Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities

Visualizing disparity trends on felony sentence-imposed months by gender and race with generative AI

Yoshiyasu Takefuji

Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan

ARTICLE INFO

Keywords:
Felony sentence
Disparity trend
Disparity by gender and race

ABSTRACT

This study analyzes trends in felony sentence disparity based on gender and race from 2010 to 2024. It utilizes a generative AI to create Python code for data visualization and employs three statistical methods (ANOVA, Chi-Square, Fisher's Exact) to assess *p*-values. The p-value signifies the probability of random chance causing the observed association. A significance level of 0.05 is used as a benchmark. The evidence-based analysis reveals a concerning trend: increasing disparities in sentences across genders and races. The findings highlight the need for further research and policy changes to address these disparities in the criminal justice system. The paper offers a novel visualization approach to depict these trends, aiding comprehension of the issue.

1. Introduction

This study presents a comprehensive analysis of disparity trends in felony sentences, focusing on gender and race, over a decade from 2010 to 2020. The analysis is based on a federal dataset released on May 29, 2024, which comprises 24,676 instances and 28 variables. The visualization of these trends was facilitated by the use of a generative AI, which produced Python code to assist in the data representation. This innovative approach allowed for a more nuanced understanding of the data. The examination of disparity or bias trends was conducted using three statistical methods: Analysis of Variance (ANOVA) (Chatzi & Doody, 2023), Chi-Square Test (Zheng et al., 2023), and Fisher's Exact Test (Heston, 2023). These methods were chosen for their ability to justify individual *p*-values (Boscardin et al., 2024), providing a robust framework for the evidence-based analysis.

This annually updated dataset (DC.GOV, 2024), sourced from the open data of the District of Columbia, encompasses all felony sentences from 2010 onwards. It incorporates demographic details of the offenders, including gender, race, and age. Additionally, it provides sentencing data, such as the nature of the offense, the group of offense severity, and the specifics of the sentence imposed, including its type and duration.

In scientific research, we often seek to understand relationships between variables. Hypothesis testing provides a framework for evaluating these relationships statistically. A key component of this framework is the *p*-value (Boscardin et al., 2024), which plays a critical role in assessing the evidence against the null hypothesis of no association between two sets of data. This paper investigates the *p*-value between sentence-imposed months and gender and that between sentence-imposed months and race.

The *p*-value represents the probability of observing a result at least as extreme as the one obtained, assuming the null hypothesis is true. In simpler terms, it tells us how likely it is to see the observed association purely by chance, if there's truly no underlying connection between the variables. However, the *p*-value alone doesn't provide a definitive answer. We need a benchmark to judge the significance of this probability. This is where the significance level, often denoted by the Greek letter alpha (α), comes in. The most widely used significance level is $\alpha = 0.05$. The choice of 0.05 reflects a balance between requiring sufficient evidence for an association and minimizing the risk of making a Type I error. A Type I error occurs when we reject the null hypothesis (concluding an association exists) when it's actually true. Setting a lower significance level (e.g., 0.01) would make it harder to reject the null hypothesis, demanding stronger evidence. Conversely, a higher level (e.g., 0.1) would increase the chance of false positives (rejecting the null hypothesis when it's true). The 0.05 level strikes a balance, ensuring we have enough evidence while minimizing the risk of Type I errors. Historically, $\alpha = 0.05$ has been adopted as a standard significance level across various scientific disciplines. This allows for easier comparison of results from different studies that use the same benchmark for judging the significance of *p*-values. It's important to remember that 0.05 is not an absolute threshold. Depending on the research context, this value can be adjusted. For instance, fields like medical research, where the consequences of errors are high, might use a stricter significance level like

E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.cities.2025.105767
Received 23 September 2024; Received in revised form 24 November 2024; Accepted 26 January 2025

0.01.

If the *p*-value is lower than the significance level, we have strong evidence to reject the null hypothesis. This suggests that the observed association is unlikely due to chance and there's evidence for a genuine relationship between the variables. Conversely, a high *p*-value indicates that the observed association could be due to random chance. In this case, we fail to reject the null hypothesis and conclude that there's not enough evidence to claim a significant association. However, it's crucial to remember that the *p*-value itself doesn't tell us how strong or in which direction the association lies. It only indicates the likelihood of observing such an association by chance.

