Bayesian Inference in Linguistic Analysis

Stephany Palmer

1 | Introduction:

Besides the dominant theory of probability, Frequentist inference, there exists a different interpretation of probability. This other interpretation, Bayesian inference, factors in prior knowledge of the situation of interest and lends itself to allow strongly worded conclusions from the sample data. Although having not been widely used until recently due to technological advancements relieving the computationally heavy nature of Bayesian probability, it can be found in linguistic studies amongst others to update or sustain long-held hypotheses.

2 | Data/Information Analysis:

Before performing a study, one should ask themself what they wish to get out of it. After collecting the data, what does one do with it? You can summarize your sample with descriptive statistics like mean, variance, and standard deviation. However, if you want to use the sample data to make inferences about the population, that involves inferential statistics. In inferential statistics, probability is used to make conclusions about the situation of interest.

3 | Frequentist vs Bayesian:

The classical and dominant school of thought is frequentist probability, which is what I have been taught throughout my years learning statistics, and I assume is the same throughout the American education system. In taking a sample, one is essentially trying to estimate an unknown parameter. When interpreting the result through a confidence interval (CI), it is not allowed to interpret that the actual parameter lies within the CI, so you have to circumvent and say that “we are 95% confident that the actual value of the

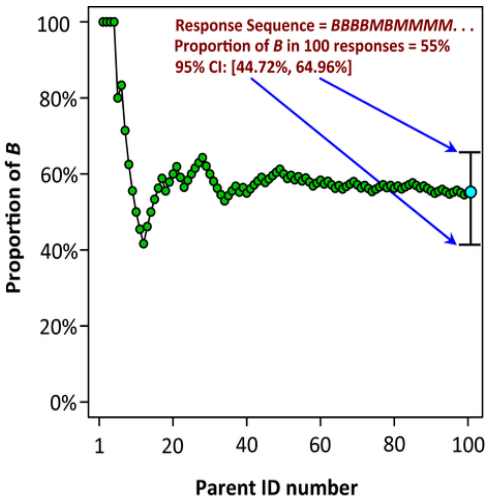
unknown parameter lies within the CI, that in infinite repeated trials, 95% of the CI's will contain the true value, and 5% will not."

Bayesian probability says that only the data is real, and as the unknown parameter is abstract, thus some potential values of it are more plausible than others. Using prior knowledge and studies, a Bayesian result is calculated and interpreted as "there is a 95% probability that the interval contains the true value" since different assumptions are used than in the frequentist approach.

A great, intuitive explanation of the difference in the two philosophies of probability:

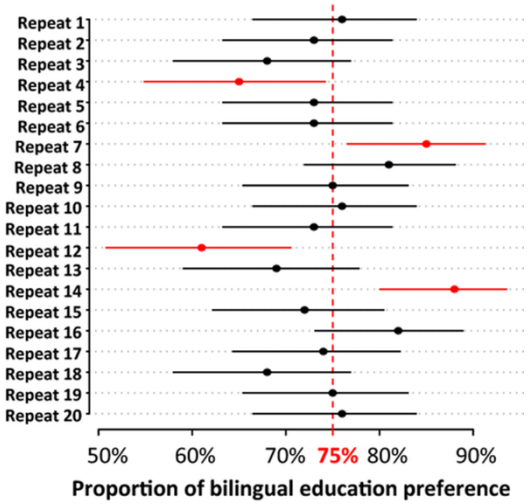
> "A frequentist is a person whose long-run ambition is to be wrong 5% of the time."
>
> "A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule." [1]

3.1 | An Example:



Given the example study used by Norouzian [4], let us determine the real proportion of parents who prefer bilingual education for their children over monolingual education. Responses from 100 randomly selected parents were collected, and the sample data showed 55% of parents prefer bilingual education and 45% prefer monolingual education. The 95% CI values are [44.72%, 64.96%].

According to frequentist theory, there is only one objective true proportion, since it is unknown, it is estimated with an associated unknown error. To account for error in the interpretation, the 95% frequentist CI of [44.72%, 64.96%] obtained from the survey is interpreted as "over infinite repetitions, 95% of the CIs constructed in this manner would contain the true population value."



