A Brief Introduction to Bayesian Statistics

David Kaplan

Department of Educational Psychology

GESIS, Mannheim, Germany, 8 November, 2017
The Reverend Thomas Bayes, 1701–1761
Pierre-Simon Laplace, 1749–1827
It is clear that it is not possible to think about learning from experience and acting on it without coming to terms with Bayes’ theorem. - Jerome Cornfield
Bayesian statistics has long been overlooked in the quantitative methods training of social scientists.

Typically, the only introduction that a student might have to Bayesian ideas is a brief overview of Bayes’ theorem while studying probability in an introductory statistics class.

Until recently, it was not feasible to conduct statistical modeling from a Bayesian perspective owing to its complexity and lack of availability.

Bayesian statistics represents a powerful alternative to frequentist (classical) statistics, and is therefore, controversial.

Bayesian statistical methods are now more popular owing to the development of powerful statistical software tools that make the estimation of complex models feasible from a Bayesian perspective.
Outline

- Major differences between the Bayesian and frequentist paradigms of statistics
- Bayes’ Theorem
- Priors
- Computation
- Bayesian hypothesis testing
- Bayesian model building and evaluation
- Debates within the Bayesian school.
Paradigm Difference I: Conceptions of Probability

- For frequentists, the basic idea is that probability is represented by the model of **long run frequency**.

- Frequentist probability underlies the Fisher and Neyman-Pearson schools of statistics – the conventional methods of statistics we most often use.

- The frequentist formulation rests on the idea of equally probable and stochastically independent events.

- The physical representation is the coin toss, which relates to the idea of a very large (actually infinite) number of repeated experiments.
The entire structure of Neyman - Pearson hypothesis testing and Fisherian statistics (together referred to as the frequentist school) is based on frequentist probability.

Our conclusions regarding null and alternative hypotheses presuppose the idea that we could conduct the same experiment an infinite number of times.

Our interpretation of confidence intervals also assumes a fixed parameter and CIs that vary over an infinitely large number of identical experiments.
But there is another view of probability as **subjective belief**.

The physical model in this case is that of the “bet”.

Consider the situation of betting on who will win the World Cup (or the World Series).

Here, probability is not based on an infinite number of repeatable and stochastically independent events, but rather on how much knowledge you have and how much you are willing to bet.

Subjective probability allows one to address questions such as “what is the probability that my team will win the World Cup?” Relative frequency supplies information, but it is not the same as probability and can be quite different.

This notion of subjective probability underlies Bayesian statistics.
Consider the joint probability of two events, $Y$ and $X$, for example observing lung cancer and smoking jointly.

The joint probability can be written as

$$ p(cancer, smoking) = p(cancer|smoking)p(smoking) \quad (1) $$

Similarly,

$$ p(smoking, cancer) = p(smoking|cancer)p(cancer) \quad (2) $$
Because these are symmetric, we can set them equal to each other to obtain the following:

\[ p(cancer|smoking)p(smoking) = p(smoking|cancer)p(cancer) \]  \hspace{1cm} (3)

\[ p(cancer|smoking) = \frac{p(smoking|cancer)p(cancer)}{p(smoking)} \] \hspace{1cm} (4)

The inverse probability theorem (Bayes’ theorem) states:

\[ p(smoking|cancer) = \frac{p(cancer|smoking)p(smoking)}{p(cancer)} \] \hspace{1cm} (5)
What is the role of Bayes’ theorem for statistical inference?

Denote by $Y$ a random variable that takes on a realized value $y$.

For example, a person’s socio-economic status could be considered a random variable taking on a very large set of possible values.

Once the person identifies his/her socioeconomic status, the random variable $Y$ is now realized as $y$.

Because $Y$ is unobserved and random, we need to specify a probability model to explain how we obtained the actual data values $y$. 
Next, denote by $\theta$ a parameter that we believe characterizes the probability model of interest.

The parameter $\theta$ can be a scalar, such as the mean or the variance of a distribution, or it can be vector-valued, such as a set of regression coefficients in regression analysis or factor loadings in factor analysis.

We are concerned with determining the probability of observing $y$ given the unknown parameters $\theta$, which we write as $p(y|\theta)$.

In statistical inference, the goal is to obtain estimates of the unknown parameters given the data.
The key difference between Bayesian statistical inference and frequentist statistical inference concerns the nature of the unknown parameters $\theta$.

In the frequentist tradition, the assumption is that $\theta$ is unknown, but that estimation of this parameter does not incorporate uncertainty.

In Bayesian statistical inference, $\theta$ is considered unknown and random, possessing a probability distribution that reflects our uncertainty about the true value of $\theta$.

