Some of the math in the November 8, 2014 draft of "Challenging the Randomness of Panel Assignment in the Federal Courts of Appeals," including the bottom-line statistical analysis, is incorrect

Henry Corrigan-Gibbs* and Keith Winstein†

November 17, 2014

A recent working-paper draft examines the randomness of assignments of U.S. federal judges to three-judge appellate panels.[1] The working paper raises important concerns about the judicial system and has attracted considerable popular interest.[2] However, the draft makes some incorrect statements and relies on an unsound analysis. We briefly note these issues.

1 The bottom-line calculation is incorrect

The working paper investigates 144 separate null hypotheses, each true if a particular federal appeals court assigned judges to three-judge panels uniformly at random. Of these 144 hypotheses, the authors found evidence sufficient to reject 21 of the hypotheses at the $p < 0.1$ level. The authors aggregate these tests and summarize them:

> Indeed, we can say with over 95% confidence that we would not have discovered as much evidence of non-randomness as we did if all the circuits were actually using a random process.[3]
>
> In the last panel of Figure 5, we simulated the likelihood of finding 21 statistically significant results at the .10 level by chance when given 144 tries to do so. Our results show that we are at the far right end of the tail, and that there is a 95% probability that this many statistically significant results would not occur by chance alone. **In other words, if we had a coin weighted to land on heads 10% of the time and tails 90% of the time, and we flipped it 144 times, there is a less than 5% chance that the coin would land on heads 21 times.** We had 21 statistically significant results, which allows us to reject the possibility that our results are due to chance with over 95% confidence.[4]

However, the bolded statement is not true:

$$\sum_{k=21}^{144} \binom{144}{k} 0.1^k (1-0.1)^{144-k} \approx 0.0506695149 > 0.05$$

We do not claim a special status for the $p < 0.05$ threshold that has been conventional in some areas of statistical practice—only that the working paper's statements about "a less than 5% chance" and "over 95% confidence" are not accurate.[5]

*Ph.D. student, Stanford University Department of Computer Science

†Assistant Professor of Computer Science and, by courtesy, of Law, Stanford University Department of Computer Science and Stanford Law School

We thank Roger Ford, John Hawkinson, and the working paper's authors for feedback in response to an earlier version of this note.

[1] Adam S. Chilton & Marin K. Levy, Challenging the Randomness of Panel Assignments in the Federal Courts of Appeals (Nov. 8, 2014) (unpublished manuscript). The working-paper draft was previously available at http://ssrn.com/abstract=2520980 (as uploaded Nov. 11, 2014). On Nov. 12 and Nov. 14, 2014, we sent the authors earlier versions of this note, and on Nov. 15, 2014, the authors replaced the original working-paper draft on SSRN with an updated version that removes some of the statements that we found to be incorrect. The original draft is no longer available from SSRN. We can provide it to interested readers on request. All references and quotations in this note are from the original, Nov. 8, working-paper draft that was the subject of press coverage.

[2] Adam Liptak, *Coalition Challenges Selection of Judges in Same-Sex Marriage Case*, N.Y. TIMES, Nov. 11, 2014, http://www.nytimes.com/2014/11/11/us/politics/after-court-loss-opponents-of-same-sex-marriage-challenge-selection-of-judges.html.

[3] Chilton & Levy, *supra* note 1, at 5.

[4] *Id.* at 45–6 (emphasis added).

[5] Our prior work has found similar small errors that caused results to appear to cross a $p < 0.05$ threshold. Keith J. Winstein, *Boston Scientific Stent Study Flawed*, WALL ST. J., Aug. 14, 2008, at B1, with technical notes at http://cs.stanford.edu/~keithw/www/Winstein-14Aug2008-TechnicalNotes.pdf.

2 The statistical analysis is unsound

In footnote 129 of the working paper, the authors correctly note that the bottom-line analysis of the main text relies on the assumption that the 144 hypotheses are statistically independent.

As the footnote points out, this assumption is not true. For example, under the paper's generative process that fixes the total numbers of panels, the number of Third-Circuit panels with one Republican appointee is exactly determined by the number that have zero, two, or three Republican appointees.

The footnote states that the authors "make this simplifying assumption so that our discussion will be accessible to a wide range of readers," but we believe the matter is more fundamental and threatens the basic validity of the technique. If not accounted for, the effect of this dependence can be to exaggerate the significance of the results by "double-counting" anomalies.