This study meticulously computes and vividly illustrates the progression of *p*-values over a span of years, utilizing 0.05 merely as a benchmark for reference. It's important to note that the choice of 0.05 is not indicative of a hard threshold, but rather serves as a conventional marker in statistical analysis for indicating the probability of falsely rejecting the null hypothesis. The graph generated as a result of this calculation provides a dynamic visual representation of the changes in the significance of evidence over time. This allows for an intuitive understanding of the temporal fluctuations in the data, highlighting periods of significant change. Therefore, not only does this paper offer a quantitative analysis, but it also provides a qualitative perspective on the evolution of the data's statistical significance. This dual approach facilitates a comprehensive understanding of the subject matter.

The results of this study reveal a concerning trend: disparities in felony sentences based on gender and race have been generally increasing over the examined period. This finding underscores the importance of continued research and policy efforts to address these disparities in the criminal justice system. Further studies are needed to delve deeper into the causes and potential solutions for this issue.

2. Methods

Generative AI is employed to craft Python code capable of visualizing a graph. This graph comprises six lines, each representing three distinct tests (ANOVA, Chi-Square Test, and Fisher's method). Each test incorporates two variables such as race and gender for the calculation of *p*-values, serving as scientific evidence. The query input to the AI system is crucial in generating the correct Python code. However, generative AI is not flawless. It often requires multiple interactions to produce the desired outcome. The process is iterative, with each conversation refining the output. The text that follows is an initial query directed to the generative AI, specifically using Microsoft's Copilot. This query serves as the starting point for the iterative process of code generation. Note that users should be familiar with variables in the dataset.

The dataset, Felony_Sentences.csv, contains 24,676 instances and 28 variables collected over a decade from 2010 to 2020. In this paper, we use the variable 'SENTENCE_YEAR' as the X-axis. The target variable 'SENTENCE_IMPOSED_MONTHS' represents the Y-axis, while 'GENDER' (x1) and 'RACE' (x2) serve as additional explanatory variables, resulting in the relationship expressed as y = f(x1, x2). Our objective is to analyze the associations between 'GENDER' and 'SENTENCE_IMPOSED_MONTHS' (x1 and y) as well as between 'RACE' and 'SENTENCE_IMPOSED_MONTHS' (x2 and y).

2.1. Initial query

show Python code to visualize a graph of 6 black lines for 'GENDER' and 'RACE' by 3 methods such as ANOVA, Chi-Square, and combined two test (Fisher's method) with 4 line styles and 2 line widths (1,2) using "Felony_Sentences.csv". calculate *p*-value between 'SENTENCE_IMPOSED_MONTHS' and 'GENDER' and p-value between 'SENTENCE_IMPOSED_MONTHS' and 'RACE'. The graph have a total of 6 lines for 3 tests, each test having two categories such as gender and race. Y-axis indicates *p*-value as significance of evidence while 'SENTENCE_YEAR' indicates X-axis. Rotate X-axis labels with 90 degrees. Plot 6 black lines and Y-axis label as *p*-value with 0.05 horizontal line for reference. Locate 6 legend box outside and under the graph.

3. Results

The final version of the code, `felony.py`, is readily accessible on our GitHub repository (GitHub, 2024). This program is responsible for generating the results depicted in Fig. 1. It computes and displays the average duration of sentences imposed, broken down by gender and race. The following table provides a detailed overview of these disparity averages by gender and race:

The results from the Table 1 show the average duration of sentences imposed, categorized by gender and race. Males receive significantly longer sentences on average compared to females, with males averaging 37.31 units and females averaging 19.03 units. When examining race, there is considerable variation in sentence durations. Asians receive an average sentence duration of 17.29 units, Blacks 35.55 units, Hispanics 29.61 units, Native Americans 26.78 units, those categorized as "Other or Unknown" 39.59 units, Pacific Islanders 6.00 units, and Whites 38.62 units. These results highlight disparities in sentencing based on gender and race, which could indicate underlying biases or systemic issues within the judicial system.

