Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 3 (2024)

Do Asians Live Longer?

Zhengchao Ying

Master of Philosophy in Economics ,Department of Economics ,Universitetet I Oslo

Abstract: The latest scientific research paper widely applied the NHST (Null Hypothesis Significance Testing) method, which has caused considerable controversy in practice, especially in the fields of biology and social sciences, where numerous critical articles have been published. This article demonstrates the drawbacks of solely applying the NHST method through a case study on Asian lifespan, while also providing suggestions for algorithm improvement.

Keywords: NHST, Traditional Chinese Medicine, Asia, Data Science.

Introduction

In China, there is an ancient saying: 'If it's a mule or a horse, it will show when you take it for a walk.' The scientific field widely applies the NHST (Null Hypothesis Significance Testing) method in confirmatory statistics to test hypotheses. However, the use of p-values in the NHST method as a basis for hypothesis testing has sparked extensive controversy among researchers. In fact, confirmatory statistics offer various algorithms, aside from the widely used NHST hypothesis testing method in experimental domains. There are also regression algorithms, maximum likelihood algorithms, and the currently popular Bayesian updating algorithm in Europe and the Americas. This article will compare the NHST method and the maximum likelihood method through an example, elucidating the intentions behind these algorithms and how they should be reasonably applied to achieve the goal of hypothesis testing for theories.

Discussion

First, let's start with a brief overview of the history of criticism against NHST. Then, we will compare the NHST algorithm and the maximum likelihood algorithm using a case study, and finally, we will provide an assessment of these two algorithms.

Literature Review

The NHST (Null Hypothesis Significance Testing) method in confirmatory statistics was first introduced by Fisher in 1921. He applied analysis of variance in his research on wheat production, laying the foundation for the NHST algorithm that we use today. However, as Keren (1993) pointed out, the NHST method we currently use is actually a combination of Fisher's method and Neyman and E. S. Pearson's methods.

In the experimental domain of physical chemistry, this NHST method has been widely adopted to test the correctness of theories. For instance, Mendenhall & Sincich (2015) authored a classic statistics textbook. However, criticism against the NHST method began to emerge in the social sciences, primarily in the 1950s. Notable articles include those by Jones (1952), Rozeboom (1960), Bakan (1966), and Lykken (1968), among others. This movement gained momentum over the following decades, as seen in the works of Cohen (1994) and Harlow et al. (2016).

Cohen's work, especially, with over two thousand citations, sparked a significant wave of criticism in the 1990s. Abelson (1995) shifted the focus to goodness-of-fit indices instead of NHST's significance tests. Subsequent social science research has increasingly relied on regression algorithms and emphasized goodness-of-fit indices like R-squared, as recommended by Byrne (2013) in the SEM (Structural Equation Modeling) method.

This issue has also stirred widespread controversy in the field of biology. NHST was applied to food research in biology starting in the 1940s, as seen in the works of Byrne (2013) and Eheart & Sholes (1946). However, in the early 1970s, researchers began to criticize the statistical testing process, as exemplified by Edwards (1976), drawing scholars' attention to the issue.In 1988, the International Committee of Medical Journal Editors recognized the issue and started to require confidence intervals to replace NHST and p-values. The American

Statistical Association also issued a statement, declaring that NHST and p-values should not be used for scientific and business decision-making."

