Using Excel to Calculate Bayes Factors: A Goodness-of-Fit Test (Chi-Square Test) as an Example

Authors: Zhong Wang, Zhong Wang

Date: 2023-07-02T00:00:00+00:00

Abstract

Taking the goodness-of-fit test (Chi-square test) as an example, this paper attempts to calculate the Bayesian factor BF10 for n-fold Bernoulli trials using Excel software (with JASP software serving as a benchmark). The results showed that within the range of 0.15–0.55 (the proportion of 'true' outcomes), Excel produced more accurate calculations, and the differences between the two software packages (Excel and JASP) were not statistically significant (P>0.3).

Full Text

Using Excel Software to Calculate Bayesian Factors: A Goodness-of-Fit Test (Chi-Square Test) Example

WANG Zhong
Beijing Doers Education Consulting Co., Ltd, Beijing Doers Institute of Clinical Education

Abstract: Taking the goodness-of-fit test (chi-square test) as an example, this paper attempts to calculate the Bayesian factor BF_{10} for n-fold Bernoulli trials using Excel software, with JASP software results serving as validation. The results demonstrate that within the range of 0.15–0.55 (the proportion of samples that are all "true"), Excel calculations achieve greater accuracy, with no statistically significant differences between Excel and JASP outputs ($P > 0.3$).

Tags: Bayesian factor, goodness-of-fit test, Excel software

1. Introduction

Numerous publications have addressed the limitations of null hypothesis significance testing (NHST). Consequently, many journals now require reporting

Bayesian factors (BF) in addition to p-values. Extensive literature already discusses the drawbacks of p-values and the advantages of BF, which will not be reiterated here. Interested readers may consult articles by Fan Wu et al. (2018), Chuanpeng Hu et al. (2018), and others. However, BF calculation remains relatively complex (Fan Wu et al., 2018) and typically requires specialized software. Currently, the most commonly used tools are JASP and the R programming language. This creates a practical challenge: traditional statistical software such as SPSS remains prevalent in biological/medical and sociological (particularly psychological) research, yet SPSS offers limited capabilities for BF computation. Researchers must therefore download and learn additional software. Although many of these tools are free and open-source, some (such as R) require learning coding syntax. This predicament sparked our curiosity: can common office software like Excel be used to calculate BF?

This paper examines the feasibility of using Excel for BF calculation through the chi-square test method in goodness-of-fit testing (hereinafter referred to as the "Chi test"). The Chi test was selected for several reasons. First, goodness-of-fit testing represents one of the most fundamental statistical concepts in social statistics. In structural equation modeling, for instance, the core principle involves attempting to fit a model to empirical data to evaluate its appropriateness. The key to good fit lies in "goodness-of-fit indices" (Jietai Hou et al., 2021, p. 8), with the chi-square value being the most critical index. Second, while BF calculation for t-tests and ANOVA has been extensively documented in the literature, the Chi test has received comparatively little attention. Most importantly, the Chi test involves recursive calculations, which align perfectly with Excel's computational strengths.

2. Deconstructing the BF Factor

Since Excel lacks built-in functions for direct BF calculation, we must first deconstruct BF to identify the core components that need to be computed. According to Fan Wu et al. (2018), BF is typically denoted as BF_{10} and expressed as:

$$\frac{P(H_1|data)}{P(H_0|data)} = \frac{P(data|H_1)}{P(data|H_0)} \cdot \frac{P(H_1)}{P(H_0)}$$

where $\frac{P(data|H_1)}{P(data|H_0)}$ represents the required BF, and $\frac{P(H_1)}{P(H_0)}$ is the prior probability ratio that can be set based on existing knowledge. However, since prior distributions are typically unknown, this ratio is often set to 1 to indicate "no bias toward either the alternative or null hypothesis" (Fan Wu et al., 2018).

