To: Professor Fiona Doherty

From: Linh Pham

Date: March 30, 2023

Re: Judge Sentencing Disparities Literature Review and Summary

To fully understand the literature on criminal sentencing disparities by judge, this review presents a summary of some of the most relevant pieces. Furthermore, it provides a deep technical understanding of the technical nuances involved in presenting claims regarding criminal sentencing in the United States. This review is structured in the following:

* Introduction
* Key Definitions
* Methodology of Anderson, Kling, and Stith
  + Discussion and importance of distributions
* Summary Table of Literature
* Conclusions and Considerations

**Introduction**

In addition to studying nationwide gender and racial disparities in sentencing, it may be useful to examine how these dynamics vary across judges and districts. Existing studies of variation in sentencing disparities across judges and districts tend to focus on two themes. First, judges are not the sole agent determining sentencing, but only one part of a system that involves many actors. Second, regional differences, including in the underlying demographics, seem to exert a strong influence on sentencing outcomes. One example of regional variation is border states’ (i.e. Texas, Arizona, etc) “fast-tracked” processes, implemented to handle the high volume of drug trafficking cases.[[1]](#footnote-1) Furthermore, sentencing outcomes depend on many factors, including the quality of available public defenders, caseload, and judicial ideology. One strong finding observed throughout the literature was that judges who sat pre-*Booker* continued to uphold pre-sentencing reform minimums even after these guidelines became advisory.

**Key definitions:**

* Negative binomial model - Models the number of failures before success occurs; in the case of the study which uses it below, “success” is a sentence length of zero, and they only model everything else.

In probability, there are three “levels” to refer to probability. I put levels in quotation marks because this is not a formal way to introduce the topic. Furthermore, the terms I use below are not formally used and may not be technically correct. However, for the purpose of this article, it would be good to standardize what I refer to.

“SET” [[2]](#footnote-2)– A set in this case, will refer to the total amount of outcomes a single probability can predict. In the case of both a binomial and negative binomial model, there are only two outcomes. Ex. A coin has heads and tails

“TRIAL” – A trial refers to the number of sets. For example, 10 trials refers to coin flips, and each coin flip has a set of two outcomes.

“DISTRIBUTION” – Describes how values are distributed across a field. For example, what is the probability of getting 1 head in 10 coin flips? 2 heads? 3 heads? A distribution often looks like a little bar graph where each point corresponds to the probability of getting all the values on the x axis, in this case, the probability of getting 1- 10 heads.

“MODEL” – a mathematical measure that makes statistical assumptions of a set of data. For example, in a binomial model, the model assumes that there are only two outcomes, and there is a mathematical formula to calculate the probability.[[3]](#footnote-3)

A binomial model measures two outcomes. For example, in a coin flip, the probability of heads or the probability of tails. The negative binomial model also measures two outcomes, but it is the inverse of the binomial model. The number of trials is fixed and the model measures the number of successes given a fixed number of trials (Ex. what is the probability of getting heads in 20 flips?). I’ve also created an example below which explains the negative binomial model in the context of sentencing data. In this case, “success” is when an offender has a sentence length of zero and failure is everything else.

|  |  |
| --- | --- |
| Offender | Sentence Length (in months) |
| A | 0 |
| B | 1 |
| C | 2 |
| D | 0 |
| E | 0 |
| F | 3 |

For a negative binomial model, we would ask “what is the probability of 80,000 offenders with a sentence length > 0 in a dataset of xx offenders?”

A negative binomial model is most different from a binomial model when there are fewer trials in total. Ultimately both are similar if there are enough trials to measure data for.

To understand this better in an actual context, see below in the section titled “Methodology of Anderson, Kling, and Stith.”