case

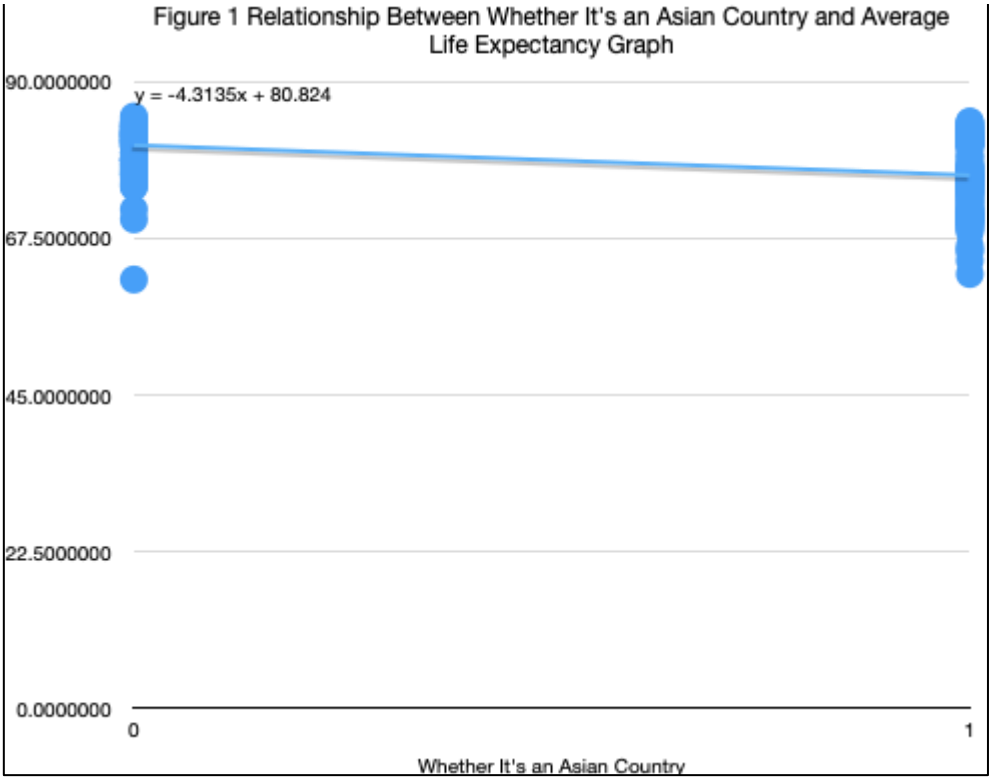
I randomly selected data for the average life expectancy, country names, and per capita GDP from 32 countries on the World Bank website from 1990 to 2018.

Table1 Average Life Expectancy and Whether it is an Asian Country, along with Per Capita GDP Data Table

Year	Country Name	Life Span	Asia or Not	GDP per capita
1990	Luxembourg	77.3804878	0	34645.14324
1990	Switzerland	77.296	0	38428.3855
1990	Norway	77.24243902	0	28242.94374
1990	Ireland	76.83756098	0	14048.1062
1990	Iceland	76.97073171	0	25423.07201
1990	USA	78.03634146	0	23888.60001
1990	Sweden	76.60731707	0	30593.67245
1990	Holand	76.53731707	0	21290.86038
1990	Austria	76.6	0	21680.98969
1990	Finland	71.59756098	0	28364.64508
1990	Canada	75.01041463	0	21448.36196
1990	New Zealand	74.80909756	0	13663.02162
1990	France	75.37804878	0	21793.84265
1990	Italy	76.93902439	0	20825.78421
1990	Spain	61.529	0	13804.8768
1990	Greece	75.21463415	0	9600.18513
…	…	…	…	…

Of course, our initial assumption was that Asian countries have traditional Chinese medicine (TCM), so life expectancy should be longer. A simple model assumes that life expectancy is determined by whether a country is in Asia. We plotted these two columns of data on a graph.

We added a trendline to the scatterplot. From the graph, it is evident that the average life expectancy in Asian countries is actually lower than that in non-Asian countries. So, modeling life expectancy based on the factor of being an Asian country and applying the frequentist statistical method of least squares regression, we can establish a predictive model that concludes that TCM may decrease life expectancy. The coefficient estimates that the Asian country factor would lead to a decrease in average life expectancy by 4.31 years and is highly statistically significant, with a t-statistic of -9.43 and a p-value of 0.0000. If you apply the principles of mathematical statistics you learned earlier, you would probably consider the theory that TCM leads to a decrease in life expectancy to be an absolute truth (of course, this is just an illustrative example and doesn't account for more complex factors like time series data).



Is the real pattern indeed like this? Unfortunately, this theory is, of course, absurd.

So, let's take a look at the definition of inferential statistics: Inferential statistics is the statistical method for studying how to use sample data to infer characteristics of the population.

Therefore, the hypothesis testing in mathematical statistics mentioned above is asking if the theory is valid after being passed? The meanings of these two sentences are completely different. According to the definition, passing a hypothesis test indicates that a theory that holds on the sample also holds in the population.

Now, the question arises: Does the theory that TCM reduces life expectancy hold on the sample? NHST doesn't address this. So, you need to learn a bit about the maximum likelihood method to answer this question.

In simple terms, the idea behind the maximum likelihood method is that if your theoretical model is correct, then real data should follow a normal distribution around your theoretical predictions, with only measurement errors present. After presetting a few parameters, you can calculate the likelihood of each data point, and by multiplying these likelihoods, you obtain the likelihood of the model being true. Then, you search for the parameters that maximize this likelihood, which gives you the parameter model with the maximum likelihood.