When $\frac{P(H_1)}{P(H_0)} = 1$, the formula simplifies to:

$$\frac{P(H_1|data)}{P(H_0|data)} = \frac{P(data|H_1)}{P(data|H_0)} = BF_{10}$$

Thus, we only need to solve for $P(H_0|data)$ and $P(H_1|data)$ respectively. Furthermore, by definition of null and alternative hypotheses, $H_0 \cap H_1 = \phi$ and $H_0 \cup H_1$ encompasses the entire sample space. From the properties of conditional probability:

$$P(H_0|x_1, \dots, x_n) + P(H_1|x_1, \dots, x_n) = P(H_0 \cup H_1|x_1, \dots, x_n) = 1 \quad \text{(Yici Zhang et al., 2000, p. 24)}$$

Therefore:

$$BF_{10} = \frac{1 - P(H_0|x_1, \dots, x_n)}{P(H_0|x_1, \dots, x_n)}$$

In other words, we need only solve for the conditional probability $P(H_0|x_1, \dots, x_n)$.

3. Calculating $P(H_0|x_1, \dots, x_n)$

Generally, Bayesian calculation of conditional probability requires knowledge of the prior distribution. However, prior distributions vary substantially across psychological experiments, complicating the computation. Recognizing that most psychological experiments involve n-fold Bernoulli trials—for example, when participants complete double-blind questionnaires, each participant's response can be treated as an independent trial—Yici Zhang et al. (2000) proposed an approach: if estimating the probability of occurrence for a particular n-fold Bernoulli event A, the prior distribution of A can be treated as a uniform distribution on (0,1), that is:

$$h(p) = \begin{cases} 1 & \text{for } p \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

where $h(p)$ is the density function for the probability p of event A occurring. Yici Zhang et al. provided the calculation method for the probability of A under these conditions:

$$\hat{p} = E[p|x_1, \dots, x_n]$$

where x_i is the ith observation, taking values of 0 or 1 (Yici Zhang et al., 2000, pp. 151–152). This reveals that the conditional probability is essentially an expectation.

However, note that probability p represents a point estimate, whereas formula requires a conditional probability containing the hypothesis $\{H_i\}$ (i = 0,1). Therefore, we must transform the problem into a point estimation framework

before applying the above method. To this end, we define event B: the frequency of event A in the first n trials equals 0.5. Let:

$$B_n = \begin{cases} 1 & \text{if the frequency of event A in the first n trials is exactly 0.5} \\ 0 & \text{if the frequency of event A in the first n trials is not 0.5} \end{cases}$$

If B_n occurs, then $\frac{\sum_{i=1}^{n} x_i}{n} = 0.5$. According to Cauchy's Convergence Test, if p is the probability of event A occurring (and p exists), then as $n \to \infty$, for all $\varepsilon > 0$:

$$\left| \frac{\sum_{i=1}^{n} x_i}{n} - p \right| < \varepsilon$$

By the Bernoulli Law of Large Numbers, $P\{|p - 0.5| > \varepsilon\} = 0$ (Yici Zhang, 2000, p. 124). Thus, we have proven that B_n converges to H_0 as $n \to \infty$. This transformation allows us to recast the question of whether event A occurs into whether event B occurs.

Now define $r_i = \left| \frac{\sum_{j=1}^{i} x_j}{i} - 0.5 \right|$, which clearly satisfies $r_i \in [0, 0.5]$ and represents the distance between the frequency of A and 0.5 in the first i trials. Additionally, let y_i be the ith observation of event B:

$$y_i = \begin{cases} 1 & \text{if } r_i = 0 \\ 0 & \text{if } r_i \neq 0 \end{cases}$$

From formula , we obtain:

$$\hat{p}_B = E[B_n | y_1, \ldots, y_n] = \frac{\sum_{i=1}^{n} y_i + 1}{n + 2}$$

where $\hat{p}_B$ estimates the overall probability of event B occurring. Since B_n converges to H_0 as $n \to \infty$, we have $\hat{p}_B \to P(H_0 | x_1, \ldots, x_n)$. Moreover, because $y_1, \ldots, y_n$ are uniquely determined by $x_1, \ldots, x_n$, for sufficiently large n:

$$P(H_0 | x_1, \ldots, x_n) = \hat{p}_B$$

As y_i is a piecewise function, expressing formulas and as a unified analytic expression proves overly complex and will not be elaborated here. Nevertheless, due to their recursive nature, formulas and can be readily implemented in Excel, as demonstrated below.

