Bayesian Statistics

JENS VOGELGESANG
University of Hohenheim, Germany
MICHAEL SCHARKOW
Zeppelin University, Germany

Introduction

Bayesian statistics allows prior knowledge to be incorporated into a statistical model. Central to this approach is Bayes' theorem, which provides a mathematical rule to update the prior knowledge with new data. The results of a Bayesian analysis are characterized by a posterior distribution, which balances the prior knowledge—specified by using a particular form of probability distribution—and the data likelihood. Bayesian statistics is conceived as a counterpart to the classical approach to statistics. The classical approach, however, is the dominant statistical paradigm in communication research.

Historically, the classical approach to inferential statistics can be traced back to R. A. Fisher (1890–1962), Jerzy Neymann (1894–1981), and Egon Pearson (1895–1980). The current statistical practice in the field of communication is a hybrid approach binding together frequentist methodology developed by Neymann and Pearson with likelihood-based methodology developed by Fisher (Gigerenzer, Porter, Daston, Beatty, & Krüger, 1989). However, introductory textbooks on statistics usually present the classical approach as if there were never any discussion about its coherence and applicability. The frequentist approach views probability as a ratio of frequencies. Most students in communication are introduced to the classical concept of probability by studying the infinite sampling properties of a coin toss or the roll of a die. This particular notion of probability can be described using axioms developed by Andrei Nikolaevich Kolmogorov (1903–1987): (a) the probability of any event is equal to or greater than zero; (b) the probability of a certain event is 1; (c) if A and B are two mutually exclusive events (events that cannot both occur), then the probability of the disjunction (the probability of either A or B occurring) is equal to the sum of their individual probabilities: $P(A \textit{ or } B) = P(A) + P(B)$. If two events A and B are independent, then the occurrence of one event does not influence the probability of another event, which can be defined as $P(A \textit{ and } B) = P(A) \times P(B)$. There are situations in which probabilities can be estimated from data (e.g., the probability of someone being a newspaper reader in a population), but sometimes they cannot (e.g., the probability, in advance of data, of the null hypothesis being correct). When probabilities cannot be thought of as a ratio of frequencies, they still need to be estimated somehow. In such situations the Bayesian approach to statistics comes into play.

The International Encyclopedia of Communication Research Methods. Jörg Matthes (General Editor), Christine S. Davis and Robert F. Potter (Associate Editors).

DOI: 10.1002/9781118901731.iecrm0013

Interestingly, the Bayesian approach to statistics predates the classical approach by more than 150 years. In the 18th century, Thomas Bayes (1702–1761) described the mathematical rule that we now call the Bayes theorem in a paper titled "An essay towards solving a problem in the doctrine of chances". This paper was found after his death and posthumously published in the *Philosophical Transactions of the Royal Society*. Pierre-Simon Laplace (1749–1827) later generalized the work by Bayes. This is why the Bayes theorem is sometimes also called the Bayes–Laplace rule. The work by Bayes and Laplace, among others, laid the foundation of modern statistical theory. Until the second half of the 18th century, probability calculation was almost entirely focused on estimating the likelihood of future uncertain events. It was the insight of Bayes that the calculus of probability could be used to assess not just the likelihood of future events, but also the likelihood of past events. Essentially, with Bayes' theorem the concept of conditional probability was introduced. This probability is denoted using a vertical bar; for example, $p(A|B)$, reads as "the probability of A given B." Conditional probabilities refer to the case of nonindependent events, which can be described using the axioms put forth by Alfréd Rényi (1921–1970), who extended the system of Kolmogorov's axioms (Kaplan, 2014; Press, 2003). The conditional probability $p(A|B)$ is determined by the joint distribution of A and B.

