**INTRODUCTION TO BAYESIAN STATISTICS: SPSS**

**by Simon Moss**

|  |
| --- |
| **Introduction** |

**Frequentist approaches versus Bayesian approaches**

 When researchers are introduced to statistics, they often learn about t-tests, correlations, ANOVAs, and linear regression. When researchers conduct these techniques their conclusions are based solely on the sample of data they have collected. These analyses are typically called frequentist techniques.

 Yet, in the last couple of decades, many researchers have advocated and applied an alternative approach, called Bayesian statistics. When researchers conduct Bayesian statistics, their conclusions are not, in general, based solely on the sample of data they have collected. Instead, their conclusions are derived from both the sample of data and their initial predictions.

**Prior and posterior probabilities**

 To illustrate the rationale that underpins Bayesian statistics, suppose you wanted to predict how long research candidates will need to complete a PhD. For example, you might want to ascertain whether women are more likely to complete more rapidly than men or vice versa. Initially, because girls tend to outperform boys at school, you might predict that women might complete their PhD more rapidly than men. The following graph depicts this prediction, called the prior distribution. In this instance

* the x axis represents the duration that men need to complete their thesis minus the duration that women need to complete their thesis
* the y axis represents the probability of each difference
* specifically, this graph shows the most probable outcome is that men need one year more than women to complete their thesis



You then access a database to determine the number of years that 250 men and 250 women needed to complete their thesis. Finally, using SPSS, R, or some other statistical package, you conduct a Bayesian technique to update this prediction. The following graph depicts this updated prediction, called the posterior distribution. In this instance

* this graph shows the most probable outcome is that men need 1.5 years more than women to complete their thesis
* this graph is not as wide as the previous graph, indicating the researcher is now more certain the difference between men and women is close to 1.5



 In short, when researchers conduct Bayesian statistics,

* they often begin with an estimate of each parameter—such as the difference between two conditions or the correlation between two measures
* in addition, they estimate their certainty of this prediction—as represented by the width of this distribution
* next, they collect data and conduct Bayesian analysis
* this analysis will update the estimate and typically increase the certainty of this estimate.

|  |
| --- |
| **Example of Bayesian analysis: Comparing two groups** |

 To appreciate the difference between frequentist statistics, such as independent t-tests, and Bayesian approaches, suppose a researcher collects the following data. In particular

* each row corresponds to one research candidate
* the first column specifies the number of years research candidates need to complete their PhD thesis
* the second column indicates their gender; 0s represent females and 1s represent males
* the data is entered into SPSS 25. Earlier versions may not include Bayesian statistics.

Even if you do not use SPSS 25, follow this example anyway to learn about Bayesian statistics. You can then learn how to conduct Bayesian statistics in R later.



**Independent t-test**

 Initially, the researcher decides to conduct an independent t-test—the test that is often utilized to compare two groups on a numerical measure. In particular, after choosing “Analyse”, “Compare means”, and then “Independent Samples t-test”, the following dialogue box appears.



 To conduct this t-test

* transfer “Years to complete” into the box called “Test Variables”
* transfer “Gender” into the box called “Grouping variable
* click “Define Groups”, and indicates the two groups correspond to the numbers 0 and 1 respectively, before choosing “Continue”
* click “OK” to generate the following output





 As these tables show

* the mean or average years to complete is higher in men—4.977—than in women—3.915
* this difference is significant, as shown in the column called Sig (2-tailed)
* the 95% confidence interval of this difference is -1.7 to -.38.

**Bayesian statistics**

 The researcher now decides to utilize Bayesian statistics to analyse these data. In particular, after choosing “Analyse”, “Bayesian statistics”, and then “Independent Samples normal”, the following dialogue box appears.



 To conduct this test, initially click the same options you chose when conducting the t-test. That is

* transfer “Years to complete” into the box called “Test Variables” and transfer “Gender” into the box called “Grouping variable
* click “Define Groups”, and indicate the two groups correspond to the numbers 0 and 1 respectively, before choosing “Continue”

However, unlike the t-test, you need to consider some additional choices. Specifically

* Click “Use Both Methods” towards the bottom of this dialogue box
* Click “Priors” to generate the following screen.



