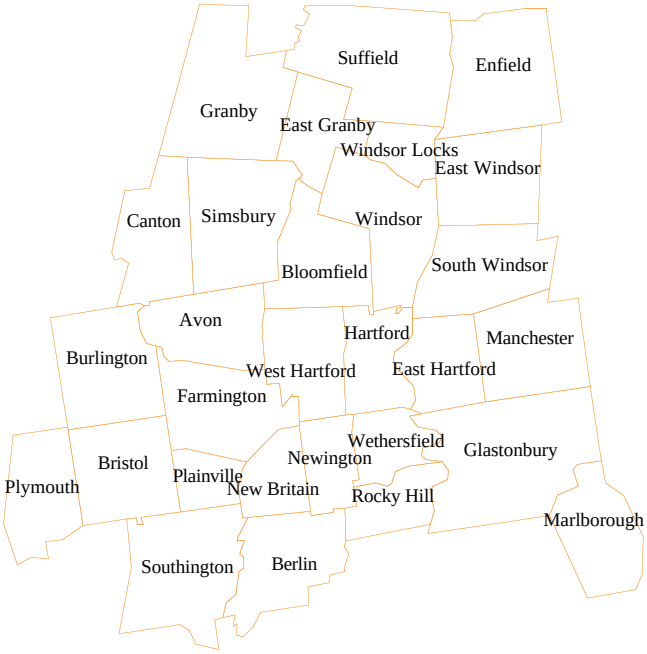
Final version 11 April 1997

Connecticut Juror Selection



Hartford-New Britain judicial district

Warning: The material regarding source lists is now out of date, because JIS changed its story on how the lists are used to create the master list. In particular, the material at the end of Section 1, and the whole of Section 7 and Appendix B need major revision.

Report to the
Connecticut Public Defender's Office

by

David Pollard
Statistics Department
Yale University
Box 208290 Yale Station
New Haven CT 06520-8290
Email: pollard@stat.yale.edu

Contents

Acknowledgements

I sought advice from several persons at JIS, from several employees of the Bureau of the Census, from numerous State and Federal officials, and from a number of subject-area experts.

I also learned a lot from my students: from Jason Cross, whose practical project first turned up many interesting leads; and from the students in a case studies graduate course—Karna Bryan, Andrew Carter, Laura McKinney, Brendan Murphy, Franklin Parlamis—who discovered many strange and surprising facts about juror selection and the statistics of the JIS data.

Catherine Sharkey, a student at the Yale Law School, suggested an interesting interpretation of the results of the geocoding and surname matching. She pointed out the weaknesses in several of my arguments.

Ann Green, of the Yale StatLab, and Jocelyn Tipton, of the Government Documents section of the Seeley Mudd Libarary at Yale, helped me to find and understand the appropriate Census data.

Tony Mein, Hartford Registrar of voters, explained some of the difficulties in maintaining the voter list. Frank De Luca and Rishi Nigum (I apologize if I have misspelled their names) answered many questions about the Hartford voter list, and took many pains in getting a backup copy of it to me.

Val and Al from the US Postal Service solved some problems involving zip-codes and street addresses.

Judge Margolis and Attorney Danaher provided data and answered many questions about the Federal juror selection procedures.

David Word, Marie Pees, Manuel de la Puente, and several other helpful folks at the Population Division of the Bureau of the Census answered many of my questions about the fine points of Census data.

Lloyd Mueller, of the Connecticut Public Health office, provided data and answered questions about population projections for Connecticut.

Marta Tienda, of the University of Chicago Sociology Department, gave a key piece of advice about surname matching.

Richard Gayer, Dana Lindner, Al Rogerson, and Lou Sapia, spent many hours discussing the JIS data and the JIS procedures with me. I appreciate their help and their patience. My study would have been impossible without their cooperation.

Tom Steahr made some informative demographic comments about the Penultimate draft of the report.

Tom Munsterman, of the National Center for State Courts, gave me valuable advice regarding the matching of sourcelists and the construction of master lists. He also referred me to his excellent little manual “Jury System Management” (Munsterman 1996). I wish I had read the manual before I had started this project. It would have saved me huge amounts of time.

My colleague Nicolas Hengartner read through a ‘final’ draft at short notice. His sharp eye rescued me from a few embarrassing inconsistencies.

Moira Buckley tracked down vital information regarding sourcelists and questionnaires, with great diplomacy and careful attention to detail.

I would also thank Mike Courtney and several other folks at the Public Defender’s Office for asking so many intelligent questions and providing valuable source material, except that it would just encourage them to think up more ways to keep me busy.

I dedicate this Report to my wife, Gai, for patience and understanding—several standard deviations beyond what one should expect from another human being—while I labored for too many months on the Report.

David Pollard

1. Introduction and summary

This report describes the results of my study of the Connecticut jury selection system, as it works in the Hartford-New Britain (HNB) judicial district. My main source materials were data from the Bureau of the Census, data obtained from questionnaires administered to jurors at several of the HNB courthouses, and summons records provided by *Judicial Information Systems* (JIS), a part of the Office of the Chief Court Administrator in the Connecticut Judicial Branch.