* Parametric model - any model that captures and describes its predictions solely using a finite set of parameters, typically control variables (as opposed to more complex models that don’t specify or limit the amount of parameters). This is a really large umbrella term to describe any probability distribution. For example, a negative binomial model is also a parametric model. The case of non-parametric models is mostly used in machine learning models, but for the purposes of this report, all the models used or mentioned are parametric models.
* Hierarchical model - Hierarchical models are models in which there is some sort of hierarchical structure to the parameters. A hierarchical structure allows greater isolation of variables of interest and allows comparison between variables without as much concern for overlap.
* ANOVA - a statistical formula that compares the variance across the mean of groups[[4]](#footnote-4); it is used to predict a continuous outcome (i.e. sentence length which can have any positive, real number that can be a whole number or a decimal) on the basis of categorical variables (i.e. gender, etc.). It is a part of the linear regression family thus a lot of studies will refer to it as a linear regression model. However, this is slightly incorrect because a true linear regression model can only use continuous variables to predict a continuous outcome.

**Methodology of Anderson, Kling, and Stith**

In the case of Anderson, Kling, and Stith, they used this to measure the proportion of offenders with zero sentence length and non-zero sentence length to see if there was an unusual amount of offenders with zero sentence length in a dataset of this size. This was also to acknowledge that the data they had taken was a sample and not a true dataset of the entire criminal justice system. They were able to determine that their sample was a good representation using this method as well. This method was mostly used to ascertain the validity of their sample dataset, and then the mathematical models used this knowledge as a base.

The following is a brief summary of how assigning a probability model goes into a larger aspect of Bayesian probability. They used this pre-existing information as a prior, created a likelihood probability function, which created a statistical prediction for the outcome. Then, they compared how their predicted outcomes compared to the actual outcomes after the Guidelines.

In the case of a binomial model, I assume that the question would have been something like the following:

“in a dataset of 100,000 offenders, what is the probability of getting xx offenders with a sentence length >0?” xx refers to any number of offenders.

The main difference between a negative binomial model and binomial model is that a negative binomial model can continuously update no matter the size of the dataset. A negative binomial model will continue find samples until the desired amount of “successes” are achieved. In this example, a negative binomial is used to measure the total amount of offenders until we have about 80,000 offenders with a sentence length of zero. The binomial model solely focuses on the successes,

In summary, a binomial model would have discarded all values with 0 since a value of 0 is not a success. However the negative binomial model keeps those values included until it reaches its desired amount of offenders with a value greater than 0.

Referencing the research paper, Anderson, Kling, and Stith do not discuss in depth why they chose a negative binomial model over others. Specifically, this is the only line found regarding this decision:

“A zero-inflated negative binomial model is developed in order to account explicitly for the fact that many cases end in dismissal or acquittal or have a sentence that involves no prison time at all” (pg 4).

Tiede matched offenders based on facts before and after the guidelines were approved to get around this issue. Smith (2021) simply removed all offenders with a sentence length of zero.

There were several questions that arose because of comparing distributions. Namely, why are distributions important? What are distributions? And how do changing distributions change the final result? The following section answers these questions.

Distributions are mathematical functions which show all the possible values for a variable and explain how they occur. They are typically used as the basis for predicting how data will act. For example, a mathematical model using a normal distribution assumes that most data will center around the average with smaller amounts of data at the far outreaches (reference the popular bell shape distribution). From there, the model can perform mathematical transformations to predict a certain result.

In this case, it is popular to say “because we applied x model and the result was y, y mean z.” Certainly, if the data follows all the assumptions the mathematical models relies on, then the conclusion is correct. However, if distributions vary, then the model should mathematically adjust for different types of data. Models preprogrammed into languages like R often have defaults distributions and require specifically specifying alternative distributions. Anderson, Kling, and Stith created their own model and assumed that the same data had a different distribution than most other pieces of research.

It is important to note that most literature examining the effects of the criminal sentencing guidelines such as Booker, etc. do not discuss the concept of distributions. Distributions are a minute aspect of mathematical models, and often, many social scientists can research and use mathematical models with no prior understanding of distributions.