4. Determining the Unbiased Interval for r_i

Although BF_{10} can theoretically be calculated from formulas and , practical implementation faces a challenge: the probability of r_i being exactly 0 is extremely small. If y_i is set to 1 only when r_i strictly equals 0, Excel's computational precision may prove too stringent, causing all y_i to be judged as 0 and rendering formula ineffective. To prevent this, we must establish an unbiased interval $[0, k]$ for r_i, where $r_i \in [0, k]$ is still considered equivalent to $r_i = 0$. The value of k must minimally impact the overall conclusions.

Drawing from the principles of goodness-of-fit testing, we define:

- G_0: Under identical conditions, chi-square test results differ significantly between $r_i = 0$ and $r_i = k$
- G_1: Under identical conditions, chi-square test results show no significant difference between $r_i = 0$ and $r_i = k$

Since n-fold Bernoulli tests generally have degrees of freedom $df = 2$, the chi-square distribution table indicates we need only demonstrate:

$$P\{\chi^2 > 3.84 \mid G_0 \text{ holds}\} < 0.05$$

From Pearson's theorem:

$$\chi^2 = \sum \frac{(\gamma_i - np_i)^2}{np_i}$$

where γ_i is the observed frequency of the first i occurrences and np_i is the theoretical frequency (Yici Zhang et al., 2000, p. 188). For this problem, $df = 2$, $np_i = 0.5n$, and $\gamma_i = n(0.5 \pm k)$. Substituting these conditions into formula yields:

$$\frac{(n(0.5 \pm k) - 0.5n)^2}{0.5n} = 3.84$$

Since this expression contains n, the value of k is not unique. Setting $n = 200$ for BF calculation, we substitute and obtain $k \approx \pm 0.069$. Thus, the unbiased interval for r_i should be $[0, 0.07)$ when $n = 200$.

5. Excel Software Implementation

Based on the above, we developed the following Excel program:

Table 1 : Implementation of the above algorithm in Excel

The Excel implementation includes: Column A contains sequence numbers, which serve as renumbering indices and provide denominators for BF_0 calculation. Row 1 contains headers for x_i, r_i, y_i, and BF_0 (see Section 3 for detailed definitions). The specific Excel formulas are:

: '=ABS(AVERAGE(B$2:B3)-0.5)" for the meaning of this sentence, please refer to the above the definition of ri, where "ABS" is the absolute value function and "AVERAGE" is the arithmetic average function. Note that the first indicator in the AVERAGE function must be "B 2"(*mustcontain*""), otherwise Excel defaults to the average of two adjacent cells, not the average of the first i items. 5.1.4 this sentence means that if the value of cell C3 is less than 0.07, it will be assigned a value of 1, otherwise it will be assigned a value of 0. The number 0.07 is the upper bound of the value of k (not necessarily the supremum). Where "IF" is the judgment function of Excel.

5.2 Procedure test

When testing the program, we will find a problem: for the same set of sample values, if we disordered the order of the sample, then the calculated BF value is may not unique. The reason is that our definition of event B contains an implicit meaning of recursion (the value of B is determined of the frequency of A in the first i samples), and the recursive sequence itself is highly sensitive to the order, which causes the difference of BF. For this reason, we arranged all the sample sequence as {1, …, 1, 0, …, 0}. The reason for this order is that if all the samples which equal to 1 are at the end (namely {0, …, 0, 1, …, 1}), it is likely to cause a large number of cases where the BF value exceeds 100, resulting in algorithm failure.

6. Verification and correction

We took JASP software as the evidence software, and calculated BF, where the total number of samples n=200, when = 10, 20, 30, … , 110 . The higher (>110) BF values was not calculated because when it was greater than 110, the BF values given by Excel software has all exceeded and equal to 201. The results of Excel and JASP software are shown in the following table (since more often, we prefer to verify the alternative hypothesis, so here JASP we selected ">Test value" and "BF10", the same below):

The sum of xi=1 Table 2 : the results of Excel and JASP The frequency of occur of A Excel It is obvious that the results of the softwares are not the same, and the difference is large. The reason for this is probably related to the algorithm we choose (the idea of Yici Zhang et al. that the default prior distribution is uniform, but there are probably more suitable prior distributions). To this end, we need to seek the possibility of converting Excel results into JASP, that is, to correct Excel results.