 In this screen, you can enter information about the prior distribution—that is, your expectations about that means and variances of each group or condition. To simplify this example, however, retain the default. This default assumes you have not developed any specific expectations about the data—sometimes called a non-informative prior. Thus, press Continue and then OK to generate the following output.









This output might look complex but is actually straightforward. Specifically,

* the first table, called Group Statistics, provides the same information as did the independent t-test
* the second table, called Bayes Factor Independent Samples Test, indicates the difference in means between men and women is 1.06. This difference is significant, because the p value is .004, the same value generated by the independent t-test.

You might, however, be unfamiliar with some of the other statistics. To illustrate, in the second table, the Bayes Factor is .094. The definition of this statistic depends on which Bayesian technique you conduct. In this example, as the note below the table indicates

* the Bayesian Factor is roughly the likelihood the null hypothesis is true divided by the likelihood the alternative hypothesis is true
* as this statistic falls below 1, the alternative hypothesis becomes more likely to be true than is the null hypothesis
* the following table can be used to interpret these Bayes Factors. In this instance, the Bayes Factor is between 1/3 and 1—indicating anecdotal or tentative evidence for the alternative hypothesis that men and women differ on completion

| **Bayes Factor** | **Evidence Category** |
| --- | --- |
| >100 | Extreme Evidence for H0 |
| 30-100 | Very Strong Evidence for H0 |
| 10-30 | Strong Evidence for H0 |
| 3-10 | Moderate Evidence for H0 |
| 1-3 | Anecdotal Evidence for H0 |
| 1 | No Evidence |
| 1/3-1 | Anecdotal Evidence for H1 |
| 1/10-1/3 | Moderate Evidence for H1 |
| 1/30-1/10 | Strong Evidence for H1 |
| 1/100-1/30 | Very Strong Evidence for H1 |
| 1/100 | Extreme Evidence for H1 |

H0: Null hypothesis

H1: Alternative hypothesis

The third table presents the posterior mean. In this example, the posterior mean is the mean difference between men and women on duration to complete a PhD—but, unlike the observed mean, combines information about the prior expectations and actual data. However,

* in this instance, the observed mean difference equals the posterior mean difference
* the reason is that we did not include any information about prior expectations in the analysis; so, the conclusion was derived from the data only.

Finally, this table also presents a 95% credibility interval—analogous to a 95% confidence interval in frequentist statistics. In this instance, the researcher can be 95% certain the difference between the two means lies between .69 and 1.4.

**Difference between frequentist statistics and Bayesian statistics**

This simple example offers some insights into key differences between frequentist statistics and Bayesian statistics. The following table outlines these differences.

|  |  |
| --- | --- |
| Frequentist statistics | Bayesian statistics |
| Derives conclusions only from the data | Can derives conclusions from the data only or from a blend of prior expectations and the data.  |
|  | Therefore, Bayesian statistics enables researchers to integrate distinct sources of information better |
| **Generates a p value** | **Generates a Bayesian Factor** |
| The p value indicates the proportion of times a sample would generate a difference of this magnitude or more had the null hypothesis been true | The Bayesian Factor indicates the likelihood of the null hypothesis relative to the likelihood of the alternative hypothesis |
| The p value is hard to understand but, because of the .05 criterion, is easy to use | The Bayesian Factor is easier to understand, although researchers have yet to develop a standard criterion to reach conclusions. Nevertheless, the previous table specifies the criteria that many researchers utilize |
| **Generates a 95% confidence interval** | **Generates a 95% credibility interval** |
| The 95% indicates that, if this sample was collected an infinite number of times, 95% of these confidence intervals would contain the true parameter—such as the actual difference between the means  | The 95% credibility interval is how certain we are the parameter lies within this interval. |
| This confidence interval is intuitively hard to understand | The credibility interval is more intuitive to most researchers |

|  |
| --- |
| **Another example of Bayesian analysis: Predicting a numerical outcome** |

 To appreciate the benefits of Bayesian analysis, you need to learn about how you can specify your prior expectations as well as the various options you can choose. However, before you learn about these complexities, another simple example could be helpful. In particular, suppose now the researcher wants to ascertain whether the age, IQ, and GPA of research candidates, in addition to their gender, affects the duration they need to complete their thesis. An extract of the data appear in the following spreadsheet.