> *My testimony at the Rodriguez trial in January 1997 was based on the "Penultimate version" of this report. As an aid to any readers who are already familiar with that version, I have retained in the final version some material that I would ordinarily have edited out, but with added comments to clarify some points that arose during my testimony.*

Questionnaires

I first began my study at the beginning of 1996, in response to a request from the Public Defender's Office for some calculations related to juror questionnaires collected for the King trial in February 1996. A court order in March 1996 mandated the collection of a modified version of the questionnaire at all the Hartford-New Britain courthouses. Over 14,000 of the new questionnaires have been filled out by persons presenting themselves for jury service up until the end of January 1997.

> *The questionnaire data were still being collected while I was preparing the final version of the report. The report analyzes the data only up to early 1997 (mid-February). The final analysis will be submitted to the court as a separate document.*

A summary of the results from the questionnaires, and a discussion of problems related to undistributed and missing questionnaires, appears in Section 2. The general conclusion that I draw from the questionnaires is that Hispanics appear to be underrepresented:

- About 4.3% of the persons filling out the juror questionnaires indicated that they were Hispanic. This figure is significantly smaller than the 6.56% of the over-18 population of the Hartford-New Britain judicicial district counted as Hispanic in the 1990 Census. Moreover, the comparison with the 1990 figure probably understates the discrepancy: demographic projections suggest that Hispanics made up about 7.8% of the over-18 population of the judicial district by mid-1996.

As I explain in Section 3, with a sample size of over 14,000, the observed 4.3% Hispanic response cannot plausibly be explained away as a mere random fluctuation. The jurors filling out the questionnaires cannot reasonably be regarded as a simple random sample from a population of over 6.56% Hispanics. Indeed, there seems to be little disagreement on this point. The real question is: Can the disparity be accounted for by 'benign influences', that is, by mechanisms that the courts should regard as fair and as an expected consequence of selection methods prescribed by the Statute?[1]

To help answer the real question, I requested access to information and data related to the selection of jurors by the State of Connecticut. Starting not

[1] Connecticut General Statutes 1995, Title-51 Chapter-884. The Statute was modified in 1996, to require a wider collection of source lists if feasible.

long after the King trial, and continuing until February of 1997, I received from JIS large quantities of data (mostly in electronic form) related to the summoning system. The data came in a number of separate transfers. My report evolved as I learned more about the system and as more data became available. Several different working drafts of the report were given limited circulation, which led to some unfortunate confusion.

JIS data

The data contain records for all persons who were sent a juror summons (for any court in Connecticut) since the 1992-93 court year.[2] Amongst other information, the records show the name and address of each person summoned, together with various codes indicating whether the person was qualified to serve or was disqualified for some reason. The full list of possible disqualification codes is described at the end of Section 4. For the sake of brevity, in the main body of this report I have compressed the disqualifications into a smaller number of categories:

01 = not US citizen
06 = can't speak/understand English
08 = older than 70, chooses not to serve
12 = extreme hardship
13 = summons undeliverable
17 = standby notice/handbook notice or other undeliverable
NS = no-show
OK = confirmed for jury service
?? = disqualification status not yet determined
xjd = not in the judicial district
rest = all other types of disqualification

A *no-show* is a person who fails to serve, or be disqualified in some way, within one year of the date of summons to serve. Such a person might have deliberately ignored the summons or follow-up communications from JIS, or might not have received the summons in the first place.

I extracted from the data the records for summonses to the five courthouses in the HNB judicial district. Persons summoned to one of those courthouses were supposed to be residents of one of the 29 towns that make up the district.

The pattern of disqualifications is not uniform across the 29 towns. Two towns—Hartford and New Britain—stand out from the general pattern, as the following four tables show.[3] Each table corresponds to a different court year of summons: HNB9293 means court year 1992-93 for the HNB judicicial district, and so on. The four rows give the percentage breakdown by disqualification category for summonses sent to persons in Hartford town, New Britain town, towns (nonHNB) outside the judicial district, or (otherHNB) one of the other 27 towns that make up the district. The bottom rows (total) give the breakdown for all summonses to courthouses in the district.

[2] Actually, the data also contained a few months of records from the 1991-92 court year, which I set aside.

[3] More detailed figures appear in Section 8. Complete listings for this table, and of all other abbreviated tables, appear in the Appendixes of the report. These listings also include data for 1996-97.