**Summary Table of Literature**

|  |  |  |
| --- | --- | --- |
| **Piece** | **Methodology Notes** | **Main Summary and Notes** |
| Anderson, Kling, Stith. Measuring Inter-Judge Sentencing Disparity Before and After the Federal Sentencing Guidelines. 1999. | Used advanced statistics to develop models which measure the extent of variation in sentence lengths across judges. The outcome of interest is therefore relative differences in the average sentence length for otherwise comparable caseloads of defendants assigned to different judges in the same district.  The math is complex but incorporates a negative binomial model (similar to a regular linear regression but models the probability of a binary outcome) to capture the chance that a judge sentences a defendant to any prison time at all. Parameters are added for judge and time, and the mean sentence is allowed to vary by district, in effect controlling for differences between districts. The model compares mean sentences for each judge to the mean for the district, and then captures the difference in magnitude of this disparity before and after the Guidelines. The study ultimately creates a statistical model which compares mean sentences for judges with the distribution of sentence lengths overall.  By using this method, they could compare the statistical likelihood of inter-judge sentencing means and disparities versus the actual numbers to give an overall account of inter-judge disparity before and after the Guidelines. | Their method is unique because it relies on a series of mathematical models.  Concluded that inter-judge disparity is less pronounced after Guidelines were implemented in 1984 than before.  Changes in inter-judge disparity are not due to changes in the types of offense in the overall caseload.  Decrease in inter-judge disparity is concentrated within the violent, weapons, and drug crimes. |
| Tiede. The Impact of the Federal Sentencing Guidelines and Reform: A Comparative Analysis. 2009 | Tiede takes issue with an older Sentencing Commission model which uses hierarchical models to compare variables. Instead, Tiede proposes using regression models on the individual variables and then comparing those regression models.  Furthermore, Tiede matches cases by facts (judge, sentence length, criminal case, etc.) before and after the Guidelines to minimize confounders and try to achieve a valid comparison. Then, Tiede applies regression modeling. Tiede calls this Fact Pattern selection.  Tiede used exclusively drug distribution cases from USSC individual offender data from 1999-2006, comprising 1,112 total cases.  Simple regression analysis did not yield any effects of legal regorm on sentence length, but ANOVA showed that some of the district courts operating in some of the circuits reacted differently to the legal changes depending on the time frame analyzed. The specificized legal changes most likely references the guidelines, but Tiede uses terms like ‘legal reforms’ and ‘legal changes’ throughout the piece.  The literature review focuses on three time periods: the PROTECT Act, Blakely v. Washington, and United States V. Booker (pg. 36). However, analysis was done on only two time periods: pre-PROTECT Act and post-Blakely. Tiede noted they did not include data in between due to insufficient data (pg. 40). | One of Tiede’s main topic points is the USSC’s use of hierarchical models in a 2004 review of 15 years of Guidelines sentencing.[[5]](#footnote-5) He disagreed with USSC’s claim that inter-judge and regional disparities did not account for sentence length differences.  Concluded that judges’ choices to depart from Guidelines had significant impact on sentence length.  Concluded that regional disparity drives inter-judge disparity. For example:  Judges in the 9th circuit departed significantly more than its counterparts. When they did so, inter-judge disparity was less pronounced (pg 39).  Judges sentence higher sentences with less variance when they apply Guideline ranges. Judges also use departures less frequently when applying Guideline-table ranges. When judges depart, sentences are lower and have higher variance (pg 39).  Before the Protect, Blakely, and Booker guidelines, judges in the 9th and 10th circuits had higher sentencing disparity than other circuits. Post-Blakely, the 9th and 10th’s sentencing disparity increased. (contrasted to the 2nd circuit which had the same level as the 9th and 10th pre-Protect but sentencing disparity decreased post-Blakely) Concluded that 9th and 10th were more independent and less perceptive to the Guidelines for sentencing (pg 42). |