In the figure to the right, the filled circles represent the observed proportion of parents who prefer bilingual education in each of the 20 repetitions of the survey. The solid horizontal lines passing through the filled circles are the 95% CIs for the obtained proportion of preferences for bilingual education in each of these 20 repetitions.

Let us assume that some higher power has assigned the real proportion of preferences for bilingual education in this population of parents to be 75%. In this case, some of the obtained proportions in these 20 repetitions have either underestimated or overestimated the real proportion of preferences for bilingual education, and do not contain the real proportion of preferences for bilingual education in their CI. In theory, if the survey is repeated infinitely many times, 95% of the obtained CIs will contain the real proportion of preferences for bilingual education.

From the Bayesian perspective, this need for infinitely many repetitions and careful language is unnecessarily complicated. A frequentist interpretation not only is not desired but it also could be a source of confusion for researchers wanting to interpret their single study's estimated results and conclusions. A better result is having x% certainty that a single obtained interval from a study has captured the actual population value.

4 | Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

If you are familiar with statistics and conditional probability, then the term Bayesian probably sounds familiar. Bayesian methods of inference are based on the Bayes' theorem, which states that the probability

of A given B can be found by multiplying the probability of B given A times the probability of A, divided by the probability of B.

4.1 | Bayesian inference:

As mentioned earlier, the Bayesian result takes prior information into account, alongside data obtained from the current sample. Thus, the Bayesian result is proportional to the prior x likelihood of getting our data given the priors, which is represented by P(A) prior and P(B|A) being likelihood. Probability of our data, P(B), is ignored, it just cannot be 0.

4.2 | Prior:

The prior summarizes our knowledge before collecting data. This probability distribution can be based on previous studies and should be justified. Priors can be either vague/uninformative or highly informative, depending on how much is known about the situation of interest.

4.3 | Choice of Priors:

In some cases, prior knowledge is minimal, or not credible. Thus, priors that concentrate their weight on a certain range for a parameter may cause a bias in the Bayesian result. So, it is important to carefully select your priors.

1) Obtain a set of priors with different levels of representativeness for past research findings.
2) Obtain a Bayesian result using each prior.
3) Compare their credible intervals.
    -If big differences exist, priors are not robust/useful.

4.4 | Likelihood:

Likelihood is the foundation of most frequentist techniques. It answers the question, “given our prior, how likely is it that we got our data?” If our prior and our data are very dissimilar, we will get a low likelihood. A low likelihood shows us that the prior is informative(in informing us that something

is missing in our knowledge). The less informative, the better our assumptions and knowledge hold.

4.5 | Posterior:

The posterior answers the question, "given our data, how likely is it that our prior beliefs accurately represent the data?" The posterior distribution contains the information needed for statistical inference and summarizes our knowledge of the parameters of interest based on:

What we assumed at the beginning of the study, and

What we have learned about our parameters after observing all the data.

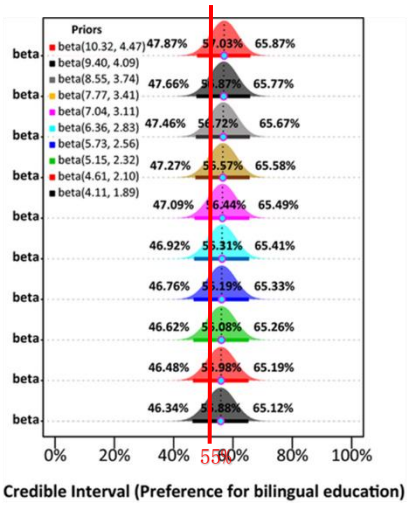
4.6 | Credible Interval:

Credible intervals are the Bayesian equivalent of frequentist confidence intervals but can use stronger language in their interpretation due to the Bayesian definition of probability.