Returning to the two columns of data, I used OxMetrics software for maximum likelihood estimation, and the results are as shown in the previous figure, with a coefficient of -4.31, which is the same as calculated by the least squares method. So, from a parameter estimation perspective, the two methods are consistent.

However, the advantage of the maximum likelihood method is that while we find the parameters with the maximum likelihood as mentioned above, we also obtain a statistical measure of how likely the model is. This is not something that can be obtained under NHST.

This indicator in the maximum likelihood method is called the log-likelihood value, which is essentially the logarithm of the likelihood calculated above. Since the logarithmic function is monotonic, the parameter model with the maximum likelihood will naturally have the maximum log-likelihood value.

OxMetrics reports the log-likelihood value of this model as -903.98. So, what seemed like a perfect model under the least squares method, almost considered an absolute truth worthy of publication in international journals, shows significant issues under the maximum likelihood method. This is because the log-likelihood value or probability of this model is too low!

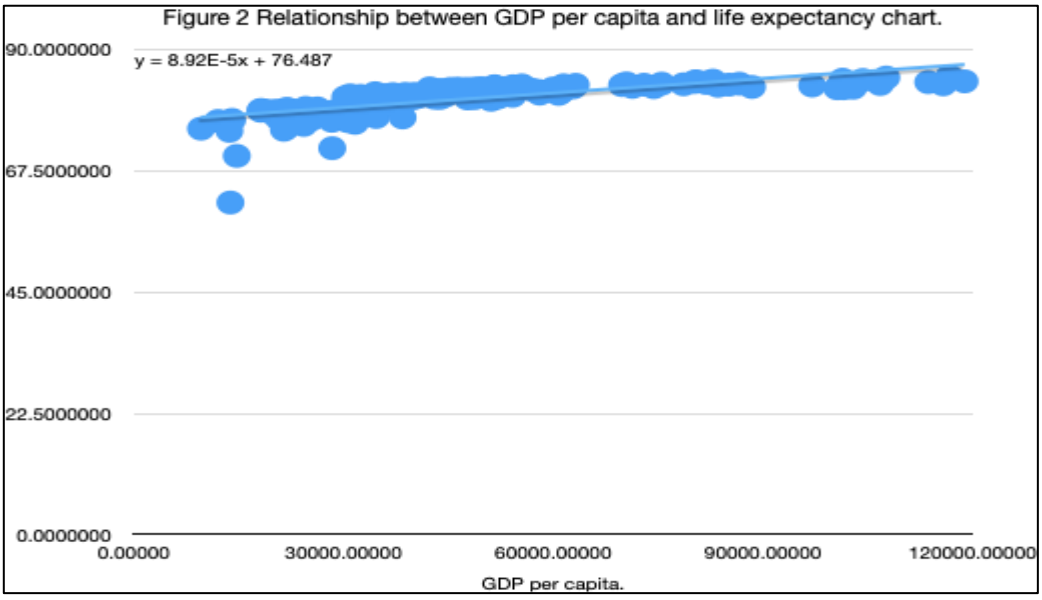
From the graph, it's also evident that this model is completely baseless. For example, the model predicts that the average life expectancy for non-Asian countries is 81 years, but the actual average life expectancy in Spain, a non-Asian country, is only 61 years. Anyone with common sense would ask, 'Why?' Is your theory really correct? Can a theory with such a large prediction error be considered true?

So, using NHST hypothesis testing alone cannot confirm the validity of a theory; we also need to consider whether the log-likelihood value reaches a reasonable level. If the log-likelihood value is high, it indicates a high likelihood of the theory being valid for that dataset. If the data is randomly sampled, you can even say that the theory holds in the population with a certain probability through likelihood ratio tests.

Using the maximum likelihood method, we can see that the aforementioned theory that TCM reduces life expectancy is absurd.

So, how do we improve our model? We need to consider the third column, which is the per capita GDP data, because economic factors clearly influence daily nutritional intake, which is, of course, one of the determining factors for average life expectancy.

All we have to do is plot a scatterplot of the per capita GDP data and the two columns of average life expectancy data, and it's easy to see this. From the graph below, we can observe that as per capita GDP increases, average life expectancy also increases, and this correlation is quite evident.



So, in reality, average life expectancy is determined by two factors. One is the economic factor, represented by per capita GDP (of course, because the impact of per capita GDP on life expectancy diminishes beyond a certain point, taking the logarithm is more appropriate). The other is the Asian factor, as the prevalence of traditional Chinese medicine (TCM) is relatively higher in Asia.

algorithm evaluation

Because from a parameter estimation perspective, maximum likelihood and NHST are consistent, OxMetrics software actually uses the least squares method internally. This algorithm is more efficient for computers and easier for machines to handle.