Figs. 1 and 2 indicate that while the results differ slightly between evaluation methods such as ANOVA, Chi-Square, and Fisher's method, similar trends are observed. However, disparities by gender and race are evident in both figures.

Upon selecting "Black" and "White" in the 'RACE' category, a significant difference becomes evident in Fig. 2 compared to Fig. 1. This figure underscores a more pronounced racial disparity, highlighting the stark contrast between these two racial groups in the context of sentencing. The visual representation in Fig. 2 serves as a powerful tool for understanding the extent of this disparity.

4. Discussion

In their comprehensive report, Shaw and colleagues presented compelling evidence that white defendants, once convicted, were often subjected to more severe and prolonged sentences compared to their Hispanic or Black counterparts (Shaw & Lee, 2019). This racial disparity in sentencing was further corroborated by the findings of Blankenship et al., who provided additional support for these observed discrepancies (Blankenship et al., 2018).

Moreover, the issue of disparity extends beyond race. Ciocanel et al. shed light on gender disparities within Federal Criminal Cases, indicating that sentencing practices also vary significantly based on the defendant's gender (Ciocanel et al., 2020).

Furthermore, Wen et al. expanded the scope of this discussion to the juvenile justice system, reporting racial disparities in youth pretrial detention (Wen et al., 2023). Their findings underscore the pervasive nature of these disparities, affecting not only adults but also young individuals in the early stages of their interaction with the justice system.

These studies collectively highlight the multifaceted nature of disparities within the criminal justice system, emphasizing the need for continued research and reform in these areas. They serve as a stark reminder that the pursuit of justice must be accompanied by a commitment to equality and fairness.

Lehmann et al. studies that inequalities in criminal punishment based on race, ethnicity, gender, and socioeconomic status challenge the justice system's moral foundations (Lehmann & Gomez, 2022). Research over the past 50 years, especially recently, showed minority defendants, males, and lower socioeconomic individuals face harsher penalties. Their chapter reviewed literature on judicial decision-making and explores conditions that exacerbate these inequalities. It highlighted implicit biases among justice actors and suggests policy reforms to promote a more equitable punishment system (Lehmann & Gomez, 2022).