HNB: DISQUALIFICATIONS BY TOWN GROUPINGS

[HNB9293]	01	06	08	12	13	17	NS	OK	xjd	rest	total
HARTFORD	4	5	8	2	27	3	12	35		5	100
NEW BRITAIN	5	5	15	3	15	1	6	44		5	100
nonHNB									97	2	100
otherHNB	2	1	12	5	9	1	3	60	1	7	100
total	2	2	11	4	12	1	5	52	5	6	100

[HNB9394]	01	06	08	12	13	17	NS	OK	xjd	rest	total
HARTFORD	3	5	8	2	31	3	10	33		4	100
NEW BRITAIN	4	5	15	3	17	2	6	42		7	100
nonHNB									98	2	100
otherHNB	2	1	12	5	10	1	3	57		8	100
total	2	2	11	4	13	1	4	49	5	7	100

[HNB9495]	01	06	08	12	13	17	NS	OK	xjd	rest	total
HARTFORD	3	5	7	3	36	4	10	27		4	100
NEW BRITAIN	4	6	17	5	16	2	5	37		7	100
nonHNB									98	2	100
otherHNB	2	1	13	8	10	1	3	53		8	100
total	2	2	11	6	14	2	4	45	6	7	100

[HNB9596]	01	06	08	12	13	17	??	NS	OK	xjd	rest	total
HARTFORD	3	5	8	2	28	5	8	5	31		4	100
NEW BRITAIN	4	6	17	5	15	2	6	2	37		6	100
nonHNB										98	2	100
otherHNB	2	1	13	7	10	1	5	2	51	1	8	100
total	2	2	12	6	13	2	5	2	44	6	7	100

It is striking that the undeliverable rates are consistently higher for Hartford (and, to a lesser extent, New Britain) than for the other towns in the district. Roughly 30% (more precisely, 27%, 31%, 36%, and 28%, for the four court years tabulated) of the summonses sent to an address in Hartford town are returned by the Postal Service as undeliverable (disqualification code 13). The figures for 1995-96 will change slightly when the disqualification status of each summons in the ?? category is finally determined.

The figures for HNB9697, which contains records for the first 4½ months of the 1996-97 court year, will be even more affected by the final resolution of the ?? category, which at present contains a mixture of undeliverables, no-shows, OK's, and other other disqualifications.

[HNB9697]	01	06	08	12	13	17	??	OK	xjd	rest	total
HARTFORD	3	4	6	1	22	3	39	18		3	100
NEW BRITAIN	4	5	13	2	11	2	37	21		5	100
nonHNB									98	2	100
otherHNB	2	1	11	3	6	1	42	27		6	100
total	2	2	10	2	9	1	40	25	4	6	100

From now on I will omit the partial results from 1996-97 from the summary listings, and refer the reader to the full counts in the Appendix.

According to the 1990 Census (see Section 6), Hartford and New Britain accounted for a large fraction of the minority population of the whole district: Hartford contained almost 60% of the Hispanic over-18 population, and almost 62% of the black over-18 population; New Britain contained more than 16% of the Hispanic over-18 population, and more than 6% of the black over-18 population. JIS should be aware of a problem:

- The two towns in the HNB judicial district that together account for a large proportion of the over-18 minority population have much the highest rates of undeliverable summonses and no-shows.

I would stress that the disqualification figures in the four tables come directly from cross-tabulations of the JIS data. They are not based on any statistical modelling.

Geocoding and Hispanic surname matching

The JIS data contain no explicit information about race or ethnicity of the persons to whom summonses are sent. To learn more about the effects of the various disqualifications (including undeliverable and no-shows) on the minority population, one must draw inferences based on the information that is contained in the JIS data.

The persons summoned are a random sample from a *master list* that JIS constructs each year. The various percentages presented below, therefore, all have interpretations as estimates of probabilities for persons on the master list being disqualified in various ways. For example, a figure of 30% for HNB summonses sent to Hispanics being returned by the Postal Service as undeliverable (code 13) corresponds to an estimate of 30% for the probability that *Hispanics who make it to the master list will be lost to the system by virtue of an undeliverable summons.*

I used two distinct methods of statistical inference. The first method is based on *geocoding*, that is, the location of addresses to within small regions—I chose to use the regions called *Census tracts*[4]—of the judicial district. I could then use Census tract data to make inferences about race and ethnicity of the persons to whom the summonses were sent. The geocoding method, and some of its limitations, are described more fully in Section 10 and Appendix C.

I put a lot of effort into the geocoding, spending many hours (by computer and manually) checking for mismatches, correcting for misspellings and abbreviations, locating new streets on tract maps, cross-checking with several other sources, and generally finetuning the matching algorithms. I consulted with Postal representatives to resolve a number of apparent inconsistencies. I particularly concentrated my efforts on Hartford and New Britain because those two towns contain a large proportion of the minority population.

I also took some pains to check for any possible patterns in the addresses I could not geocode, and used surname matching (see below) as a safeguard against systematic error. With sample sizes as large as for the JIS data, systematic error will be more important than random error caused by sampling fluctuations.

With the finetuning, the method was able to get unique matches for well over 90% of the addresses for Hartford and New Britain towns, with a more modest 80+% matching rate over the whole HNB district, as shown in the small table. My geocoding method could not give 100% matching of juror records to geographical locations, partly because of ambiguous addresses, partly because it would be deceptive to locate all Post Office boxes (POB) at a single point in a town, and partly because disqualifications coded ‘xjd’ should not correspond to a location in the judicial district. If the Post Office boxes and xjd records are excluded, the matching rates are higher. (The column headed ‘excl. POB/xjd’ excludes all addresses that are only POB’s and all addresses outside the judicial district from the denominator, with only unique tract matches in the numerator.)