Bayesian statistics and probability

Bayesian and classical (hybrid approach) statistics are based on the assumption that statistical methods allow valid inference statements when there is random variability in the data. In both approaches, sample data are used to make inference statements about unknown population parameters. However, Bayesian statistics challenges many of the assumptions underlying classical statistics. One key difference between both approaches concerns the nature of the unknown parameter. According to the frequentist approach, an unknown population parameter is a fixed, nonrandom quantity. It is assumed that there is one true population parameter. As a consequence, no probability statements can be made about its value. In the Bayesian view, in contrast, the true value of a population parameter is conceived as uncertain and is therefore considered a random variable. According to the Bayesian approach, the unknown, random population parameter should be described by a probability distribution. Unlike the frequentist approach, the Bayesian counterpart allows a probability statement to be made about the value of an unknown parameter. Both approaches also differ in their notion of probability. Frequentist procedures are based on a concept of probability that is associated with the idea of long-run frequency (e.g., a coin toss). Frequentist inference, which employs sampling distributions based on infinite repeated sampling, is focused on the performance over all possible random samples. Therefore, a frequentist probability statement does not relate to a particular random sample that was obtained. Rather, the sampling distribution, which describes the probability distribution of the sample statistic over all possible random samples from the population, is used to make a confidence statement about the unknown population parameter. The name "confidence statement" is chosen because the inference probability is based on all possible datasets that could have occurred for the fixed but unknown population parameter.

The Bayesian approach, in contrast, has a different interpretation of probability. According to this view, a probability statement about an unknown parameter mirrors a subjective degree of belief or experience of uncertainty. This uncertainty is captured by a probability distribution that is defined before observing the data. In the Bayesian terminology, this particular distribution is called the prior distribution, or simply prior. The idea of a prior is best described as being analogous to placing a bet. The bet comprises the amount of certainty that a bettor has about a random outcome before knowing the outcome's realization. The Bayesian approach provides a mathematical rule called Bayes' theorem describing how to change existing prior beliefs about the value of an unknown random parameter in the light of new evidence, such as empirical (sample) data. The data can be expressed in terms of a likelihood function, sometimes simply called the likelihood. Using Bayes' theorem as a formal rule to weigh the likelihood of the actual occurred data with the beliefs held before observing the data gives the posterior distribution. The posterior distribution allows researchers to make probability statements concerning the unknown parameter of interest.

Bayes' theorem

Bayes' theorem contains three essential elements. It balances a prior state of knowledge and the data likelihood to a more informed posterior distribution, that is:

$$\text{posterior distribution} \propto \text{prior distribution} \times \text{data likelihood}, \tag{1}$$

where the symbol $\propto$ means "is proportional to." More specifically, Bayesian inference always begins with some prior probability statement about an unknown parameter, that is $f(theta)$, for example. Recall that, in contrast to the frequentist approach, all unknown parameters of a Bayesian model are treated as random variables that can be described in terms of a distribution. Unlike the frequentist approach, the Bayesian approach allows incorporation of prior knowledge before observing the data. The prior probability statement $f(theta)$ is nothing but a summarized expression of the current state of knowledge and the subjective degree of belief about *theta* before gathering or seeing any new data. The functional form of the prior is usually chosen to facilitate the calculation of the posterior. For example, the mean of a normally distributed prior would represent an informed guess about the location of the unknown population mean, whereas the variance would reflect the amount of uncertainty about that particular parameter. The smaller the prior variance, the more certain one is that the prior mean mirrors the population mean. In Bayesian terminology, the prior variance is called precision. Parameters such as the prior mean or the precision are also referred to as hyperparameters. When a probability distribution is specified on the hyperparameters in a fully hierarchical Bayesian model, they are referred to as hyperpriors (Kaplan, 2014). As can be seen from Equation 1, the prior distribution is weighted with the data likelihood (which is not a probability distribution). Let, for example, $data = (x_1, \ldots, x_i)$ be a sample from a density f_{theta} with an unknown parameter *theta* and with an associated likelihood function:

$$l(theta|data) = \prod_{i=1}^{n} f_{theta}(x_i). \tag{2}$$

The likelihood function summarizes the sample information about *theta* and provides some value of *theta* that makes the data most likely to have occurred. The information from the likelihood function is weighted with the prior probability distribution by employing Bayes' theorem to calculate an updated distribution posterior to the former state of knowledge:

$$f(theta|data) = \frac{f(data|theta) \times f(theta)}{f(data)}, \tag{3}$$

where $f(theta|data)$ denotes the posterior distribution for the parameter *theta,* $f(data|theta)$ is a sampling density for the data, $f(theta)$ is the prior distribution for the parameter, and $f(data)$ is the marginal probability of the data. Whereas classical inference about *theta* follows from inspection of the likelihood, Bayesian inference, in contrast, relies on inspecting the posterior distribution using descriptive measures. The shape of the posterior distribution can be described by calculating the location parameters, such as the posterior mean, which is the expected value of *theta* under $f(theta|data)$, or the posterior mode, which is the most likely value under $f(theta|data)$ or, in addition, by some variability measures.