 Typically, to assess whether a series of predictors are related to some numerical outcome, researchers often conduct a multiple regression analysis, sometimes called a linear regression analysis. Specifically, the researchers could

* choose the “Analyse” menu, and then “Regression” and finally “Linear”
* transfer “Years to complete” into the box called “Dependent”
* transfer the other predictors into the box called “Independent(s)”.

After they press OK, they would receive a series of tables. One of the most important tables is labelled “Coefficients”. In this instance, the table shows that

* gender is significantly associated with years to complete after controlling the other predictors
* none of the other predictors, however, are significant.



**Bayesian regression**

 To conduct the Bayesian variant of this linear regression, select the “Analyze” menu and choose “Bayesian” and then “Linear Regression” to generate the following screen. As illustrated in this screen

* transfer “Years to complete” into the box called “Dependent”
* transfer the categorical predictors into the box called “Gender” but the numerical predictors into the box called “Covariates”
* choose “Use both methods”



The most important, and perhaps challenging, task is to specify your prior expectations. To achieve this goal, click the Priors button to generate the following screen.



You can then adjust some of these defaults. For example, in the column called “Mean”, you could enter some values. For example

* if you feel that age should be marginally, but positively, associated with years to complete, you could enter .2 in the box alongside age
* enter 0 in boxes if you have not formed any expectations







 To interpret these findings, you can largely apply the same rationale that you would utilize to interpret linear regression. For example

* the Bayesian Estimates of Coefficients, in the column called posterior mean, are tantamount to regression coefficients
* yet, these Bayesian coefficients differ from the regression coefficients, derived from the previous regression, because they are shaped by prior expectations as well
* if the credible interval does not include zero, you can be 95% certain the coefficient is not zero.

|  |
| --- |
| **Other Bayesian analysis** |

Thus far, this document has illustrated the Bayesian variants of independent t-tests and linear regression. Researchers have actually developed Bayesian variants of many statistical techniques. To illustrate, the first column of the following table presents a series of common statistical tests. The second column specifies the option you would chose in SPSS to conduct the Bayesian equivalent of each test.

|  |  |
| --- | --- |
| Frequentist statistical technique | To conduct the Bayesian equivalent in SPSS, choose “Analyze”, “Bayesian Statistics”, and then… |
| One sample t-test | One sample normal |
| Paired-samples or dependent t-test | Related samples normal |
| Independent samples t-test | Independent samples normal |
| Pearson correlation | Pearson correlation |
| Linear or multiple regression | Linear regression |
| One-way ANOVA | One-way ANOVA |
| Chi-square test of independence | Log-linear models |

In addition, SPSS also includes some other options. For example

* one sample binomial can assess whether the proportion of one category on a dichotomous variable differs from some expectation
* one sample Poisson can assess whether some variable conforms to a Poisson distribution

|  |
| --- |
| **The underlying rationale** |

 To apply Bayesian statistics effectively, you need to understand the rationale that underpins this approach more comprehensively. To understand this rationale, consider the following example. You have written a textbook entitled “How to complete a thesis in 5 easy steps”. You then want to estimate the percentage of research candidates who would purchase this textbook. In particular, you want to generate a graph, or distribution, that resembles the following pattern. Note

* To simplify the example, I have drawn this graph as a series of columns
* I could have drawn this graph as a continuous line, similar to the first graph in this document