Phillips reported that many studies have examined the impact of

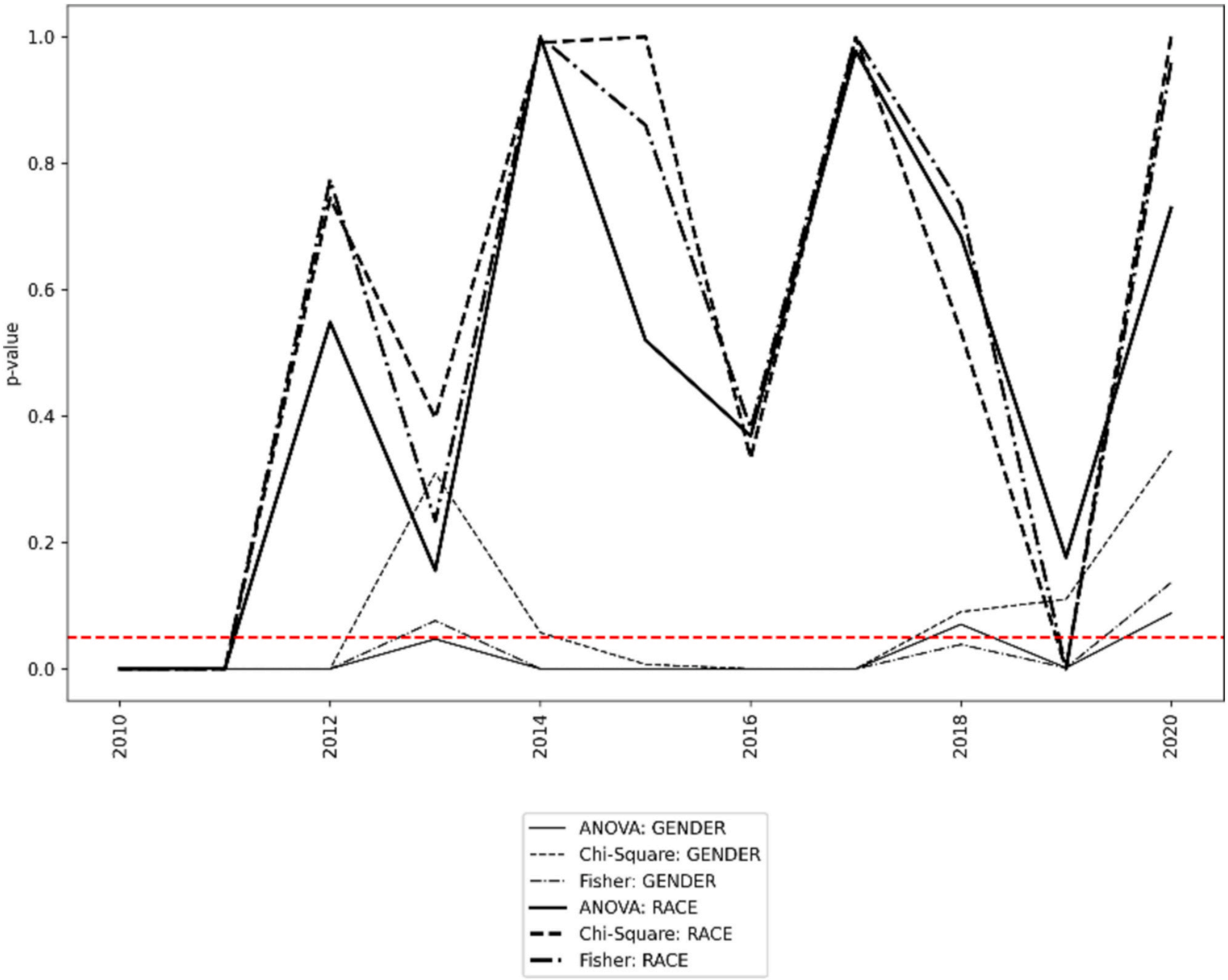

Fig. 1. disparity trends by gender and race on sentence-imposed months.

Table 1
Disparities on felony sentence-imposed average months by gender and race.

Gender	Imposed months	Instances
Female	19.3	1893
Male	37.31	22,556
Race		
Asian	17.29	17
Black	35.55	22,487
Hispanic	29.61	54
Native American	26.78	9
Other or Unknown	39.59	1196
Pacific Islander	6.00	1
White	38.62	912

victim race and gender in capital punishment, but victim social status has been largely overlooked (Phillips, 2009). This research analyzed 504 capital murder cases in Harris County, Texas, from 1992 to 1999 and found that social status significantly influences the district attorney's choice to seek the death penalty and the jury's decisions. High-status victims, perceived as integrated, sophisticated, and respectable, were more likely to have the death penalty sought or imposed (Phillips, 2009).

Over the past two decades, researchers have highlighted the deterring influence of religion on crime-related attitudes and behaviors (Adamczyk et al., 2017). However, limited work has assessed the overall state of this research. Their study addressed this gap by systematically reviewing empirical journal articles from 2004 to 2014, analyzing qualitative and quantitative studies. It highlighted prevalent theoretical perspectives, strengths and weaknesses of existing research, and offers directions for future studies in this area (Adamczyk et al., 2017).

This paper pioneers a novel approach by utilizing generative AI to visualize trends in sentencing disparities over time, specifically focusing on variations by gender and race. This innovative approach facilitates a nuanced understanding of the data by depicting the changing significance levels (*p*-values) of the relationships between sentence length and both gender and race across the years. The visualizations not only reveal the existence of disparities but also provide insights into how these disparities may be evolving over time.

The results demonstrate a substantial gender disparity in sentence length. Males receive sentences on average nearly twice as long as females (37.31 months vs. 19.03 months). This stark contrast underscores a concerning level of gender bias within the system.