Bayes' theorem is fundamental, for example, to the Naïve Bayes (NB) algorithm that is commonly used in automatic text classification. The attribute naïve comes from the assumption that the features (such as words) in a text are mutually independent, and therefore probabilities for all features can simply be multiplied to yield a combined probability. The NB algorithm is used to assign documents to prespecified categories. Specifically, the probability that a document belongs in a category given its features is computed using (a) the prior probability of the category and (b) the frequency of the occurring features in documents previously assigned to the category. The basic idea of this algorithm is to maximize the posterior probability for a category given some training data to formulate a classification or—more general—a decision rule.

Bayesian interpretation of the *p*-values and credible intervals

The different notions of probability of Bayesian and classical statistics have far-reaching ramifications for how conclusions are drawn. Scientists formulate hypotheses based on theoretical reasoning and the question is whether or not the hypothesis is correct. Unlike the classical approach, the Bayesian approach provides an answer to this question. Bayesian statistics is concerned with the probability of a hypothesis given the data; that is, $p(hypothesis|data)$. By contrast, classical statistics is interested in the probability of obtaining data as extreme, or more extreme, than those observed if the hypothesis is correct, that is $p(data|hypothesis)$. Correspondingly, different notions of probability also affect the interpretation of Bayesian credible intervals and classical confidence intervals. Both types of intervals provide a measure of uncertainty with respect to an estimated parameter. Assuming that a posterior density is approximately normal, derivation of a 95 percent confidence interval, for example, is straightforward:

$$\text{posterior mean } \pm 1.96 \times \text{ posterior standard deviation.} \tag{4}$$

The Bayesian 95 percent credible interval is expected to contain the unknown population parameter with a probability of 95 percent. Unlike the Bayesian credible interval, which refers to the parameter space, the classical confidence interval refers to the sample space. A probability statement associated with a 95 percent confidence interval can only be made with reference to the procedure, not the unknown population parameter itself. Since the population parameter is a fixed, nonrandom quantity, a frequentist 95 percent confidence interval means nothing other than that the procedure of interval construction is expected to construct intervals that include the population parameter approximately 95 percent of the time.

Prior distributions

Scientific progress rests on the idea of learning; gathering scientific knowledge is a cumulative process. Most studies, if not all, are conducted in the light of previous research. When planning a study, it is rational to collate as much theoretical thought as possible, and to become familiar with methodological and statistical standards, and existing empirical findings. Bayesian statistics requires researchers to take into account all existing knowledge when choosing a prior distribution. The notion that statistical inference is a way of updating existing knowledge given new data, and that, consequently, data do not speak for themselves, might be irritating at first sight. The incorporation of preexisting information adds a seemingly more subjective flavor to the scientific method than pure likelihood-based inference. Consequently, the subjectivity of the chosen priors is the most prominent objection to the Bayesian approach. Generally speaking, the choice of a prior is based on how much information one believes oneself to have prior to the data collection and how certain one is about this subjective belief. There is considerable dissent on the issue of specifying priors. There are two general types of priors: uninformative and informative prior distributions. This distinction is also referred to as objective and subjective priors (Press, 2003). The former are chosen in such a way that the data—that is, the likelihood according to Bayes' theorem—speak for themselves. Although there is consensus that no statistical or other scientific method can be truly objective, many scholars argue that using uninformative priors is more justifiable and transparent to colleagues, students, or reviewers. For those, the notion of objectivity might be more important than the statistical efficiency gained by using informative priors. The public policy prior denotes a special case of using uninformative prior distributions and concerns reporting results as general as possible; that is, minimizing the impact of the researcher's subjective belief on the posterior (Press, 2003). Noninformative priors are also referred to as vague or diffuse priors. Statistically, using noninformative priors yields about the same results as classical inference but still allows Bayesian interpretation. The simplest way of expressing prior ignorance is often seen in using a uniform distribution for a parameter of interest. A uniform distribution deems every value equally probable, possibly within some bounds like [−1, 1] for a correlation. Because this kind of prior is a constant, the posterior distribution is computed only from the likelihood, yielding a pseudofrequentist result. As easy as