If you have learned the maximum likelihood method, you will get much better results this time. The log-likelihood value is -614.74. Physics and chemistry is highly predictable, resulting in very high log-likelihood values. However, in the fields of biology and social sciences, such high values are generally not achievable, so -614.74 is acceptable. This means that the theoretical model is valid for the sample.The conclusion of this theory is completely opposite to the previous one. The coefficient for the Asian factor is 0.93, meaning that the average life expectancy in Asian countries is actually one year higher than in non-Asian countries. The coefficient for the logarithm of per capita GDP is 3.98, indicating that for every 1% increase in per capita GDP, average life expectancy increases by 3.98 years. Both coefficients are significant, with the former having a t-statistic of 4.10 and the latter 40.2. The p-values for both coefficients are below 0.001. So, according to the perspective of NHST, the theory that traditional Chinese medicine (TCM) increases average life expectancy is almost an absolute truth.

However, just by introducing another variable, the effect of TCM on life expectancy completely changes direction. The theories of TCM increasing and decreasing average life expectancy cannot both be true. Therefore, the sole use of NHST has serious limitations.

So now that we've learned about the maximum likelihood method, we understand that the first theory cannot pass the log-likelihood test, whereas the second theory can. Therefore, the first theory is not valid in the sample, while the second theory is valid in the sample. Since the second theory also passed the NHST test, assuming random sampling, it not only holds in the sample but can also be extended to the population with a 99.9% confidence level (again, this is just an illustrative example and does not consider more complex factors like time series data).

You might wonder why this serious flaw in NHST hasn't been discovered in the research of physics and chemistry. That's because the factors in physics and chemistry are relatively simple and most experiments are controlled. So, those theories are valid in the sample, and the main consideration is whether they can be extended to the population to form scientific theories, mainly using NHST method.

However, factors in biology and the social sciences are very complex, and controlled experiments can easily miss a variable or rely on observational data where it's not even known how the dependent variable is determined by the independent variables. Therefore, the log-likelihood value from the maximum likelihood method becomes a necessary prerequisite indicator.

Conclusion

In summary, it is clear that there are significant drawbacks to using NHST on its own, which is why it has faced criticism from many researchers. The original intention behind NHST was to probabilistically extend the parameters from a sample to the population and it does not inherently possess the capability for model evaluation. However, due to the complexity of the technology and its evolution, NHST, like a game of telephone, gradually acquired the role of model evaluation, leading to significant protests. Therefore, the maximum likelihood method provides a more comprehensive approach to confirmatory statistics, where the log-likelihood value can be used to evaluate the model, and likelihood ratio tests can provide p-values to determine whether sample parameters can be extended to the population.

Of course, there is room for further improvement in the maximum likelihood method. For instance, the log-likelihood value decreases as the number of observations increases, making it challenging to evaluate which model is better when sample sizes are different. Comparison can only be made within the same batch of sample data. To address this, the log-likelihood value can be divided by the number of observations to calculate the average log-

likelihood value. This way, there is an intuitive numerical value to indicate what level of average log-likelihood value is acceptable.

Works Cited

[1] Abelson, R. P. (1995). *Statistics as Principled Argument*. Psychology Press.
[2] Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437.
[3] Byrne, B. M. (2013). *Structural Equation Modeling with Mplus: Basic Concepts, Applications, and Programming*. Routledge.
[4] Cohen, J. (1994). The earth is round (p < .05). *The American Psychologist*, *49*(12), 997–1003.
[5] Edwards, A. W. F. (1976). *Likelihood: An account of the statistical concept of likelihood and its application to scientific inference*.
[6] Eheart, M. S., & Sholes, M. L. (1946). Effects of method of blanching and temperature of storage on nutritive value of dehydrated cabbage. *Food Research*, *11*(4), 298–304.
[7] Fisher, R. A. (1921). Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. *The Journal of Agricultural Science*, *11*(2), 107–135.
[8] Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (2016). *What If There Were No Significance Tests?*
[9] Jones, L. V. (1952). Tests of hypotheses: one-sided vs. two-sided alternatives. *Psychological Bulletin*, *49*(1), 43–46.
[10] Keren, G. (1993). *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*. Lawrence Erlbaum Assoc Incorporated.
[11] Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*(3), 151–159.
[12] Mendenhall, W. M., & Sincich, T. L. (2015). *Statistics for Engineering and the Sciences, Sixth Edition*. Chapman and Hall/CRC.
[13] Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.