For example, a 95% frequentist confidence interval for the difference in two group means might be (−1.25, −0.55). The frequentist interpretation is strictly worded such that if constructed in the same manner, infinite confidence intervals will capture the true value 95% of the time, and thus "we are 95% confident that the true difference in means lies within the interval (-1.25,-0.55)." We cannot interpret the likelihood of the true difference being captured in this one CI.

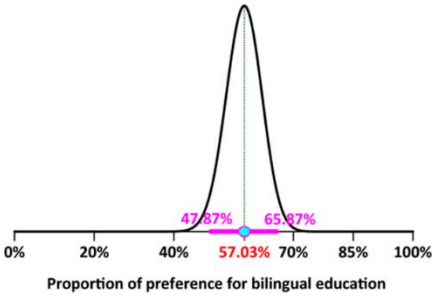
If we obtained the same values from a Bayesian credible interval, we are able to interpret it as: "given the prior and model, we believe that the true mean difference lies between −1.25 and −0.55 with 95% probability."

4.7 | Example continued:



Looking back at past studies shows that the proportion of parents who prefer bilingual education has stayed in the range of 60%-80% and that there have been increasing efforts to promote bilingual education over time. Based on this knowledge, the values of proportions from different surveys found to be smaller than 60% or larger than 80% are less likely to represent the real proportion of bilingual education preference in the population. The likelihood function assigns the highest weight to the obtained proportion of bilingual education from the current survey (55%). With the prior distributions and the likelihood, we can calculate the Bayesian resulting credible interval.

The researcher is able to state that there is a 95% probability that the real proportion of preferences for bilingual education in the population of parents could credibly range between 47.87% and 65.87%.



5 | Advantages:

The Bayesian approach is more of a commonsense interpretation of probability, and the use of priors presents a formal means to incorporate scientific knowledge and expertise into statistical models. The "updating" nature of Bayesian statistics is natural to many fields, especially in machine learning.

Also, Bayesian methods are no longer bogged down because they are computationally intensive; Bayesian methods can now be easily implemented using analytic tools and programming languages such as SAS, Open BUGS, JAGS, Stan, and multiple packages in R.

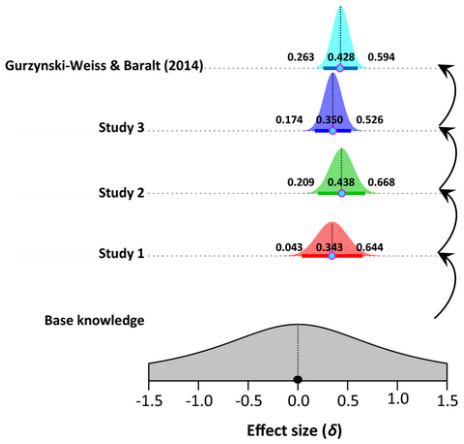
5.1| Disadvantages:

Frequentists' main objection to the Bayesian approach is the use of prior probabilities. Their criticism is that there is always a subjective element in assigning them and thus the Bayesian result is always biased.

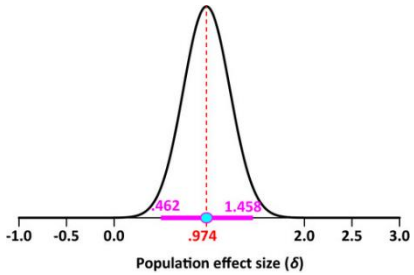
Frequentist methods took the lead since they are not as computationally intensive. Which has led to inferential statistics being mainly taught through the frequentist approach and has led to less people aware of a different philosophy of probability.

6 | Another Example:

Let us consider a study by Gurzynski-Weiss (2014). One of the questions researched was the effect of the interaction mode (face-to-face vs over computer) when they allowed 24 intermediate-level learners of Spanish to correct themselves during interactional feedback meetings with their teacher. After collecting the data, the authors conducted a paired-samples t-test to answer their research question, finding their critical t-statistic to be 5.03, with results showing better performance with face-to-face interaction.