| Surprising Judge-to-Judge Variations Documented in Federal Sentencing. TRAC Reports.  <https://trac.syr.edu/tracreports/judge/274/> (March 5, 2012) | This report used data from FY2007 -FY2011 from Transactional Records Access Clearinghouse (TRAC).  It only examined drug and white collar cases because those had the largest discrepancies. Appeared that judges which had large sentencing disparities for drug cases also had the same for white collar cases  TRAC used bar charts and tables as their methodology. The simplicity offers valuable information that is easy to interpret but lacks depth about the multivariate nature of sentencing  Excluded cases handled by magistrate judges or special judges sitting by designation. Also excluded judges who sentenced less than 50 total defendants. In total, there were 885 district judges with 372,232 defendants in the five years. The average number of defendants per judge was 420. | “Long term efforts to improve the consistency of the federal sentences through the adoption of complex sentencing guidelines have not been entirely successful in curtailing large judge-to-judge differences in sentencing practices”  Smaller courts showed more similar sentences than larger courts and therefore smaller sentencing disparities between judges within the district as a whole.  Northern district of Texas had the largest discrepancy in sentencing (100 months difference in between medians of highest sentencing judge and lowest sentencing judge). Northern district of Georgia and Eastern district of Virginia followed with 90 months of difference.  Hypothesized pre-existing predilections of the judge assigned were the cause for wide ranges in drug sentencing cases |
| Schanzenbach, Max M. “Racial Disparities, Judge Characteristics, and Standards of Review in Sentencing.” *Journal of Institutional and Theoretical Economics (JITE) / Zeitschrift Für Die Gesamte Staatswissenschaft*, vol. 171, no. 1, 2015, pp. 27–47. *JSTOR*, http://www.jstor.org/stable/24549079. Accessed 2 Mar. 2023. | Schanzenbach interacted judge and offender characteristics to test for unexplained racial bias; specifically, ANOVA to predict the effect of unexplained racial bias on sentencing outcomes (p. 32).  Arizona, Southern California, New Mexico, south and west Texas face uniquely large caseloads. Drug trafficking cases are quickly passed through in “fast track” departures. There are “substantial discounts” available in these border districts, and many cases are conducted without trials in this area. Thus, Schanzenbach excluded these areas due to their unique circumstance. As a result, the significance level of political variables generally increased, but it did not change the qualitative conclusions. (pg 34-35)  Used Guidelines[[6]](#footnote-6) data from 1992-2009. Limits to the following categories: murder, manslaughter, sex crimes, robbery (primarily bank robbery), drug trafficking, firearms offenses, racketeering, arson, and auto theft (55% of all prison sentences). Did this so that there were fewer opportunities for judges to depart or provide alternative prison sentences. | Democrats and Republicans sentence differently from each other, with Democratic appointees more lenient than Republican appointees. However, racial disparities do not vary by political party. Moreover, Black, and Hispanic judges do not sentence differently from their white counterparts.  He found that changes in standards of review did not affect the interaction between judge characteristics and racial disparities and therefore concluded that the Guidelines did not affect racial disparities  Noted that judicial ideology as proxied by the party of the president who appointed the judge, “is an important predictor of judicial decisions” → also that other judicial characteristics such as race, sex, and professional background are significant predictors of judicial decisions in circumstances in which the characteristic itself is relevant such as racial or sex discrimination.[[7]](#footnote-7) |