	Unique	excl. POB/xjd
HNB9293	82.3%	88.4%
HNB9394	82.4%	88.6%
HNB9495	82.3%	88.5%
HNB9596	81.8%	87.9%
HNB9697	83.3%	88.0%

The second method of inference applies only to Hispanics. It uses data collected by the Bureau of the Census, in the form of a “Spanish Surname List”, to draw inferences about Hispanic origin based on a person’s surname. The method, which I refer to by the acronym SSL, and its limitations, are described more fully in Section 9.

The following four tables present my estimates in the form of percentages of all summonses (sent to various groupings of persons) for the same disqualification categories as above. The tables refer to the whole HNB judicial district, for the four court years 1992-93 through 1995-96.

[4] For example, the town of Hartford is divided into 49 disjoint Census tracts.

ESTIMATES FOR THE WHOLE OF HNB

HNB9293	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	4	3	8	3	20	2	13	43	5		100
Hgeo	3	14	3	3	29	3	12	27	5		100
SSLgeo	3	13	2	2	31	3	12	32	3		100
ALLgeo	3	2	12	4	12	1	5	54	6		100
SSL	3	13	2	2	29	3	12	32	3	2	100
nonH	2	1	12	4	11	1	4	53	6	6	100
ALL	2	2	11	4	12	1	5	52	6	5	100

HNB9394	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	4	2	8	3	22	3	12	41	5		100
Hgeo	3	13	3	3	33	3	10	26	5		100
SSLgeo	3	11	2	2	33	4	11	31	3		100
ALLgeo	2	2	12	5	14	1	5	51	7		100
SSL	3	11	2	2	32	4	10	31	3	2	100
nonH	2	1	12	5	12	1	4	50	7	6	100
ALL	2	2	11	4	13	1	4	49	7	5	100

HNB9495	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	3	3	8	4	25	4	11	37	5		100
Hgeo	3	13	3	4	37	4	10	22	5		100
SSLgeo	2	11	2	3	38	5	10	26	3		100
ALLgeo	2	2	12	7	15	2	5	47	7		100
SSL	2	11	2	3	36	5	10	26	3	2	100
nonH	2	1	12	7	13	1	4	47	7	6	100
ALL	2	2	11	6	14	2	4	45	7	6	100

HNB9596	01	06	08	12	13	17	??	NS	OK	rest	xjd	total
Bgeo	3	3	8	3	21	4	9	6	39	5		100
Hgeo	3	15	3	4	29	5	8	5	24	5		100
SSLgeo	2	13	2	3	32	6	7	4	28	3		100
ALLgeo	2	2	12	6	13	2	5	2	47	7		100
SSL	2	13	2	3	31	5	7	4	28	3	2	100
nonH	2	1	12	6	11	2	5	2	45	7	6	100
ALL	2	2	12	6	13	2	5	2	44	7	6	100

As I explain in Appendix D, the effect of sampling fluctuations is not large enough to account for the more striking differences in disqualification rates shown by the tables. For example, at worst, sampling fluctuations could account for something on the order of one or two percentage points in the code 13 estimates.

The first four rows of each table—the rows labelled with an abbreviation ending in "geo"—give percentages only for those summonses that I could geocode into a unique Census tract.

Bgeo → For the first row (Bgeo), I estimated using **geocoding** the number of summonses for each disqualification category sent to a **black person**. The sum across all disqualifications gave an estimate of the total number of summonses (whose addresses I could uniquely geocode) sent to blacks. For example, for the 1992-93 court year, over the whole HNB judicial district, I estimate that 20% of the summonses sent to blacks were undeliverable (code 13).

Hgeo → The second row (Hgeo) similarly estimates the disqualifications of Hispanics for each disqualification category, expressed as percentages of the total number of summonses sent to **Hispanics**. Again the calculations were based only on the 80+% of uniquely **geocoded summonses**, but I was able to refine the method of estimation by drawing on more Census data to better identify the "eligible population" within each tract for each disqualification. The details are given in Appendix C.

SSLgeo → The third row (SSLgeo) provide a valuable cross-check on the geocoding estimates for Hispanics. For that row I applied **surname matching** to the uniquely **geocoded summonses**—the same summonses as used for the geocoding estimates. That is, I have applied two distinct methods of estimation to the same set of summonses, in order to test the two methods directly against each other. Comparison of the corresponding percentages for the Hgeo and SSLgeo rows gives a good cross-check of the two methods of estimation.

ALLgeo → The fourth row (ALLgeo) expresses the counts of **all** uniquely **geocoded summonses** of each disqualification category as percentages of the total count of uniquely geocoded summonses. By contrast, the last row in the table (ALL) gives percentages for **all summonses**, not just those that were uniquely geocoded. The close agreement between the ALLgeo and ALL rows (except 'xjd') gives me confidence that the geocoding is selecting out a large representative subset of the summonses. The 'xjd' percentages were different because I chose not to geocode summonses based on addresses outside the judicial district.