the interpretation and justification of this prior may seem, uniform priors are not equally well suited for all kinds of parameters, mainly because they are not robust to simple transformations (Gill, 2008). For example, a uniform prior for a variance is different than that for a standard deviation, and both yield different posteriors. The construction of uninformative priors that are both robust to transformations and lead to proper posteriors has challenged many Bayesian statisticians. If one accepts the idea of informative priors, the question of obtaining them arises. The answer is both simple and complicated: Any knowledge or belief, be it from an expert, a preexisting study, a meta-analysis, theoretical reasoning, or just an educated guess, can be used as long as it can be transformed into a probability distribution of some kind. It is mathematically convenient to choose conjugate informative priors. A prior distribution is called conjugate if the posterior is in the same distributional family as the prior. If the prior is not conjugate, the resulting posterior distribution cannot be solved analytically. Alternatively, numerical simulation methods such as Markov chain Monte Carlo (MCMC) estimation need to be used to find approximate solutions for the posterior (Gilks, Richardson, & Spiegelhalter, 1996). Together with the advent of fast computer hardware in the 1990s, MCMC algorithms and freely available software programs helped to popularize Bayesian data analysis methods in the social sciences. Generally speaking, there are three types of informative priors: empirical priors, weakly informative priors, and subjective priors. Empirical priors refer to previous observation or other forms of data collection, including expert interviews. The latter are often referred to as elicited priors. Eliciting a prior is very demanding because transforming qualitative or vague expert opinions into prior distributions for parameters such as regression coefficients or between-group variances is a challenging task for any applied researcher. Another kind of empirical prior comes from the incorporation of previous single results or meta-analyses. Those results can either be directly incorporated into strict replication studies or somewhat discounted, depending on the similarity of the previous and the current study. While this can be intuitively accomplished by using equivalent sample sizes to relate prior and likelihood, a more versatile and rigorous approach to this problem involves power priors. In the presence of historical data or data from previous similar studies with large sample size, a power prior can be realized by raising the likelihood function based on the prior data to a suitable power δ ($0 \leq \delta \leq 1$) to downweight the historical data relative to the current data. Subjective priors express personal theories or beliefs about a phenomenon, and as such they are highly debatable. However, it may often be desirable to check the consequences of different priors on the posterior distribution or compare models with optimistic or pessimistic priors (Press, 2003). While both empirical and subjective priors are domain specific, a third kind of informative prior distribution is based on statistical convenience and common sense about model parameters. The rationale behind weakly informative priors is that extreme values of parameters such as correlations or regression coefficients are highly unlikely and should therefore be given less prior probability (Gelman, Carlin, Stern, & Rubin, 2013). This approach to using default priors leads to efficient and stable estimation without overly affecting the likelihood.

Bayesian data analysis

Applications using Bayesian data analysis have recently grown in the social sciences. One reason for this growth, among others, is that Bayesian methods were successively implemented in software packages that are frequently used by social scientists such as R, Mplus, and SPSS Amos. These software packages, which were originally designed with respect to the frequentist approach, now also allow estimation of the posterior distribution using MCMC methods.

Predefined routines of software packages nowadays support researchers in evaluating the posterior distribution, the results of a Bayesian data analysis. This can be done, for example, by using point estimates of the posterior, such as the posterior mean or variance. Another summary statistic that is commonly used is the mode of the posterior distribution. This mode is also referred to as the maximum a posteriori (MAP) estimate. The MAP is the Bayesian analogy to the classical maximum likelihood estimator (MLE). In addition to point estimates, intervals summaries can be obtained to characterize the posterior distribution (Kaplan, 2014). As pointed out above, credible intervals—also sometimes called posterior probability intervals—of point estimates can be directly obtained from the quantiles of the posterior distribution. Another alternative is the so-called highest posterior density (HPD) interval, which has the property that the density within the interval region is never lower than the density outside. It is recommended to use the HPD in favor of standard credible intervals when the posterior density is asymmetric or not unimodal (Box & Tiao, 1973).