| Daly, Kathleen, and Michael Tonry. “Gender, Race, and Sentencing.” *Crime and Justice*, vol. 22, 1997, pp. 201–52. *JSTOR*, http://www.jstor.org/stable/1147574. Accessed 2 Mar. 2023. | Examined data from Prosecution of Felony Arrests, Offender- Based Transaction Statistics (OBTS), the National Judicial Reporting Program. This piece is mostly a literature review spanning across statistical review and other public analysis on the issue of gender and race in sentencing. | Suggested that gender embeddedness in offense variables may be even more pronounced. “Specifically, the sources of variation in the character and content of men’s and women’s offenses and their criminal histories may be especially poorly measured.” For example, men were higher ranked in drug dealing operations  Largest increases for female sentences were in drug sales with aggravated assault and robbery following.  More imprisoned women were in for drug offenses than were men, noted a drastic increase in gap between women and men in this regard five years earlier |
| Yang, Crystal S. “Free at Last? Judicial Discretion and Racial Disparities in Federal Sentencing.” *The Journal of Legal Studies*, vol. 44, no. 1, 2015, pp. 75–111. *JSTOR*, https://doi.org/10.1086/680989. Accessed 8 Mar. 2023.[[8]](#footnote-8) | Yang constructed her own dataset using data from the USSC (FY1994-2010), Transactional Records Access Clearinghouse (TRAC), and the Federal Judicial Center.  -From the USSC data, Yang excluded cases with missing criminal records and race. Matching USSC to TRACT retained 60% of data. The final dataset included FY2000-2010 and totaled 381,361 cases.  Used ANOVA to predict sentencing outcomes. | Concluded that racial disparities after *Booker* were greater among judges appointed after *Booker*.  Controlled for many aspects such as education, non-random assignment of cases to judges , and many more. |
| Smith, C. M., et al. Racial Disparities in Criminal Sentencing Vary Considerably Across Federal Judges. SocArXiv, 29 July 2021, doi:10.31235/osf.io/j2gbn. | Utilized JustFair Data but also appended FY 2018 and 2019 data to that made available by JustFair. This caused an increase of 30,000 cases to the original 600,000. Excluded immigration cases in the southwestern states. Only examined cases after 2006 because they were only interested in data after Booker. Eliminated cases where sentence length was zero, resulting in removing 30,000 cases.  They used a hierarchical linear model of log transformed sentence lengths. | The average judge assigns Black defendants sentences that are 13% longer than white defendants.  In general, judges who existed 2 standard deviations from the average sentenced Black and Hispanic defendants much more than white defendants (almost 85-95% more).[[9]](#footnote-9)  Not focused on examining the impact of the guidelines, more so on the presence of disparities for specific judges. |
| Statement on Behalf of the Federal Community Defender Office  for the Eastern District of Pennsylvania  https://pae.fd.org/ | Noted that Justfair is a sample and does not include the entire case load of any one judge included. For some judges, may only have certain caseloads and thus resulted in extreme bias for specific judges.  Criticized for the lack of caution around sentence length of zero because these cases may include probationary and time served sentences (pg.2).[[10]](#footnote-10)  Also criticized the lack of acknowledgement towards other structural racism that affects a judges’ decision. | Primarily a response to Smith, et al (2021) above, criticizing the piece for being a flawed empirical “study,” specifically for naming Judges C. Darnell Jones II and Timothy J. Savage as among the most biased against racial minorities. Within the Eastern District of Pennsylvania, these judges hold high reputations among the legal community in Pennsylvania for fair and just sentencing. |