ALL →

SSL →

The fifth row (SSL) gives percentages of disqualifications for **Hispanics**, based on **surname matching** applied to **all summonses**. The close agreement between the SSLgeo and SSL rows (except 'xjd') again suggests that the geocoding is selecting out a large representative subset of the summonses sent to Hispanics.

nonH →

The sixth row (nonH) was obtained by subtracting the estimates for Hispanics based on surname matching from the counts of all summonses for each disqualification category. It effectively estimates disqualifications for **nonHispanics** by **surname matching**. There is little difference between the nonH percentages and the ALL percentages, because Hispanics are only a small fraction of the whole population.

Each of the four tables for HNB contains three estimates for the percentage of undeliverable summonses sent to Hispanics. The twelve percentages are nearly all greater than 30%. Similarly the estimates for nonHispanics are all close to 13%.

Nicolas Hengartner has suggested that differences between the first four rows and the last three rows of each table would be easier to interpret if the xjd numbers were excluded from the denominator for 'SSL', 'nonH', and 'ALL' rows. In principle he is correct, but actually the modification has only a tiny effect on the tables: the 'ALL' row and 'ALLgeo' rows become almost identical if xjd are excluded. The change has no effect on my overall interpretation.

The corresponding estimates for summonses sent to addresses in Hartford town (abbreviated HAR) tell a similar story.

ESTIMATES FOR HARTFORD TOWN

HAR9293	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	4	3	6	2	25	3	17	37	4		100
Hgeo	3	15	2	2	36	4	14	20	4		100
SSLgeo	2	14	1	1	40	4	14	22	2		100
ALLgeo	4	5	8	2	27	3	12	35	5		100
SSL	2	14	1	1	39	4	14	22	2		100
nonH	4	2	10	2	23	2	11	39	5		100
ALL	4	5	8	2	27	3	12	35	5		100

HAR9394	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	4	3	6	2	28	4	14	36	4		100
Hgeo	3	13	2	2	40	4	12	20	4		100
SSLgeo	2	12	1	1	42	5	12	23	2		100
ALLgeo	3	5	8	2	31	3	10	32	4		100
SSL	1	12	1	1	42	5	12	23	2		100
nonH	4	2	10	2	28	3	10	35	5		100
ALL	3	5	8	2	31	3	10	33	4		100

HAR9495	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	3	3	6	3	32	4	14	31	4		100
Hgeo	2	12	2	2	45	5	12	16	3		100
SSLgeo	1	11	1	2	48	6	12	17	2		100
ALLgeo	3	5	7	3	36	4	11	27	4		100
SSL	1	11	1	2	47	6	12	18	2		100
nonH	3	2	10	3	32	3	10	31	5		100
ALL	3	5	7	3	36	4	10	27	4		100

HAR9596	01	06	08	12	13	17	??	NS	OK	rest	xjd	total
Bgeo	3	3	6	2	25	5	10	7	34	4		100
Hgeo	3	15	2	2	35	6	9	6	18	4		100
SSLgeo	1	13	1	2	40	7	7	5	21	2		100
ALLgeo	3	5	8	2	28	5	8	5	31	4		100
SSL	1	13	1	2	40	7	7	5	21	2		100
nonH	4	2	11	2	24	4	8	5	34	5		100
ALL	3	5	8	2	28	5	8	5	31	4		100

Notice that the xjd columns are empty: the presence of the Hartford towncode in a juror record automatically eliminates the xjd classification.

In (rough) summary:

- For the whole Hartford-New Britain judicial district, over 30% of the summonses sent to Hispanics are undeliverable (code 13), compared with about 13% for nonHispanics. The undeliverable problem is particularly bad for the town of Hartford, which contains a large fraction of the minority population. Moreover, the problem is even worse if one adds in the code 17 disqualifications—the second form of undeliverable classification.

Similarly, I summarize the OK columns by asserting:

- The qualification rate (percentage on the master list who actually qualify to serve as jurors) for Hispanics over the whole HNB judicial district is mostly under 30% compared with almost 50% for nonHispanics. For Hartford town, the qualification rates are even worse.

Similarly, I summarize the forty-eight percentages (rows Hgeo, SSLgeo, SSL; columns 01 and 06; four tables for each of HNB and HAR) for disqualifications of Hispanics for noncitizenship or inability to speak/understand English by asserting that:

- Of the Hispanics who make it to the master list, about 13% to 15% are disqualified on language grounds and about 3% for noncitizenship.

In the Penultimate report I continued: *These disqualifications are largely counterbalanced by an unusually low disqualification for Hispanics over-70 (about 2% for Hispanics compared with about 12% for nonHispanics). The undeliverable problem accounts for over 30% of the Hispanic disqualifications, compared with about 13% for nonHispanics.*

My choice of the word "counterbalanced" was unfortunate because apparently it suggested to some readers that some disqualifications are cancelling out the existence of other disqualifications. It has been proposed, for example, that the language disqualifications account for most of the shortfall in the qualification rates for Hispanics, and that therefore the undeliverable problem can be ignored. One could just as well argue that the severity of the undeliverable problem is being masked by the fortuitously low disqualification rate for over-70, and that the underrepresentation will get worse as the Hispanic population ages.