Racial disparities are also evident, with a distinct pattern observed in the average sentence lengths imposed. Sentences are typically longest for White offenders, followed by Hispanic, Native American, Asian, and Pacific Islander offenders. These findings suggest a complex interplay of race and ethnicity in sentencing outcomes. Further research is necessary to delve deeper into the underlying causes of these disparities.

By shedding light on these trends, this paper contributes valuable insights to the ongoing discourse on equality and fairness in the justice

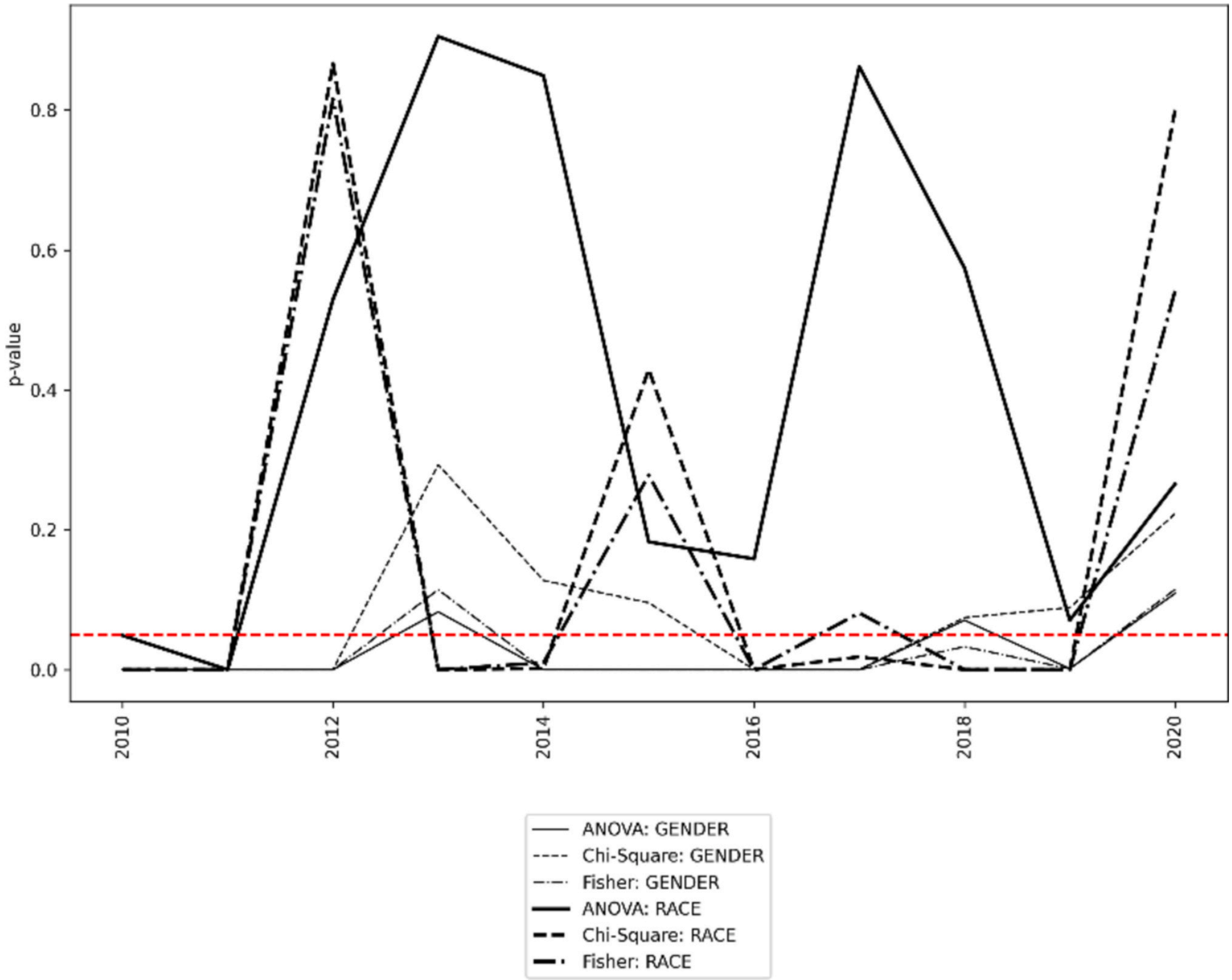

Fig. 2. disparity trends by gender and race on sentence-imposed months with filtered of 'Black' and 'White'.