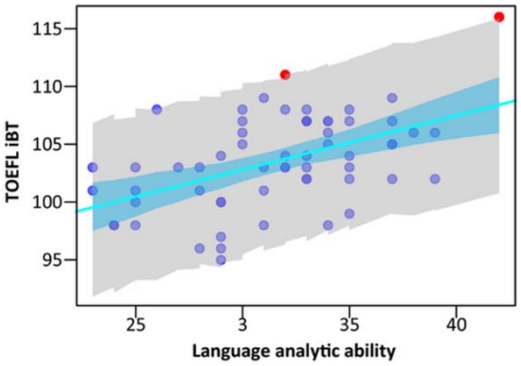
If they had used Bayesian methods, the results could have been interpreted as "there is a 95% probability that the real superiority of the face-to-face interaction over computer-mediated communication in providing more opportunities for intermediate Spanish foreign language learners to correct their output is found to be between .462 and 1.458.



7 | Other Uses of Bayesian Methods - Bayesian Regression:

The set-up of the linear regression model is the same; the role of Bayesian inference starts once we set out to find the unknown parameters in our overall statistical model. Instead of thinking that there is one true answer to each of

the unknown parameters (βs in standard linear model $E(y) = \beta_0 + \beta_1 x$), we can follow the Bayesian approach by assigning to each a prior.



Let us suppose that we want to know if language proficiency measured via TOEFL iBT scores of 60 advanced EFL learners can be predicted by their language analytic ability.

The values eventually used to draw the Bayesian regression line are only one set of likely values for predicting EFL learners' TOEFL iBT scores from their language analytic ability scores. When we consider other likely values for the intercept (α) and slope (β) from their posteriors, many other possible regression lines begin to appear to form a halo around the original regression line.

8 | Conclusion:

Thanks to modern advancements, Bayesian methods have become readily accessible. This allows for stronger interpretations to be made from data using previous studies and updating them to reflect a more accurate understanding of the situation of interest. As the difference in the two schools of probability is mainly philosophical, both are valid. However, there are certain circumstances and fields that lends themselves to using one approach over the other.

8.1 | Reflection:

It was a lot harder than I thought it would be to find a specific mathematical concept to talk about from the broad view of statistics in linguistic studies. Thus, I chose to explain the relatively simple difference in philosophies using linguistic study examples. I had hoped to find a narrower focus, but since my linguistics prior knowledge was based off of an introductory class, my resulting presentation was biased. It still conveyed what I learned about Bayesian inference, but tying in some linguistics concepts would have made the presentation better.

References:

[1] Annis, Charles. Frequentists and Bayesians, 8 June 2014, www.statisticalengineering.com/frequentists_and_bayesians.htm.

[2] Cthaeh, The, et al. “Frequentist and Bayesian Approaches in Statistics.” Probabilistic World, 19 Apr. 2020, www.probabilisticworld.com/frequentist-bayesian-approaches-inferential-statistics/.

[3] Gurzynski‐Weiss, L., & Baralt, M. (2014). Exploring learner perception and use of task‐based interactional feedback in FTF and CMC modes. Studies in Second Language Acquisition, 36, 1‐37. https://doi-org.proxy.wm.edu/10.1017/s02722631130003

[4] Hsu, Anne S., et al. “The Probabilistic Analysis of Language Acquisition: Theoretical, Computational, and Experimental Analysis.” Cognition, vol. 120, no. 3, 26 Mar. 2011, pp. 380‐390. ScienceDirect, doi:https://doi.org/10.1016/j.cognition.2011.02.013.

[5] Norouzian, R., de Miranda, M. and Plonsky, L. (2018), The Bayesian Revolution in Second Language Research: An Applied Approach. Language Learning, 68: 1032-1075. doi:10.1111/lang.12310

[6] Oleson, J. J., Brown, G. D., & McCreery, R. (2019). The evolution of statistical methods in speech, language, and hearing sciences. Journal of Speech, Language and Hearing Research (Online), 62(3), 498-506. doi:http://dx.doi.org.proxy.wm.edu/10.m44/2018_JSLHR-H-ASTM-18-0378