I have more confidence in the estimates for the Hispanic population than the estimates for the black population, because:

(i) I was able to make better use of tract data to identify the "eligible populations" for several of the disqualification categories, and the fractions of them that were Hispanic, within each tract;

(ii) I had two distinct methods to apply to the estimation of Hispanic counts.

Nevertheless, the geocoding estimates do suggest some race effect:

- Blacks also have a higher undeliverable rate and a lower qualification rate than the general population of the judicial district, but the differences are not as extreme as for the Hispanic population.

The data from the questionnaires appears to contradict the suggestion about lower qualification rates for blacks. However, the questionnaires were filled out by some persons who were later disqualified after signing in at the courthouses. If minorities were more likely to turn up at a courthouse despite cancellation (a possibility suggested by the tabulations for HHD at the end of Section 2), or if minorities were more likely to be disqualified after appearing at the courthouse, then their qualification rates would be lower than suggested by the questionnaire responses.

Other stages in the jury summoning process

The surname matching and geocoding estimates, based on JIS records of actual summonses sent, reflect only the workings of the disqualifications process after the construction of the master list of potential jurors. There are two earlier steps in the process that also affect minority representation.

The master list is constructed by sampling from two source lists: the voter list from each town, and the DMV list of licenced drivers. The surname matching and geocoding methods also gives estimates of the number of Hispanics who made it to the master list. However, I know that geocoding (see Section 10) will tend to underestimate Hispanic proportions if they have increased significantly since the 1990 Census, while surname matching seems to have only a slight sytematic error (at least for populations like the sample taken for the questionnaires.) If the over- or under-estimation effects are not heavily concentrated in any particular disqualification code, the methods will still be valid for estimation of relative proportions or percentages of total counts. The evidence from the tabulations suggests that there is no such concentration. For estimates of total counts, I would expect a widening gap (until new Census data became available) between surname matching and geocoding estimates. The figures in the next table, which gives the estimated counts for the various methods described above for HNB in each of four court years, show just such an effect.[5] Compare the Hgeo and SSLgeo lines:

	HNB9293	HNB9394	HNB9495	HNB9596
Bgeo	6707	5621	7116	7198
Hgeo	4160	3362	4522	4379
SSLgeo	4714	3823	5468	5645
ALLgeo	73248	55898	71272	71959
SSL	5337	4295	6086	6400
nonH	83647	63528	80523	81578
ALL	88984	67823	86609	87978

To get an estimate of the fraction of Hispanics on the master list, one has only to divide either the Hgeo or SSLgeo figures by the corresponding ALLgeo figure, or divide SSL by the corresponding ALL.

	HNB9293	HNB9394	HNB9495	HNB9596
Hgeo	5.7	6.0	6.3	6.1
SSLgeo	6.4	6.8	7.7	7.8
SSL	6.0	6.3	7.0	7.3
CB/S	6.88	7.10	7.29	7.56

For comparison, I have added a row (CB/S) showing the Census Bureau/Steahr estimates for the percentage Hispanic in the over-20 population of Hartford County, as of July 1 of each year. The over-18 proportions corresponding to CB/S would be slightly larger.

The existence of the widening gap between Hgeo and the surname matching methods has just been explained. The reason for the gap between SSLgeo and SSL is less obvious. The answer is to be found in the xjd counts in the full listing at the end of Section 3 in Appendix C. For example, for HNB9495, the SSL method applied to all summonses estimated only a very low count of Hispanics ($95 + 38 + 16 = 149$) out of a total of $3339 + 1244 + 290 = 4873$ summonses classified xjd (codes 02, 15, or 16). That is, SSL estimates only 3.1% Hispanic amongst the xjd. If the xjd estimate/count were removed from both numerator and denominator, the SSL estimate would decrease to 7.3%, which is much closer to the SSLgeo figure. By the same token, the SSLgeo is overestimating the total proportion of Hispanics, because the ALLgeo denominator does not include the xjd counts.

Given the (not unexpected) range between the estimates, I can draw less precise conclusions concerning the proportion of Hispanics on the master lists

[5] Sharp-eyed readers will detect the tiny effects of rounding error if they compare the first table with the corresponding tables in Appendix C.

than I can draw about the proportions of those Hispanics disqualified in various ways. For the Penultimate Report, I tried to summarize the comparisons between my three estimates of Hispanic proportions on the master list and the Census Bureau/Steahr estimates in a suitably cautious way:

- *Hispanics are slightly underrepresented on the combined source lists—from somewhere between* $1/2$ *to* 1 *percentage point out of the 6% to a 7% of the over-18 population of the Hartford-New Britain judicial district was Hispanic. (The figures changed over the four court years covered by the JIS data.) That is, a moderate fraction of the Hispanic population is lost to the system even before summonses are sent out. (I cannot be more precise about this assertion, because I am steering between two estimates, one of which I expect to give a slight underestimate and the other a slight overestimate.)*

The "6% to 7%" was my attempt to summarize roughly a change from 6.56% in 1990 to something over 7% by 1995-96. I regret that my attempt at summary caused confusion. For the Final version of my report, I would prefer to let the reader draw his own conclusions about coverage of the Hispanic population by the source lists from the the evidence presented in the table and from the reasons I have given for the differences between the estimates.