system. The visualizations serve as a powerful tool for understanding and communicating these complex issues to a broader audience. These findings highlight the critical need for continued scrutiny of sentencing practices to ensure they are free from bias and discrimination. Future research efforts should explore the root causes of these disparities and develop evidence-based interventions to promote a more just and equitable criminal justice system.

CRediT authorship contribution statement

Yoshiyasu Takefuji: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Funding

This research has no fund.

Declaration of competing interest

The author has no conflict of interest.

Data availability

The authors do not have permission to share data.

References

Adamczyk, A., Freilich, J. D., & Kim, C. (2017). Religion and crime: A systematic review and assessment of next steps. *Sociology of Religion, 78*(2), 192–232. http://www.jstor.org/stable/44654404.

Blankenship, K. M., Del Rio Gonzalez, A. M., Keene, D. E., Groves, A. K., & Rosenberg, A. P. (2018). Mass incarceration, race inequality, and health: Expanding concepts and assessing impacts on well-being. *Social Science & Medicine, 1982*(215), 45–52. https://doi.org/10.1016/j.socscimed.2018.08.042

Boscardin, C. K., Sewell, J. L., Tolsgaard, M. G., & Pusic, M. V. (2024). How to use and report on p-values. *Perspectives on Medical Education, 13*(1), 250–254. https://doi.org/10.5334/pme.1324

Chatzi, A., & Doody, O. (2023). The one-way ANOVA test explained. *Nurse Researcher, 31*(3), 8–14. https://doi.org/10.7748/nr.2023.e1885

Ciocanel, M. V., Topaz, C. M., Santorella, R., Sen, S., Smith, C. M., & Hufstetler, A. (2020). JUSTFAIR: Judicial System Transparency through Federal Archive Inferred Records. *PLoS One, 15*(10), Article e0241381. https://doi.org/10.1371/journal.pone.0241381

DC.GOV. (2024). Felony sentences. Accessed on July 23. https://opendata.dc.gov/datasets/DCGIS::felony-sentences/explore.

GitHub. (2024). felony.py for examining disparity trends by gender and race. Accessed on July 23. https://github.com/y-takefuji/felony.

Heston, T. F. (2023). Statistical significance versus clinical relevance: A head-to-head comparison of the Fragility Index and Relative Risk Index. *Cureus, 15*(10), Article e47741. https://doi.org/10.7759/cureus.47741

Lehmann, P. S., & Gomez, A. I. (2022). Racial, ethnic, gender, and economic sentencing disparity. In E. Jeglic, & C. Calkins (Eds.), *Handbook of issues in criminal justice reform in the United States*. Cham: Springer. https://doi.org/10.1007/978-3-030-77565-0_8.

Phillips, S. (2009). Status disparities in the Capital of Capital Punishment. *Law & Society Review, 43*(4), 807–838. https://doi.org/10.1111/j.1540-5893.2009.00389.x

Shaw, J., & Lee, H. (2019). Race and the criminal justice system response to sexual assault: A systematic review. *American Journal of Community Psychology, 64*(1–2), 255–276. https://doi.org/10.1002/ajcp.12334

Wen, A., Gubner, N. R., Garrison, M. M., & Walker, S. C. (2023). Racial disparities in youth pretrial detention: A retrospective cohort study grounded in critical race theory. *Health & justice, 11*(1), 14. https://doi.org/10.1186/s40352-022-00203-8

Zheng, Y., Mao, Y., Tsao, M., & Cowen, L. L. E. (2023). Minimum chi-square method for estimating population size in capture-recapture experiments. *PLoS One, 18*(10), Article e0292622. https://doi.org/10.1371/journal.pone.0292622