As in classical statistics when using the likelihood ratio test, Bayesian data analysis allows the testing of model fit, the comparison of competing models or hypotheses. For example, the Bayes factor (BF) is a weighted average likelihood ratio. It is often interpreted as the relative evidence in the data, indicating the odds that the data favor one model or hypothesis over another. Konijn, van de Schoot, Winter, and Ferguson (2015), who illustrated the use of BF by reanalyzing data of published communication studies, argue that the BF offers more meaningful results than frequentist null hypothesis significance testing (NHST). BF null hypothesis testing is supposed to have advantages over classical NHST, because it allows a more intuitive interpretation, considers likelihood under both the null and the alternative hypothesis, and can also provide evidence for and not just against the null hypothesis. It should be noted, however, that the BF is sensitive with regard to the choice of priors used for parameters in each model. Therefore, any BF should be used with caution. To overcome the sensitivity of the BF to the prior, Gelman et al. (2013) recommend weakly informative default priors like the Cauchy distribution when estimating parameters of common statistical models like analysis of variance (ANOVA) or regression analysis. Kass and Raftery (1995) suggest that if BF is 1 to 3, the evidence is not worth more than a bare mention; if BF is 3 to 20, it is positive; if BF is 20 to 150, it is strong; and if BF is more than 150, it is very strong. When the exact BF is impossible to calculate (e.g., due to computational limitations or when it is difficult to specify reasonable priors), the Bayesian information criterion (BIC) is said to provide a reasonable approximation to the BF. Note that the BIC is a widely used measure in model selection when using classical statistics such as time series analysis or structural equation modeling (SEM). Unlike the BF, there are no rules of thumb with

respect to the size of the BIC difference between two models. A common guideline is to favor the model with the smallest BIC. The deviance information criterion (DIC) is a Bayesian alternative to the BIC. Based on MCMC estimation, the DIC uses the posterior density, implying that the prior information is taken into account. Among a candidate set of models, the one with the lowest DIC value is chosen.

It is important to recognize that there is no unique Bayesian solution to a statistical problem. However, the Bayesian approach provides a versatile, flexible toolkit that might help to overcome the limitations of classical statistical approaches. Generally speaking, the Bayesian approach is particularly well suited to model complex data structures. For example, Bayesian multilevel modeling allowed estimating a proportional hazards model to investigate the number of seconds that commercials are viewed before being zapped, while accounting for unobserved heterogeneity across both consumers and commercials (Gustafson & Siddarth, 2007). Likewise, marketing researchers used a Bayesian hidden Markov model to identify visual attention states to magazine advertisements from individual eye-tracking data (Liechty, Pieters, & Wedel, 2003). Moreover, Bayesian parameter estimates have favorable efficiency and bias properties relative to MLE in small samples. For example, when few countries are available in comparative research designs, using a Bayesian approach yields far more stable and precise estimation results than frequentist techniques (Stegmueller, 2013).

SEE ALSO: Amos (Software); Comparative Research Methods; Computational Simulation Methods; Mplus; Probability Distributions; R (Software); Statistical Significance (Testing)

References

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). London: CRC Press.

Gigerenzer, G. S., Porter, Z., Daston, T., Beatty, L., & Krüger, J. L. (1989). *The empire of chance: How probability changed science and everyday life*. New York: Cambridge University Press.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.) (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Gill, J. (2008). *Bayesian methods: A social and behavioral sciences approach* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gustafson, P., & Siddarth, S. (2007). Describing the dynamics of attention to TV commercials: A hierarchical Bayes analysis of the time to zap an ad. *Journal of Applied Statistics*, *34*(5), 585–609. doi:10.1080/02664760701235279

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. New York: Guilford Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi:10.2307/2291091

Konijn, E. A., van de Schoot, R., Winter, S. D., & Ferguson, C. J. (2015). Possible solution to publication bias through Bayesian statistics, including proper null hypothesis testing. *Communication Methods and Measures*, *9*(4), 280–302. doi:10.1080/19312458.2015.1096332

Liechty, J., Pieters, R., & Wedel, M. (2003). Global and local covert visual attention: Evidence from a Bayesian hidden Markov model. *Psychometrika*, *68*(4), 519–541. doi:10.1007/bf02295608

Press, S. J. (2003). *Subjective and objective Bayesian statistics: Principles, models, and applications* (2nd ed.). New York: John Wiley & Sons.

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, *57*(3), 748–761. doi:10.1111/ajps.12001

Further reading

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Amsterdam: Academic Press.

McGrayne, S. B. (2011). *The theory that would not die: How Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*. New Haven, CT: Yale University Press.

Jens Vogelgesang (PhD, Free University of Berlin) is a professor of communication at the University of Hohenheim, Germany. His main interests concern audience research, media effects, and methodology.

Michael Scharkow (PhD, University of the Arts Berlin) is professor of communication at Zeppelin University, Germany. His research interests include empirical research methods, online communication, and media use.