The interpretation of the coverage figures for the source lists is complicated by two other problems (discussed in Section 7 and Appendix B), whose existence I discovered only after many months of analysis of the data. The first problem is caused by a failure of JIS to follow the Statute governing the construction of the master list:

- JIS uses an inappropriate sampling procedure in the construction of its master list, from which summonses are drawn. The procedure over-samples persons who are on both the voter list and the DMV list.

The second problem is more mysterious. The information I have about the JIS procedures implies that the set of summonses for each town in the summary files should be (roughly) a simple random sample from the combined DMV and voter samples for the town; but the actual proportions of summonses originating in the voter samples, as identified by JIS sourcecodes, are wildly inconsistent with such an hypothesis.

- Somehow, at some stage between the sampling from source lists and the sending of summonses, JIS is systematically oversampling jurors drawn from the DMV lists for each town.

I suspect the problem is caused by an inappropriate method of randomization applied when JIS merges the DMV and voter samples. The problem is discussed in Section 7.

2. The questionnaire data

My initial involvement with the problem of jury selection arose from a request by the Public Defender's Office that I analyse a batch of questionnaires distributed to potential jurors for the Kevin King trial at Hartford Superior Court. The questionnaires asked jurors to check off one of five race categories, and also to answer an ethnicity question asking whether they were Hispanic or not. (They were also asked to give their juror id numbers, and sign their names, but I made no use of those two pieces of information.)

	Hisp	Non-Hisp	??	total
1 (= Black)	2	122	69	193
2 (= White)	11	1519	319	1849
3 (= AmerInd)		3	1	4
4 (= Asian)		12	5	17
5 (= Other)	19	7	5	31
1+2		1		1
1+2+3			1	1
1+3		3		3
2+3		4	2	6
2+5		1		1
??	51	1	8	60
total	83	1673	410	2166

The table contains a slight rearrangement of data I presented at a preliminary hearing in Hartford Superior Court, 8th and 13th February 1996. The column headings in the table indicate answers to the ethnicity question, with ?? denoting a nonresponse. I coded the race responses as 1 = black, 2 = white, . . . , 5 = other. Thus 1+2+3 corresponds to a juror who checked three race categories: black, white, and amerind. The ?? again denotes a nonresponse. For example, 8 jurors answered neither question, and 319 jurors identified themselves as white but did not answer the Hispanic question. The data were incomplete, possibly because of the order in which the two questions appeared on the questionnaire—a sizeable fraction of jurors did not answer the ethnicity question.

The large number of nonresponses to the questionnaires for the King trial made it difficult to draw convincing conclusions about minority representation on the jury panels. Similar problems of nonresponse are well known to the Bureau of the Census: In answer to a question at the August 1996 Joint Statistical Meeting at Chicago, Manuel de la Puente (Chief, Ethnic and Hispanic Statistics Branch, Population Division of the US Bureau of the Census) explained that response rates for questions regarding ethnicity are known to be affected by previous questions regarding race. Further explanation appears in a paper of Gerber & de la Puente (1996). For example, (pp. 3–4): '. . . many survey respondents tend to use the terms "race" and "ethnic origin" interchangeably, and they do not clearly distinguish between the two concepts.', and (p. 5):

> In the 1990 census the race question preceded the Hispanic origin question on the census form. In the 1990 census, 373,100 persons who provided a Hispanic write-in response (such as "Mexican", "Puerto Rican" or "Spanish") in the race question did not respond to the Hispanic origin question. Cognitive research, as well as in-depth interviews and focus groups, with Hispanics of different national origins show that some Hispanics find the race and Hispanic origin questions redundant because these questions are viewed as asking for the same information (. . .). These findings were confirmed in our research.

based on 6.56Hispanic over-18 based on 7.5Hispanic over-18

A new court order on 26 March 1996 required jurors subsequently appearing at any of the Hartford-New Britain (HNB) courthouses to complete a new questionnaire. One courthouse declined to participate. With a rearranged questionnaire form and more careful supervision by court personnel, there

have been fewer missing answers. For the questionnaires from April 1996 through early 1997, the responses appear in the next table on the left; the separate column (headed Race %) on the right gives the responses for the race question expressed as a percentage of the total number (22719) of questionnaires returned.