First, let's look at the regression function of the overall results of the two. As shown in Figure 1 [Figure 1: see original paper]-2, the scatter plots of Excel and JASP results are shown respectively, and the dotted line are the regression functions. It can be seen from the regression function that the difference between the two are still large. The most obvious difference was that the Excel results have obvious upward in the range of 0-0.15 (corresponding to = 10, 20).

If we draw the scatter diagram of ri, it is not difficult to find the key of the problem (Fig. 3 [Figure 3: see original paper]-4): for = 10, 20 , because we have previously placed all the samples that A happened (that is, xi=1) at the beginning, this results in all the values of r10 and r20 being 0.5, and then the values of r20 and r40 will quickly decline to 0 (because ri is the arithmetic mean of the first i). Therefore, the essence of this deviation is caused by the order of samples mentioned above ({1, …, 1, 0, …, 0}). When > 20 , because the number of xi equal to 0 and 1 are gradually close, the bias effect also gradually disappears. Therefore, we rejected the values in [0, 30) (that is ri>0.15), let's look at the overall regression of the two of this time. As shown in Figure 5 [Figure 5: see original paper]-6, it can be seen that the shape of the two regression functions has been closer this time, and the R2 of the two regression functions are greater than 0.7, which is acceptable, indicating that the overall trend of the two functions has been very close.

Figure 1 the regression result of Excel EXCEL y = 330.68x2 - 216.23x + 36.793 R^{2} = 0.797 Figure 2 [Figure 2: see original paper] the regression result of JASP y = 3.3279x2 - 1.5514x + 0.1358 R^{2} = 0.7018 Figure 3 the value ri of when Σxi=10 (n=200) ri, Σxi=10 Figure 4 [Figure 4: see original paper] the value ri of when Σxi=20 (n=200) ri, Σxi=20 Figure 5 the regression result of Excel when rejected the value in [0, 30) EXCEL' y = 278.51x2 - 187.17x + 33.747 R^{2} = 0.7393 Figure 6 [Figure 6: see original paper] the regression result of Excel when rejected the value in [0, 30) JASP' y = 4.0663x2 - 2.3411x + 0.3081 R^{2} = 0.7471 But the problem now is that the R2 of the two is still far less than 0.9, that is, the error is still large. This means that, if the results of Excel are corrected according to these two regression functions at this time, it is likely that there is still a large gap between the results of JASP and Excel. Therefore, we cut the define domains of two regression functions into two parts respectively, 0.15-0.45 and 0.45-0.55, and recalculate their regression functions respectively, which may greatly reduce the error. Figure 7 [Figure 7: see original paper]-10 shows the scatter diagram and regression function of Excel and JASP in 0.15-0.45 and 0.45-0.55 respectively. In order to further reduce the error, we used the cubic function to carry out regression. It can be seen that R2 were greater than 0.99, and the fitting were very good.

Figure 7 the regression result of Excel in 0.15-0.45 EXCEL-1 y = -156.89x3 + 241.2x2 - 121x + 23.393 R^{2} = 0.9959 Figure 8 [Figure 8: see original paper] the regression result of JASP in 0.15-0.45 JASP-1 y = 2.2222x3 - 1.5524x2 + 0.3811x - 0.0231 R^{2} = 0.9955 Figure 9 [Figure 9: see original paper] the regression result of Excel in 0.45-0.55 EXCEL-2 y = 45797x3 - 66450x2 + 32165x - 5187.8 R^{2} = 0.9987 Figure 10 [Figure 10: see original paper] the regression result of JASP in 0.45-0.55 JASP-2 y = 1113.8x3 - 1619x2 + 784.47x - 126.62 R^{2} = 0.9993 It is also known that if we set two functions: = 1 3 + 2 2 + 3 + 4 = 1 3 + 2 2 + 3 + 4 then the above formula can be converted into: 1 3 + 2 2 + 3 + 4 = 1 2− 2 1 1 3− 3 1 1 4− 4 1 Where 1 0. That is, the form of upper formula can be transformed as the lower formula through algebraic operation. Calculate the corrected Excel value from formula , as

shown in the following table (Table 3):

