# Statistical Lessons Learned in Personal Informatics Research

Zilu Liang
liang.zilu@kuas.ac.jp
Kyoto University of Advanced Science (KUAS)
Kyoto, Japan

## Abstract

Personal informatics, a field that blends human-computer interaction (HCI) and health research, has seen significant growth. This paper reflects on key lessons learned from analyzing personal data and offers insights to enhance the rigor of statistical practices within this research community. Three core lessons are discussed: (1) the distinction between ordinal data and interval data, especially with Likert scales; (2) the importance of considering data structure, such as nested data in longitudinal studies, and the necessity of using multilevel analysis; and (3) the pitfalls of relying on the p-value. These reflections aim to spark a broader discussion on the adoption of better statistical practices in personal informatics. By fostering statistical rigor, the field can advance towards more meaningful conclusions that can benefit individuals in their pursuit of self-knowledge and well-being.

**ACM Reference Format:**
Zilu Liang. 2025. Statistical Lessons Learned in Personal Informatics Research. In *Proceedings of (CHI '25 Workshop on Envisioning the Future of Interactive Health).* ACM, New York, NY, USA, 3 pages.

## 1 Introduction

Personal informatics has emerged as one of the most popular topics in the HCI + health community [13, 14], with its motto of "self-knowledge through numbers" [22] resonating deeply with many researchers (myself included). As personal informatics researchers, we are fortunate to have an abundance of analytical methods at our disposal to make sense of continuously collected personal health data. It has been a thrilling journey learning from various disciplines like statistics, psychometrics, and fundamental science. What I have learned has come from analyzing my own data collected with early generations of Fitbit wristbands, from digging through literature, and yes, from those painful yet valuable journal and conference reviews. In this paper, I hope to share some lessons I have learned and spark reflections on how we, as an emerging research community, can improve the rigor of our statistical practices.

## 2 Three Lessons I Have Learned (Among Others)

### 2.1 Not All "Numbers" Are Numbers

Psychometric questionnaires, particularly the Likert scale, have been widely used in personal informatics research. Simple yet versatile, these scales are often employed to capture subjective perceptions, such as how much participants agree with statements like "Fitbit sleep data reflect my true sleep quality." These scales commonly feature five levels (e.g., "strongly agree," "agree," "neutral," "disagree," and "strongly disagree"), which are coded with symbols from 1 to 5, giving the appearance of natural numbers.

However, this numerical representation can be misleading. It is tempting to compute means, standard deviations, or perform t-tests on these 'numbers'. Yet, in psychometrics, individual Likert scale items are not treated as numbers. The distance between 2 (which denotes "disagree") and 3 ("neutral") is not necessarily equal to the distance between 3 ("neutral") and 4 ("agree"). This is a key distinction: Likert scales provide ordinal data, while natural numbers represent interval data.

This subtle difference has major implications for data analysis. For instance, calculating an average Likert score, such as 4.45, feels odd—what does this number actually signify? Is it 1.1% more "agree" than 4.40? The decimal place lacks meaningful interpretation and could be better rounded down (as Cohen advocates in the "less is more" principle) [9].

Despite this, there is a silver lining. When multiple ordinal Likert items are aggregated into a scale, the Central Limit Theorem comes to the rescue. Although the data remain ordinal, the summed scores can approximate a normal distribution, thus allowing us to apply traditional statistical methods [8]. This is why well-validated psychometric instruments often combine multiple Likert items to measure underlying constructs.

Given these considerations, I now resist the temptation to calculate means and standard deviations for individual Likert items. Instead, I use the median and quartiles, which are more appropriate for ordinal data [8, 17]. I have also developed the habit of inspecting data distributions visually before performing any statistical analysis. In my experience, a histogram or other graphical representation of Likert scale data can often reveal far more than summary statistics alone [16].

### 2.2 Many Datasets Are Nested

The rise of wearable devices, such as Fitbit, Apple Watch and Oura Ring, has made it easier and more cost-effective to collect large volumes of physiological, behavioral, and contextual data [7, 19]. With this new opportunity comes a shift from traditional cross-sectional studies to more sophisticated longitudinal designs, where multiple data points are collected from each participant over time.