 This graph is called the **posterior** distribution. To interpret this graph

* the highest column corresponds to a proportion between 0.6 and 0.7
* hence, the most likely proportion of candidates who purchase the book is about 0.65
* the proportion of candidates who purchase the book could be as low as .3 and as high as .9
* yet, the pattern is quite narrow, indicating a fairly high level of certainty
* in contrast, the following pattern is more dispersed, indicating quite a low level of certainty



**Step 2 Collect a sample of data**

 So, how do you generate this posterior distribution? You need to collect some data. That is

* you might ask 10 candidates whether they would like to purchase your book
* you might then discover that 7 candidates purchase the book
* hence, in this sample, 0.7 or 70% of candidates purchased the book
* therefore, your best estimate of the proportion in the population is also 0.7 or 70%
* however, this value of 0.7 does not clarify the certainty of your prediction—that is, whether the distribution is narrow or dispersed.

**Step 1 Specify the prior distribution**

 **Before** you collect your data, however, you should specify the prior distribution. The prior distribution, in essence, is your prediction of the posterior distribution. For example, before you collect data, you might predict the distribution is likely to resemble the following graph:



This graph implies you believe that a proportion of .5 to .6 is the most likely, but also implies you are quite uncertain. Nevertheless, if you have not received any information to indicate that one proportion is more likely than any other proportion, your prior distribution could resemble the following graph instead—sometimes called a uniform distribution or noninformative prior.



**Step 3 Identify the generative model**

 To conduct Bayesian analysis, you need to specify a method that converts a specific parameter, such as the proportion of people in the population who would purchase the book, to the data that is likely to emerge, such as the proportion of people in your sample who would purchase the book. This method is called a generative model. Which generative model you would choose depends on the circumstances. To illustrate

* suppose the proportion of people in the population who would purchase the book is 60%
* if you asked a sample of 10 candidates whether they want to purchase the book, you might expect 6 individuals to say yes
* but actually, every time you repeat this procedure, the number of candidates who purchase the book would vary—such as 6, 4, 9, 7, and so forth—as the following figure shows



 In this circumstance, you can use a special model, called a binomial distribution, to predict the likelihood that a specific number of candidates will purchase the book. This model is illustrated in the following table. To clarify, in this table

* 10! indicates 10 x 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1 and so forth
* according to the first row, the probability that 0 individuals in a sample of 10 candidates will purchase the book is .000105
* according to the second row, the probability that 1 individual in a sample of 10 candidates will purchase the book is .00157
* and so on

|  |
| --- |
| p (0 purchases) = 10! /(0! X 10!) x 0.60 (1 – 0.610) = .000105  |
| p (1 purchases) = 10! /(1! X 9!) / x 0.611 (1 – 0.69) = .00157 |
| p (2 purchases) = 10! /(2! X 9!) / x 0.612 (1 – 0.68) = .01060 |
| p (3 purchases) = 10! /(3! X 9!) / x 0.613 (1 – 0.67) = .00424 |
| and so forth |

 This pattern conforms to a formula. You do not need to know the formula, but merely be aware that researchers have developed a formula—as specified in the following table. This formula is called the binomial distribution. Often, if you want to predict the likelihood of a certain number of successes from a particular number of attempts, this formula is a suitable generative model.

|  |
| --- |
| p(k successes from n attempts) = n!/k!(n-k)! x pk x (1-p)1-k |

**Step 4 Apply the generative model repeatedly**

 Finally—and here is the clever part—you apply the generative model repeatedly. However, you retain only the proportions that generated simulated data that match the actual data. This principle, although confusing now, will seem insightful in a moment. Here is the procedure:

|  |
| --- |
| First, from the prior distribution—such as the uniform distribution in the following graph—randomly choose one proportion, called a parameter. |

In particular, the likelihood you would choose a specific proportion, such as .42 or .62, should depend on the frequency or height of the graph at this point. In this instance, because the distribution is uniform, you would choose each proportion, such as .42 or .62, with equal probability. To illustrate, you might chose **0.7**



|  |
| --- |
| Second, apply the generative model—in this instance, the binomial distribution—to generate data that you might have collected. These data are sometimes called simulated data.  |