Because both the observed data $y$ and the parameters $\theta$ are assumed random, we can model the joint probability of the parameters and the data as a function of the conditional distribution of the data given the parameters, and the prior distribution of the parameters.
More formally,

\[ p(\theta, y) = p(y|\theta)p(\theta). \]  

(6)

where \( p(\theta, y) \) is the joint distribution of the parameters and the data. Following Bayes’ theorem described earlier, we obtain

\[ p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta), \]  

(7)

where \( p(\theta|y) \) is referred to as the posterior distribution of the parameters \( \theta \) given the observed data \( y \).
Equation (7) represents the core of Bayesian statistical inference and is what separates Bayesian statistics from frequentist statistics.

Equation (7) states that our uncertainty regarding the parameters of our model, as expressed by the prior distribution \( p(\theta) \), is \textit{weighted} by the actual data \( p(y|\theta) \) yielding an updated estimate of our uncertainty, as expressed in the posterior distribution \( p(\theta|y) \).
Bayesian statistics requires that prior distributions be specified for **ALL** model parameters.

There are three classes of prior distributions:

1. **Non-informative priors**
   - We use these to quantify actual or feigned ignorance. Lets the data maximally speak.

2. **Weakly informative priors**
   - Allows some weak bounds to be placed on parameters.

3. **Informative priors**
   - Allows cumulative information to be considered.
The increased popularity of Bayesian methods in the social and behavioral sciences has been the (re)-discovery of numerical algorithms for estimating the posterior distribution of the model parameters given the data.

It was virtually impossible to analytically derive summary measures of the posterior distribution, particularly for complex models with many parameters.

Instead of analytically solving for estimates of a complex posterior distribution, we can instead draw samples from $p(\theta|y)$ and summarize the distribution formed by those samples. This is referred to as *Monte Carlo integration*.

The two most popular methods of MCMC are the Gibbs sampler and the Metropolis-Hastings algorithm.
A critically important component of applied statistics is hypothesis testing.

The approach most widely used in the social and behavioral sciences is the Neyman-Pearson approach.

An interesting aspect of the Neyman-Pearson approach to hypothesis testing is that students (as well as many seasoned researchers) appear to have a very difficult time grasping its principles.

Gigerenzer (2004) argued that much of the difficulty in grasping frequentist hypothesis testing lies in the conflation of Fisherian hypothesis testing and the Neyman-Pearson approach to hypothesis testing.
Fisher’s early approach to hypothesis testing required specifying only the null hypothesis.

A conventional significance level is chosen (usually the 5% level).

Once the test is conducted, the result is either significant ($p < .05$) or it is not ($p > .05$).

If the resulting test is significant, then the null hypothesis is rejected. However, if the resulting test is not significant, then no conclusion can be drawn.

Fisher’s approach was based on looking at how data could inform evidence for a hypothesis.
The Neyman and Pearson approach requires that two hypotheses be specified – the null and alternative hypothesis – and is designed to inform specific sets of actions. It’s about making a choice, not about evidence against a hypothesis.

By specifying two hypotheses, one can compute a desired tradeoff between two types of errors: Type I errors ($\alpha$) and Type II errors ($\beta$).

The goal is not to assess the evidence against a hypothesis (or model) taken as true. Rather, it is whether the data provide support for taking one of two competing sets of actions.

In fact, prior belief as well as interest in “the truth” of the hypothesis is irrelevant – only a decision that has a low probability of being the wrong decision is relevant.
The conflation of Fisherian and Neyman-Pearson hypothesis testing lies in the use and interpretation of the $p$-value.

In Fisher’s paradigm, the $p$-value is a matter of convention with the resulting outcome being based on the data.

In the Neyman-Pearson paradigm, $\alpha$ and $\beta$ are determined prior to the experiment being conducted and refer to a consideration of the cost of making one or the other error.

Indeed, in the Neyman-Pearson approach, the problem is one of finding a balance between $\alpha$, power, and sample size.

The Neyman-Pearson approach is best suited for experimental planning. Sometimes, it is used this way, followed by the Fisherian approach for judging evidence. But, these two ideas may be incompatible (Royall, 1997).
However, this balance is virtually always ignored and $\alpha = 0.05$ is used.

The point is that the $p$-value and $\alpha$ are not the same thing.

This is made worse by the fact that statistical software packages often report a number of $p$-values that a researcher can choose from after having conducted the analysis (e.g., .001, .01, .05).

This can lead a researcher to set $\alpha$ ahead of time, but then communicate a different level of “significance” after running the test.

The conventional practice is even worse than described, as evidenced by nonsensical phrases such as results “trending toward significance”.
The problem with NHST, therefore, seems to lie with the interpretation of the \( p \)-value.

To be precise, the \( p \)-value rests on a form of argumentation referred to as *reductio ad absurdum*.

This form of argumentation attempts to establish an assertion by deriving an absurdity from its denial, thus arguing that the assertion must be accepted because its rejection would be indefensible.

The \( p \)-value as reduction ad absurdum is thus properly interpreted as the probability of observing data at least as extreme as the data that was actually observed, computed under the assumption that the null hypothesis is true.
The $p$-value is based on data that were never observed.