Table 3 the results of Excel corrected The frequency of occur of A Excel Excel corrected It can be seen that "Excel corrected" and "JASP" were very close. But is that "close" statistically significant? For further confirmation, we used the independent sample t- test method to calculate the significance of the difference between the two groups of data (calculation software: SPSS24). As shown in Table 4 , the p value is greater than 0.3, which means that there is no significant difference between the two.

Table 4 the significant between Excel corrected and JASP Levene test Independent t-test Significance Significance Average SE interplation 95% CI of interpolated (two tailed) interpolation Lower limit Upper limit Un-EV *"EV" means the equal variance, abbreviations are used because of insufficient space of table cell.

To sum up, we corrected Table 1 as follows: Table 5 : corrected about Table 1 BF10 修 Among them, and have the same meaning as in Table 1. : "=(SUM(D3 : $D201$)+1)/($A201$+2)`, $which implements formula\ .Caremustbetakennottoomitthe$"" symbols.

: `=(1-E201)/(E201)`, which implements formula .

: `=(-0.01416)*F201+1.863972*AVERAGE(B2:B201)^2+(-1.33275)*AVERAGE(B2:B201)+0.35444`, which implements formula for the interval 0.15–0.45.

: `=0.02432*F201+(-2.91139)*AVERAGE(B2:B201)^2+2.205288*AVERAGE(B2:B201)+(-0.45078)`, which implements formula for the interval 0.45–0.55.

7. Discussion

In summary, this paper's overarching approach constructs event B based on a Bayesian point estimation algorithm, proves that $B \to H_0$ as $n \to \infty$, and thereby transforms the problem of solving $P(H_0|x_1, \ldots, x_n)$ into a point estimation problem. Finally, formula adjusts the results to align with JASP outputs.

However, this study leaves unresolved a critical question: what should be done when $\sum x_i > 110$? Beyond this threshold, Excel values saturate at 201, rendering the correction meaningless. To address this, we first determined the point at which Excel fails—specifically, when $\sum x_i$ exceeds approximately $0.57n$. We therefore calculated JASP values at $0.57n$, presented in Table 6 .

Table 6: JASP Bayesian factors when n > 200

$0.57n$	JASP BF
1.689+10^{7}	

The JASP values increase substantially with n, and when $n \in (325, 350)$, the JASP value approximately exceeds 3. Since values greater than 3 are generally

considered reliable evidence, when $n > 325$ and $r_i > 0.57n$, no further JASP verification is needed—the results can be considered strong evidence. For $n \leq 325$, however, JASP or R software verification is recommended to ensure accuracy.

Finally, this solution relies on several preconditions, such as the unbiased interval estimation assuming $n = 200$ and JASP's default option of ">Test value." If these assumptions are modified, the algorithmic framework should remain applicable. We leave further solution development and model refinement to interested readers.

References

Yici Zhang et al. (2000). *Probability Theory and Mathematical Statistics*. Beijing: Science and Technology Press, 2000.

Jietai Hou, Zhonglin Wen, Zijuan Cheng (2004). *Structural Equation Modeling and Its Applications*. Beijing: Educational Science Press, 2004.

Fan Wu, Quan Gu, Zhuanghua Shi, et al. (2018). Escaping the trap of traditional hypothesis testing methods: The application of Bayesian factors in psychological research. *Applied Psychology*, 2018, 24(03).

Chuanpeng Hu, Xiangzhen Kong, Eric-Jan Wagenmakers, et al. (2018). Bayesian factor and its implementation in JASP. *Advances in Psychological Science*, https://doi.org/10.3724/SP.J.1042.2018.00951

Note: Figure translations are in progress. See original paper for figures.