To illustrate, suppose the proportion you choose was 0.7, and your sample comprises 10 individuals. According to the binomial distribution formula

* the probability that 0 candidates purchase the book is .0000059
* the probability that 1 candidate purchases the book is .00013
* the probability that 2 candidates purchase the book is .00114
* and so forth

You would then use a computer program to randomly choose the number of candidates who purchased the book—but proportionate to these probabilities. For example, this program might estimate that 4 of 10 candidates approached decided to purchase the book. That is, the simulated data indicates that 40% of candidates purchased the book.

|  |
| --- |
| Third, discard this proportion—also called a parameter—if this estimated number differs from the actual data you collected |

 To illustrate, when you collected actual data, you discovered that 6 of the 10 candidates you approached purchased the book. In this simulation, when the proportion was .7, only 4 of the candidates purchased the book. The actual data diverged from the simulated data. You would thus discard this .7.

|  |
| --- |
| Fourth, repeat this procedure thousands of times. Construct a frequency distribution from the proportions or parameters you did not discard.  |

 To illustrate, consider the following table. Each row corresponds to one simulation. In addition

* the first column presents the proportion or parameter you randomly generated each time
* the second column presents the simulated data, generated when the proportion or parameter was subjected to the binomial distribution
* the third column presents the actual data—a number that is always 6 out of 10
* the fourth column indicates whether the simulated data is equal to the actual data
* the final column indicates whether the proportion or parameter was retained, because the simulated data equals the actual data, or discarded

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Proportion | Simulated data – estimated number of people out of 10 who purchase the book | Actual number of people out of 10 who purchased the book | Simulated data equals actual data | Retained or discarded |
| 0.6 | 6 | 6 | Yes | Retained |
| 0.4 | 5 | 6 | No | Discarded |
| 0.7 | 8 | 6 | No | Discarded |
| 0.5 | 6 | 6 | Yes | Retained |
| … | … | … | ,,, | … |

 The computer would then generate a frequency distribution from the proportions in the first column, but only from the retained values. The following graph illustrates this frequency distribution that might emerge. This frequency distribution is the posterior distribution.



**Overview of rationale**

 You might have followed these steps, but not quite appreciated the underlying rationale. So, to outline this procedure, the computer program

* begins with a prior distribution, corresponding to your prediction of the posterior distribution
* repeatedly draws a proportion, also called a parameter, from this prior distribution
* subjects each of these proportions or parameters to a generative model, such as a binomial distribution, to generate simulated data
* retains only the proportions or parameters that generated simulated data that match the actual data
* in essence, only a subset of the prior distribution is retained—the subset that generated data that matches the actual data

|  |
| --- |
| **Caveats to the rationale** |

 The previous section outlined the rationale that underpins Bayesian analyses. In contrast, this section introduces some of the complexities or nuances of this rationale.

**Multiple parameters**

 The previous example was designed to estimate one parameter: the proportion of candidates who would purchase the book. A similar rationale, however, can be applied to estimate more than one parameter—such as the intercept and slope in a regression equation. Yet, despite this complexity, the rationale is the same. To illustrate

* the computer derives a possible intercept and slope from a prior distribution—such as 1.3 and -.04 for the intercept and slope respectively
* the computer subjects these numbers to a generative model, such as a regression equation
* the two parameters—1.3 and -.04—are subjected to this generative model.
* if this model generates the same intercept and slope as did the actual data that were collected, these parameters are retained
* the retained parameters accrue to become the posterior distribution.