The $p$-value is obtained by referencing the value of the observed test statistic (e.g. the $t$-test) based on the study at hand to hypothetically replicated data generated under the assumption that the null hypothesis is true.

The null hypothesis is never true in reality, and moreover, it is typically not the research question of interest.

Thus, after Cohen (1994), researchers are typically testing a “nil” hypothesis that is hopefully rejected.

However, it is important to emphasize that there is nothing within the NHST paradigm that requires testing a null hypothesis of no effect. In fact, any theoretically justifiable hypothesis can serve as the null hypothesis.
To summarize, the \( p \)-value does not quantify evidence for a hypothesis.

Rather, it provides a measure of the probability of an outcome not actually observed, computed under the assumption of a null hypothesis that will likely not to be true in any population and, moreover, is typically not of interest.

To quote Jeffreys (1961, pg. 385)

“This seems a remarkable procedure.”
Bayesian hypothesis testing begins first by obtaining summaries of relevant distributions.

The difference between Bayesian and frequentist statistics is that with Bayesian statistics we wish to obtain summaries of the posterior distribution.

The expressions for the mean (EAP), variance, and mode (MAP) of the posterior distribution come from expressions for the mean and variance of conditional distributions generally.
In addition to point summary measures, it may also be desirable to provide interval summaries of the posterior distribution.

Recall that the frequentist confidence interval requires two methods to imagine an infinite number of repeated samples from the population characterized by $\mu$.

For any given sample, we can obtain the sample mean $\bar{x}$ and then form a $100(1 - \alpha)\%$ confidence interval.

The correct frequentist interpretation is that $100(1 - \alpha)\%$ of the confidence intervals formed this way capture the true parameter $\mu$ under the null hypothesis. Notice that the probability that the parameter is in the interval is either zero or one.
Posterior Probability Intervals (cont’d)

- In contrast, the Bayesian framework assumes that a parameter has a probability distribution.

- Sampling from the posterior distribution of the model parameters, we can obtain its quantiles. From the quantiles, we can directly obtain the probability that a parameter lies within a particular interval.

- So, a 95% posterior probability interval would mean that the probability that the parameter lies in the interval is 0.95.

- The posterior probability interval can, of course, include zero. Zero is a credible value for the parameter.

- Notice that this is entirely different from the frequentist interpretation, and arguably aligns with common sense.
The frequentist and Bayesian goals of model building are the same.

1. Model specification based on prior knowledge
2. Model estimation and fitting to data obtained from a relevant population
3. Model evaluation and modification
4. Model choice
Despite these similarities there are important differences.

A major difference between the Bayesian and frequentist goals of model building lie in the model specification stage.

Because the Bayesian perspective explicitly incorporates uncertainty regarding model parameters using priors, the first phase of modeling building requires the specification of a full probability model for the data and the parameters of the model.

Model fit implies that the full probability model fits the data. Lack of model fit may be due to incorrect specification of likelihood, the prior distribution, or both.
Another difference between the Bayesian and frequentist goals of model building relates to the justification for choosing a particular model among a set of competing models.

Model building and model choice in the frequentist domain is based primarily on choosing the model that best fits the data.

This has certainly been the key motivation for model building, respecification, and model choice in the context of structural equation modeling (Kaplan 2009).

In the Bayesian domain, the choice among a set of competing models is based on which model provides the best posterior predictions.

That is, the choice among a set of competing models should be based on which model will best predict what actually happened.
A very simple and intuitive approach to model building and model selection uses so-called *Bayes factors* (Kass & Raftery, 1995)

These could be two regression models with a different number of variables, or two structural equation models specifying very different directions of mediating effects.

The Bayes factor provides a way to quantify the odds that the data favor one hypothesis over another.

A key benefit of Bayes factors is that models do not have to be nested.

The Bayes’ factor gives rise to the popular Bayesian information criterion (BIC)
The selection of a particular model from a universe of possible models can also be characterized as a problem of uncertainty. This problem was succinctly stated by Hoeting, Raftery & Madigan (1999) who write:

"Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are." (pg. 382)

An approach to addressing the problem of model selection is the method of *Bayesian model averaging* (BMA).
The major advantages of Bayesian statistical inference over frequentist statistical inference are:

1. Coherence
2. Flexibility in handling very complex models (through MCMC)
3. Inferences based on data actually observed
4. Quantifying evidence
5. Intuitive probability statements
6. Incorporating prior knowledge
To conclude, the Bayesian school of statistical inference is, arguably, superior to the frequentist school as a means of creating and updating new knowledge in the social sciences.

The Bayesian school of statistical inference is:

1. the only coherent approach to incorporating prior information thus leading to updated inferences and growth in knowledge.

2. represents an internally consistent and rigorous way out of the $p$-value problem.

3. approaches model building, evaluation, and choice, from a predictive point of view.

As always, the full benefit of the Bayesian approach to research in the social sciences will be realized when it is more widely adopted and yields reliable predictions that advance knowledge.
VIELEN DANK