**Conclusions and Considerations**

* Departure. When reading, it appears that the most variance derived from the judge’s decision whether to depart or not, but variation from cases which lied solidly within a sentencing range did not change much after Booker guidelines.
* Hierarchical modeling vs Fact Pattern Matching: The only article which discusses hierarchical modeling is Tiede. Thus, it is a little confusing for both Charlotte and I. Tiede argues that the USSC uses hierarchical modeling and as a result, incorrectly concludes that judges do not have a role in sentencing disparity. Instead, Tiede proposes a “Fact Pattern matching” where Tiede continuously matches defendants until they are as similar as possible, removing those who differ. Specifically, one drug crime with the same fact pattern is analyzed. This process is similar to how Schanzenbach constructed their data set to only include sentences of serious crimes so that there would be few rows with no sentence length/alternative sentence length.
* Sentence Length of Zero/Minimal Sentencing Length: Anderson, Kling, Stith used a negative binomial distribution model to counteract this. Schanzenbach took out crimes where there was a possibility of having no sentence length or alternative sentence length. Smith removed cases with zero sentence length.
* Southwestern states and border states (parts of Texas, parts of California, Arizona, etc.) appear to be so varied that each paper which mentions these areas appear to drop this region entirely.

There are two questions that this review hopes to address:

1. How significant are differences in disparities in sentence lengths across judges? And specifically in terms of racial and demographic disparities, what does the evidence say about how large those differences are?
2. How did judge disparities change because of the Guidelines or Booker?

To answer the first question, all of the papers mentioned in this review discussed the presence of sentencing disparity due to racial bias. While some areas of the United States had smaller variance between their judges within a district, there were others which varied over a period of 100 months (TRAC, 2012). In terms of demographic disparities, certain regions appeared consistently as a cause of concern. Firstly, the southwestern states on the border between the US and Mexico were often excluded from datasets due to the systemic treatment of cases in that area. Smith, 2021 excluded immigration cases which specifically existed in this region. Schanzenbach excluded this area entirely. Tiede and TRAC explored this region and both discovered large sentencing disparity.

In regards to the second question, there was dispute about the Guidelines’ effect on sentencing disparity. Anderson, Kling, and Stith concluded that inter-judge sentencing disparity was less pronounced after the Guidelines. Tiede also concluded there was some impact; meaning most districts saw no change except the 2nd which decreased sentenced disparity and the 9th and 10th which increased after the Guidelines. TRAC and Schanzenbach concluded that there was no effect. Yang concluded that disparities after Booker increased.

1. Schanzenbach, Max M. “Racial Disparities, Judge Characteristics, and Standards of Review in Sentencing.” Journal of Institutional and Theoretical Economics (JITE) / Zeitschrift Für Die Gesamte Staatswissenschaft, vol. 171, no. 1, 2015, pp. 27–47. JSTOR, [↑](#footnote-ref-1)
2. A set refers to any collection of objects. It can be used in multiple different ways as well. Statistics has a bad habit of switching terminology with terms like “set” which are purposely vague in order to encompass a large umbrella of different possibilities. A lot of sentences may be phrased weirdly in order to avoid using set and trial in cases when I am not specifically referring to these definitions. [↑](#footnote-ref-2)
3. Model and distribution are also inter-changed frequently. In general, the distribution describes the attributes and characteristics of the data, where as the model calculates a prediction from these characteristics. [↑](#footnote-ref-3)
4. This is a technical definition of ANOVA. This means that first we calculate the mean of each group, whether it be the mean sentencing length, etc. Then we calculate how the means vary between each group. [↑](#footnote-ref-4)
5. USSC Appendix D, Section B Pg. D-3 <https://www.ussc.gov/sites/default/files/pdf/research-and-publications/research-projects-and-surveys/miscellaneous/15-year-study/appendix_D.pdf> [↑](#footnote-ref-5)
6. Assuming USSC Guideline data. [↑](#footnote-ref-6)
7. Their footnote had other resources, but some include (Cox and Miles 2008), (Peresie, 2005). (Schanzenbach, 2005) failed to find differences [↑](#footnote-ref-7)
8. There are many more conclusions to draw from Yang, but in the sake of time, there are not many in here. Charlotte’s literature review also discusses Yang’s article in depth. [↑](#footnote-ref-8)
9. Standard deviation is a measurement of variance. There is a mean value. Then, we usually add or subtract two standard deviations around the mean in order to describe 95% of the data. In this case, this means that judges who had sentencing averages on the upper end (top 5%) ended up sentencing Black and Hispanic defendants much more than white. [↑](#footnote-ref-9)
10. There wasn’t any more description available regarding this issue in the paper, but this matters because if an author is dropping all zero length case lengths, that means the only cases considered are ones which had sentencing. The study this response was addressing did not hold for things such as the individual case type. Thus, the implication is that judges who typically sentence probation for crime A would only be judged for sentences for Crime C and Crime D which are extreme cases. The amount of crime A cases this judge would oversee would be dropped entirely, making it appear as if the majority proportion of their case load involves incredibly high sentencing. While Smith 2021 does account for crime type, the specific nuances of each case are not accounted for. This is a broad generalization, but this is what I think this response is trying to imply. [↑](#footnote-ref-10)