**Approximate Bayesian computation**

 According to the previous section, the computer, when conducting Bayesian analysis, extracts a parameter, such as a proportion, from a prior distribution. The computer will then retain only the parameters that, when subjected to the generative model, generate data that mirror the actual data. This particular approach is called Approximate Bayesian Computation. Unfortunately

* this approach, although quite easy to understand, may consume a long time in practice
* the computer might generate lots of data that diverges from the actual data
* each of these repetitions will discard the parameter and, therefore, waste time

In short, Approximate Bayesian Computation can be inefficient and slow, especially if the model comprises many parameters. Instead, researchers have developed a range of other algorithms to generate the posterior distribution. These algorithms generate very similar results to the Approximate Bayesian Computation, but are faster. To clarify

* one class of models is called Markov Chain Monte Carlo
* this class includes many specific kinds, such as Metropolis-hastings, Gibbs sampling, Hit-n-run, the t-walk, and the Hamiltonian Monte Carlo

**Rationale of the Markov Chain Monte Carlo**

To illustrate a Markov Chain Monte Carlo, suppose you did want to estimate the intercept and slope of a regression equation. Initially, the computer might choose an intercept and slope from the range of possibilities. In this example

* as the following graph shows, the computer first chooses an intercept of about .2 and a slope of about -.15, as represented by the red dot
* the computer then enters these parameters into the generative model to generate simulated data
* the computer then chooses another intercept and slope, as represented by the blue dot, and so forth



 Interestingly, however, the computer learns to choose parameters that generate simulated data that match the actual data. In particular, the computer somehow chooses maximum likelihood parameters—parameters that are most likely to generate simulated data that resemble the actual data. We will not discuss the precise algorithm now, however. You merely need to appreciate that other approaches can be used to generate the posterior distribution.

**Why is this approach called Bayesian?**

 Many researchers believe that Bayesian analysis should be called something else, like uncertainty or probability modelling. Nevertheless, this term evolved because the computations utilize a formula called Bayes theorem. The following table specifies this formula.

|  |  |
| --- | --- |
| p ( | data) =  |  p (data | x p ( |
|  |  p (data | x p (  |

 To decipher this formula

* p ( | data) represents the “probability a specific set of parameters are correct given the data that was collected. represents a specific set of parameters. | represents the word “given”
* p (data |  represents the “probability of this data given these parameters”
* p (represents the probability of a specific set of parameters
*  p (data | x p (is the same as the numerator, but summed across all possible sets of parameters

Perhaps this description does not help. In essence, the left hand side represents what you want to discover: how likely is a specific parameter, such as a proportion of 0.6, given the data. The right hand side represents what you know from the data and priors. This formula might look confusing but actually, somehow, describes the rationale that was presented earlier in mathematical terms.

|  |
| --- |
| **Prior distributions** |

**How to estimate the prior distribution**

 In the previous example, the prior was uniform or non-informative. Bayesian statistics is more effective—and the posterior distribution is likely to be narrower—if the prior distribution is not uniform, such as the example below.



But, how do researchers generate these prior distributions? Which information should researchers utilize to inform these prior distributions?

* This prior distribution could simply be the posterior distribution from a previous study or a pilot study
* This prior distribution could be an estimate, based on comments from experts. In this example, an expert might have indicated that “the proportion of people who will purchase your book is likely to be around .55 but, although unlikely, could be as low as .05 or as high as 0.95.

**How to specify the prior distribution in SPSS**

 If using SPSS, or even other software, you cannot merely draw this prior distribution. You need some other way to convey this information to the computer. For example, rather than draw this prior distribution, you might

* specify the approximate mean and standard deviation of these numbers
* specify other parameters, depending on the circumstances.

For each circumstance or test, you need to specify different parameters. The following examples display the screens that enable you to specify these parameters—for comparing two groups, correlating two variables, and comparing more than two groups respectively. Sometimes, the information you need to enter is intuitive. Sometimes, the information you need to enter is not intuitive—and you might need to seek advice, Google the options, or even guess.







|  |
| --- |
| **R** |

 Many packages have been developed to conduct Bayesian statistics in R. For example

* many researchers use BUGS or JAGS
* some researchers use STAN—a programming language that interfaces with R and applies a specific algorithm called Hamiltonian Monte Carlo

These options will be discussed in another document.