For example, if we examine 72 independent quantities and find that 11 of them lie in pre-specified improbable regions (each with $p < 0.1$), the "overall" p-value by the authors' method is nonetheless greater than 0.1.[6] By contrast, if we duplicate each measurement, making for 144 quantities with 22 lying in pre-specified improbable regions, the overall p-value is less than 0.03—now an apparently "significant" result.[7]

Double-counting related hypotheses, as the working paper does, can therefore overstate the significance of results. There is a vast literature on multiple hypothesis testing that we do not claim to have mastered and do not attempt to summarize here. The issue is more complex than footnote 129 appears to acknowledge, because in addition to the interdependence between hypotheses related to panels with one vs. two Republican-appointed judges, there is also a dependence across different categories of test statistics. For example, in the First Circuit, a panel with a majority of Republican-appointed judges is also a majority-male one, because there are no female Republican-appointed judges on this court.

A further issue concerns the multiple-counting of panels within a circuit. The working paper acknowledges this issue:

> [I]n order to be consistent across circuits, we determined that a panel was defined by three-judges who sat to hear cases during a particular session on a particular day. Accordingly, if three judges were listed as hearing a set of cases at 9am and then later at 1pm, even on the same day, **we would count this grouping as two separate panels**. This approach seemed consistent with the interpretation of most circuits. However, some circuits had different measurement units. That is, in some circuits, panels were defined as units sitting together for a court week, and so Judges A, B, and C might be listed together for hearings on three consecutive days. **Again, to be consistent, we counted such sittings as three different panels.** Ideally, we would be able to identify what constituted a "draw" of a panel for each circuit. For example, some circuits may draw a panel and then have those judges sit together for three consecutive days, but others circuits may draw a panel and then have those judges sit together for just a morning. Since we were not able to reliably identify what constituted a draw consistently—both within circuits and across circuits—we made this judgment to identify panels in this way because we believed it to be both both justifiable based on our knowledge of circuits and objective.[8]

The concern with the bolded statements is that such practices will cause each "random sample" in that appeals court to be double- or triple-counted, which may lead to an excess of spurious findings and will make the p-values for that circuit incorrectly low.

The Ninth Circuit, for which the authors found the strongest apparent effect, appears to be such a circuit, where one panel sits together for multiple days. Depending on the degree of over-counting, this error alone could account for the working paper's entire findings.

[6] $\sum_{k=11}^{72} \binom{72}{k} 0.1^k (1-0.1)^{72-k} \approx 0.1018656057$

[7] $\sum_{k=22}^{144} \binom{144}{k} 0.1^k (1-0.1)^{144-k} \approx 0.02961445095$

[8] Chilton & Levy, *supra* note 1, at 23–4 (emphasis added).

3 The claimed finding of non-randomness in a majority of circuits is spurious

The working paper says:

> Our results show evidence of non-randomness in the majority of the federal courts of appeals.[9]
>
> Whatever the underlying causal mechanism, the point is still the same: the majority of the circuits appear to have panels that are different than what would be produced by a truly random process.[10]
>
> [T]here was evidence of non-randomness in seven of the twelve region circuits. More specifically, there is evidence of non-randomness in the D.C., Second, Third, Fourth, Fifth, Eighth, and Ninth Circuits.[11]

As the paper says, in 7 of the 12 geographic circuits, the authors were able to reject at least one hypothesis of uniform random assignment at a nominal p-value of less than 0.1. We believe this finding is entirely spurious, because *such a result would be expected even if all circuits used uniform random assignment and the working paper's reasoning were otherwise sound.*

Each of the 12 circuits was the subject of 12 separate hypothesis tests, where each test attempted to reject a null hypothesis along the lines of, e.g., "The event of a panel drawn with exactly two Republican appointees occurred with probability equal to what would have been expected under uniform random assignment."

Under the null hypothesis, a particular test will have up to a 10% probability of mistakenly rejecting the hypothesis at the $p < 0.1$ level. Following the working paper's reasoning in assuming that each hypothesis is statistically independent, then the probability that *any* of 12 hypothesis tests will produce a false rejection is as much as $1 - (1 - 0.1)^{12}$, or about 72%.

In other words, even when a circuit's panel assignments are perfectly random, this approach nonetheless makes it more likely than not that the circuit will be wrongfully identified as non-random. Across 12 separate circuits, still assuming the null hypothesis, the expected number of circuits that would be wrongfully fingered by this methodology is $12 \cdot (0.72)$, or more than 8 out of 12 on average.

In this light, the fact that the authors only found evidence for non-randomness in **7** of the 12 circuits is actually lower than would be expected from chance alone.[12]

[9] *Id.* at 1.

[10] *Id.* at 5.

[11] *Id.* at 42.

[12] Of course, this discrepancy is well within the range of the natural play of chance.