verbal scale, the most experienced preferred a purely numerical scale, while the groups in between preferred a combined verbal-numerical scale.

Impact on Medical Decision Making

The use of verbal expressions of probability poses a serious problem as a potential source of errors, particularly in the case of informed consent. The first problem is the risk of misunderstanding. Let us imagine a patient suffering from a disease that will cause his death. His doctor offers him a treatment that may save his life but may have side effects. The decision of accepting the treatment depends on the probability of survival and on the probability and severity of side effects. In this context, verbal expressions of probability entail an obvious danger of misunderstanding: The doctor's estimate that there is a 60% probability of survival, conveyed as "It is likely that you will get cured," might be interpreted by the patient as a

90% probability, and the assertion that "sometimes the treatment causes severe adverse effects" may be interpreted as having a probability lower than 2% or higher than 15%. The danger is even higher in the case of extreme probabilities, because an expression such as *very unlikely* may mean .1 as well as .00001 probability. However, there is ample evidence that people, including investigators, underestimate the variability of subjective estimations.

The second issue is how patients process the information that they receive from their doctors. On this point, the empirical evidence is contradictory. Some studies seem to demonstrate that subjects are more effective at reasoning with verbal expressions than with numerical expressions, even if the tasks performed rely on frequency information. However, other studies arrived at the opposite conclusion, and others have found no significant difference.

Disadvantages and Advantages

Verbal expressions of probability are often used in medical communications—in fact, much more often than numerical expressions. Experts in the field defend contradictory opinions about their usefulness and their peril.

The main drawback of verbal probabilities is the risk of misunderstanding, because the value interpreted by the listener can be very different from that intended by the speaker. Some researchers have proposed using a very limited number of linguistic probabilities, whose meaning should be explicitly determined beforehand. However, this proposal does not solve two of the main problems: that the interpretation of such expressions varies with the context (the aforementioned base-rate effect) and that verbal probabilities are not able to discriminate extreme values, such as .1 and .0001. On the contrary, this lack of precision of linguistic expressions turns into an advantage in the usual case of imprecise knowledge: In that case it may be very difficult for the speaker to utter a numerical probability, and, even worse, that precise probability may mislead the recipient of the information.

On the other hand, a disadvantage of verbal expressions of uncertainty is the lack of a normative calculus, in contrast with the well-defined principles and techniques of probability theory, which play an essential role in medical diagnosis and decision making. Additionally, some experiments have shown that numerical probabilities lead to better judgments and to better decisions. However, other studies have arrived at the opposite conclusion, or at least a tie.

The directionality of verbal expressions, which carries additional information, has been put forward as one of their advantages. In contrast, numerical expressions, because of their neutrality, should be chosen when the speaker does not wish to bias the listener.

As an attempt to combine the advantages of both, some experts advocate using them together, by appending to each linguistic expression its intended meaning—for instance, "It is very likely (80%–90%) that. . . ."

The movement of evidence-based medicine and the use of computerized decision support systems will give an increasingly prominent role to numerical probabilities, to the detriment of verbal expressions, but because of the above arguments and the strong human preferences, it is clear that the use of linguistic probabilities will never disappear from medical communications, either oral or written.

Francisco J. Díez and Marek J. Druzdzel

See also Human Cognitive Systems; Probability; Risk Communication; Subjective Probability

Further Readings

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes, 41,* 307–314.

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. In J. R. Busemeyer, R. Hastie, & D. Medin (Eds.), *The psychology of learning and motivation: Decision making from the perspective of cognitive psychology* (pp. 275–318). New York: Academic Press.

Clark, D. A. (1990). Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology, 9,* 203–235.

Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Medical Decision Making, 27,* 696–713.

Mazur, D. J., & Merz, J. F. (1994). Patients' interpretations of verbal expressions of probability: Implications for securing informed consent to medical interventions. *Behavioral Sciences and the Law, 12,* 417–426.

Nakao, M. A., & Axelrod, S. (1983). Numbers are better than words: Verbal specifications of frequency have no place in medicine. *American Journal of Medicine, 74,* 1061–1065.

Teigen, K. H., & Brun, W. (1999). The directionality of verbal probability expressions: Effects on decisions, predictions, and probabilistic reasoning. *Organizational Behavior and Human Decision Processes, 80,* 155–190.

Teigen, K. H., & Brun, W. (2003). Verbal expressions of uncertainty and probability. In D. Hardman & L. Macchi (Eds.), *Thinking: psychological perspectives on reasoning, judgement and decision making* (chap. 7, pp. 125–145). Chichester, UK: Wiley.

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *Knowledge Engineering Review, 10,* 43–62.

Witteman, C. L. M., Renooij, S., & Koele, P. (2007). Medicine in words and numbers: A cross-sectional survey comparing probability assessment scales. *BMC Medical Informatics and Decision Making, 7,* 13.

Probability Errors

Physicians and patients deciding on treatment options often need to estimate the probability of various outcomes (e.g., death, partial recovery, full recovery) in order to make effective decisions. However, human probability estimation is often fraught with errors that may interfere with making correct decisions. There are a number of probability errors that physicians ought to be aware of.

Human probability judgments often exhibit a rather basic bias: We overestimate low probabilities and underestimate high probabilities. For instance, people overestimate the likelihood of dying from rare diseases such as smallpox or botulism while underestimating the likelihood of dying from more common afflictions such as strokes or heart disease. One explanation for these biases might be that our probability estimates tend to regress toward the mean. That is, estimates of low probabilities have nowhere to go but higher while estimates of high probabilities can more easily go lower.

Another reason for these biases has to do with the *availability heuristic*, which states that we base our estimates of an event's likelihood on how easy it is to imagine that event occurring. In other words, we overweight information that comes easily to mind. Because rare events get a large amount of graphic news coverage (e.g., plane crashes, the contracting of rare diseases), these events are easier to imagine than mundane and less publicized events (e.g., car accidents, asthma attacks). Given this, it is perhaps not surprising that surgeons from high-mortality specialties give higher estimates of dying in the hospital than do surgeons from low-mortality specialties.

Relatedly, the way in which we think about information can also affect the ease with which we envision certain events and our subsequent estimates of these events' probabilities. If composite events are "unpacked" into their components, then these components become easier to envision, and people estimate their probabilities to be higher. For instance, the risk of dying from "natural causes" can be unpacked into heart attacks, cancers, and other natural causes. People who were asked about the risk of dying from natural causes responded that it was 58%. However, people who were asked about the probability of the unpacked events responded that the risk of dying from heart attacks was 22%; from cancer, 18%; and from other natural causes, 33%. This sums to 73%, considerably higher than the 58% estimate. Probability estimates increase when individual components are considered. When this occurs, such estimates are considered to be *subadditive*, as the judgment for the composite event is less than the sum of the judgments for its parts.

Decision makers also overweight surface similarities between events when judging probabilities, a strategy known as the *representativeness heuristic*. For instance, given a description of a woman who is bright, outspoken, and concerned with social justice, Amos Tversky and Daniel Kahneman found that people are more likely to state that this woman is a feminist bank teller rather than just a bank teller. However, as all feminist bank tellers are also bank tellers, one cannot be more likely to be a feminist bank teller than a bank teller. This bias occurs because the description of the woman