However, this transition from "1-of-N" to "N-of-1" [12, 21] (and eventually "N-of-N") necessitates a reevaluation of the statistical analysis techniques we apply. While many personal informatics studies naturally adopt repeated measures designs, researchers often mistakenly apply statistical methods intended for cross-sectional data, which can lead to erroneous conclusions.

To illustrate this, in one of my previous studies, I tested three different correlation analysis techniques on a nested dataset: (1)


*CHI '25 Workshop on Envisioning the Future of Interactive Health, Yokohama, Japan*

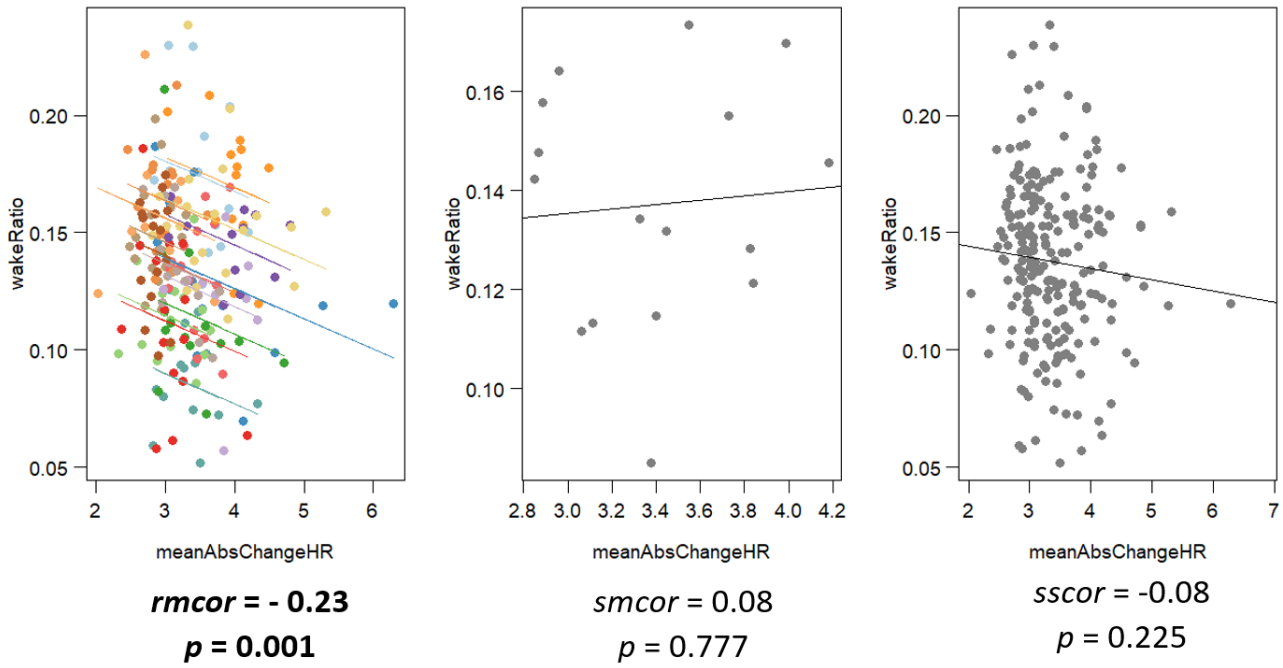


**Figure 1: Correlation analysis between the mean absolute change in heart rate and wake ratio using a nested dataset comprising 229 days of sleep, steps, and heart rate data from 16 participants using Fitbit wristbands [18]. Left: The repeated measure correlation method correctly accounts for within-subject variance, revealing a weak negative correlation with a coefficient of -0.23. Middle: When the dataset was flattened by averaging each participant's data, and Pearson's correlation was applied, no correlation was found. This averaging masked valuable insights. Right: Ignoring the nested structure and applying Pearson's correlation directly to the raw data also resulted in no correlation. This example highlights how improper statistical techniques and the failure to consider data structure can lead to misleading results.**

Pearson's correlation on individual data points (*sscor*), (2) Pearson's correlation after averaging each subject's data (*smcor*), and (3) repeated measures correlation (*rmcor*) [4]. The results were revealing: *sscor* and *smcor* might have shown correlations where none existed (or vice versa), whereas the *rmcor* technique, which properly accounted for the nested data structure, provided more accurate insights [18]. This reinforced the importance of accounting for the nested structure in the dataset, as ignoring it can lead to faulty conclusions.

The key takeaway here is that we must account for within-subject variance when analyzing repeated measures data. Instead of removing this variance through averaging, which can obscure meaningful patterns, multilevel analysis techniques should be employed (e.g., mixed-effects models [3]).

The challenges become even more pronounced when applying machine learning to nested datasets. Many machine learning algorithms (tree-based models being notable exceptions [15]) assume that data points are independently and identically distributed (i.i.d), a fundamental assumption often violated in nested data structures. While recent deep learning techniques, such as long short-term memory (LSTM) neural networks [23], can accommodate temporal correlations in longitudinal data, they still fall short in capturing both within-person and interpersonal variances. This highlights the need for new algorithms that integrate multilevel analysis principles with machine learning techniques, a direction that future personal informatics studies should pursue.

## 2.3 The *P*-Value Is Often Misinterpreted

Another important lesson that has come to the forefront of my research is the growing criticism of the *p*-value and its role in statistical analysis. Like many researchers, I relied heavily on the 0.05 threshold to determine statistical significance without fully appreciating its limitations. I initially believed that if $p > 0.05$, we could safely conclude that the null hypothesis was true.

However, I recently learned that all I could conclude from $p > 0.05$ was that I could not reject the null hypothesis; in other words, I could "*hardly conclude anything*" [9] at all.

The over-reliance on the *p*-value has contributed to the replication crisis in science and is now considered one of the major pitfalls in modern statistical practice [10]. The backlash against the *p*-value has been ongoing for decades, perhaps ever since the concept was invented [9]. But it was not until very recently, that consensus is finally emerging across fields such as psychology [11], statistics [5], and beyond [2]. Multiple publications in *Nature* have advocated for redefining [6] or even abandoning the idea of statistical significance altogether [1, 2, 20].

Rather than discarding the p-value entirely, I now interpret it with caution. A better approach, supported by the statistical community, is to report effect sizes and confidence intervals [2, 11]. These metrics provide a more nuanced understanding of the data and help mitigate the risks of relying solely on *p*-values when drawing conclusions.

## 3 Staying Current With Statistical Practices

Personal informatics is a rapidly evolving field that draws heavily from more traditional research disciplines. However, this "methodological mimicry" can be a double-edged sword: while it opens up new application avenues for statistical techniques, it also carries the risk of perpetuating outdated or incorrect statistical practices. The pace at which new methods are adopted can be slow, particularly in fields like HCI, where researchers may not always keep up with the latest developments in statistical analysis.

I have personally encountered this challenge. While my training in informatics provided me with strong analytical skills, it left me with gaps in statistical knowledge, and I have continued to learn and grow in this area. As the field of psychology, for example, undergoes significant reforms to improve statistical practices, I encourage personal informatics researchers to collaborate and share updated best practices. By doing so, we can develop guidelines that will help newcomers in the field. By embracing better statistical practices, we can improve the quality of our research and ensure that we use the best possible methods to derive meaningful conclusions from the data we collect. Incorporating statistical rigor into our work can not only enhance the credibility of our findings but also help advance the field of personal informatics in ways that benefit individuals.

## References


[1] Valentin Amrhein and Sander Greenland. 2018. Remove, rather than redefine, statistical significance. *Nature Human Behavior* 2, 4 (2018), 4.

[2] Valentin Amrhein, Sander Greenland, and Blake McShane. 2019. Scientists rise up against statistical significance. *Nature* 567 (2019), 305–307.

[3] R.H. Baayen, D.J. Davidson, and D.M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59, 4 (2008), 390–412. doi:10.1016/j.jml.2007.12.005 Special Issue: Emerging Data Analysis.

[4] Jonathan Bakdash and Laura Marusich. 2017. Repeated Measures Correlation. *Frontiers in Psychology* 8 (04 2017). doi:10.3389/fpsyg.2017.00456

[5] Monya Baker. 2016. Statisticians issue warning over misuse of P values. *Nature* 531 (2016), 151.

[6] Daniel Benjamin, James Berger, Magnus Johannesson, Brian Nosek, E.-J Wagenmakers, Richard Berk, Kenneth Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher Chambers, Merlise Clyde, Thomas Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, and Valen Johnson. 2018. Redefine Statistical Significance. *Nature Human Behaviour* 2 (2018), 6–10. doi:10.1038/s41562-017-0189-z

[7] Lauriane Bertrand, Nathan Cleyet-Marrel, and Zilu Liang. 2021. Recognizing Eating Activities in Free-Living Environment Using Consumer Wearable Devices. *Eng. Proc.* 6 (2021), 58. doi:10.3390/I3S2021Dresden-10141

[8] James Carifio and Rocco J. Perla. 2007. Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. *Journal of Social Sciences* 3, 3 (2007), 106–116.

[9] Jacob Cohen. 1990. Things I have learned (so far). *American Psychologist* 45, 12 (1990), 1304–1312. doi:10.1037/0003-066X.45.12.1304

[10] Lincoln J. Colling and Dénes Szűcs. 2021. Statistical Inference and the Replication Crisis. *Review of Philosophy and Psychology* 12 (2021), 121–147. doi:10.1007/s13164-018-0421-4

[11] Geoff Cumming, Fiona Fidler, Martine Leonard, Pavel Kalinowski, Ashton Christiansen, Anita Kleinig, Jessica Lo, Natalie McMenamin, and Sarah Wilson. 2007. Statistical Reform in Psychology: Is Anything Changing? *Psychological Science* 18, 3 (2007), 230–232.

[12] Nediyana Daskalova, Karthik Desingh, Alexandra Papoutsaki, Diane Schulze, Han Sha, and Jeff Huang. 2017. Lessons Learned from Two Cohorts of Personal Informatics Self-Experiments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 46 (Sept. 2017), 22 pages. doi:10.1145/3130911

[13] Daniel A. Epstein, Clara Caldeira, Mayara Costa Figueiredo, Xi Lu, Lucas M. Silva, Lucretia Williams, Jong Ho Lee, Qingyang Li, Simran Ahuja, Qiuer Chen, Payam Dowlatyari, Craig Hilby, Sazeda Sultana, Elizabeth V. Eikey, and Yunan Chen. 2020. Mapping and Taking Stock of the Personal Informatics Literature. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 126 (Dec. 2020), 38 pages. doi:10.1145/3432231

[14] Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. 2015. A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 731–742. doi:10.1145/2750858.2804250

[15] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on tabular data? arXiv:2207.08815 [cs.LG] https://arxiv.org/abs/2207.08815

[16] Richard Heiberger and Naomi Robbins. 2014. Design of Diverging Stacked Bar Charts for Likert Scales and Other Applications. *Journal of Statistical Software* 57, 5 (2014), 1–32. doi:10.18637/jss.v057.i05

[17] Malcolm Koo and Shih-Wei Yang. 2025. Likert-type scale. *Encyclopedia* 5, 1 (2025), 18.

[18] Zilu Liang. 2022. Correlation Analysis of Nested Consumer Health Data: A New Look at an Old Problem. In *2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech)*. 46–47. doi:10.1109/LifeTech53646.2022.9754805

[19] Zilu Liang, Bernd Ploderer, Wanyu Liu, Yukiko Nagata, James Bailey, Lars Kulik, and Yuxuan Li. 2016. SleepExplorer: a visualization tool to make sense of correlations between personal sleep data and contextual factors. *Personal Ubiquitous Comput.* 20, 6 (Nov. 2016), 985–1000. doi:10.1007/s00779-016-0960-6

[20] Blakeley B. McShane and Andrew Gelman. 2017. Abandon statistical significance. *Nature* 551 (2017), 558.

[21] Peter Molenaar. 2004. A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement: Interdisciplinary Research Perspective* 2 (10 2004), 201–218. doi:10.1207/s15366359mea0204_1

[22] Amon Rapp and Maurizio Tirassa and. 2017. Know Thyself: A Theory of the Self for Personal Informatics. *Human–Computer Interaction* 32, 5-6 (2017), 335–380. doi:10.1080/07370024.2017.1285704

[23] Sofia Yfantidou, Pavlos Sermpezis, Athena Vakali, and Ricardo Baeza-Yates. 2023. Uncovering Bias in Personal Informatics. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 139 (Sept. 2023), 30 pages. doi:10.1145/3610914