SUPPLEMENTAL QUESTIONNAIRES: APRIL 1996 THROUGH EARLY 1997

	Hisp	Non-Hisp	??	total	Race %
1 (= Black)	24	1950	30	2004	8.90
2 (= White)	214	18947	70	19231	85.41
3 (= AmerInd)	8	37	1	46	0.20
4 (= Asian)	3	212		215	0.95
5 (= Other)	374	106	8	488	2.17
1+2		12		12	0.05
1+2+3		9		9	0.04
1+3		20		20	0.09
1+4		1		1	0.00
1+5	2	7		9	0.04
2+2		1		1	0.00
2+3		53	2	55	0.24
2+3+5		3		3	0.01
2+4	1	7		8	0.04
2+4+5			1	1	0.00
2+5	11	24	2	37	0.16
3+5		1		1	0.00
4+5		2		2	0.01
5+1	1			1	0.00
??	300	23	48	371	1.65
n				0	0.00
total	938	21415	162	22515	100

The 8.82% of the questionnaires for the jurors who identified themselves as black is close to the figure obtained from the 1990 census: 9.09% of the over-18 population of the HNB judicial district was counted as black. (The figure is derived from STF1A, as explained in Section 6.) The percentages of the totals for the answer to the ethnicity question are more suggestive of some underrepresentation:

	Hisp	Non-Hisp	??	total
q'naires	4.13	94.26	0.71	100

The 4.13% of the questionnaires for the jurors who identified themselves as Hispanic is significantly smaller (in the senses explained in the next Section) than the figure obtained from STF1A of the 1990 census: 6.56% of the over-18 population of the HNB judicial district identified itself as Hispanic. Moreover, population changes since the 1990 Census can only strengthen the conclusion: as shown by the data in Section 6, the figure 6.56% is undoubtedly an underestimate of the current proportion of Hispanics in the over-18 population of HNB. According to demographic projections carried out by Dr. Thomas Steahr,[6] the figure is probably over 7.8%.

The responses to the ethnicity question by month suggest that the Hispanic representation drops off during the court year: from about 4.6% in April 96 (=9604) to about 3.7% in August, followed by a jump at the start of the

[6] The Steahr projections were entered into evidence at the Rodriguez trial on 17 January 1997.

new court year in September. The monthly counts are subject to random fluctuations large enough to produce some of the observed differences. (Smaller sample sizes make random error relatively more important.) I subjected the monthly data to no formal statistical testing, because I regarded the apparent downward trend merely as a hint of what I might expect to see in the more extensive JIS data.

SUPPLEMENTAL QUESTIONNAIRES BY MONTH

	Hisp	Non-Hisp	??	total
7905		1		1
9604	80	1659	8	1747
9605	66	1741	15	1822
9606	35	1050	7	1092
9607	57	1421	6	1484
9608	15	387	1	403
9609	80	1436	13	1529
9610	94	1847	23	1964
9611	60	1386	26	1472
9612	59	1140	8	1207
9701	77	1527	8	1612
9702	85	1812	2	1899
9703	95	2137	13	2245
9704	87	2322	18	2427
9705	61	1719	15	1795
9706		18		18
total	951	21603	163	22717

	Hisp	Non-Hisp	??	total
7905	0.00	100.00	0.00	100.00
9604	4.58	94.96	0.46	100.00
9605	3.62	95.55	0.82	100.00
9606	3.21	96.15	0.64	100.00
9607	3.84	95.75	0.40	100.00
9608	3.72	96.03	0.25	100.00
9609	5.23	93.92	0.85	100.00
9610	4.79	94.04	1.17	100.00
9611	4.08	94.16	1.77	100.00
9612	4.89	94.45	0.66	100.00
9701	4.78	94.73	0.50	100.00
9702	4.48	95.42	0.11	100.00
9703	4.23	95.19	0.58	100.00
9704	3.58	95.67	0.74	100.00
9705	3.40	95.77	0.84	100.00
9706	0.00	100.00	0.00	100.00
total	4.19	95.10	0.72	100.00

Of course there will be some argument about the race or ethnicity of those jurors who did not answer the questions, but I would strongly maintain that most of those who did not answer the Hispanic question should not be regarded as Hispanic. My evidence consists chiefly of their full names, their answers to the race question, and an assessment of their Hispanic origin based on the surname matching method described in Section 9. It suggests that only four or five of those jurors were Hispanic. To respect the privacy of persons who filled out the questionnaires, the evidence is not included in the present report.

In summary: The questionnaires suggest very strongly that, for whatever reasons, Hispanics really are underrepresented in the pool of qualified jurors, at least by comparison with their proportion of the over-18 population of HNB judicial district. In the next Section I explain some of the formal ways of quantifying the underreprestation. In the context of jury selection, one cannot rely exclusively on the results from questionnaires that are administered after jurors have already passed through various disqualification filters. Unless one adjusts for the known differences in the effects of the disqualifications on different subgroups of the population, it is unwise to infer anything beyond the existence of a significant difference.

April 1997 update

During my testimony in January 1997, some possible problems with regard to the validity of the questionnaire data (for the April 1996 through November 1996) were identified. The main difficulties were:

(i) The Manchester courthouse (H12M) did not participate in the distribution of the supplemental questionnaires.

(ii) The Bristol courthouse (H17B) summoned no jurors from April 1996 until February 1997.

(iii) The Enfield courthouse apparently failed to collect supplemental questionnaires during part of the period.

(iv) Different courthouses followed different procedures regarding which jurors filled out the questionnaires.

(v) There are different rates of disqualifications and different no-show rates for the different courthouses.

(vi) Some of the persons who signed questionnaires were later disqualified by the courts, for various reasons.

(vii) Some jurors whose service had been cancelled turned up at the courthouse and filled in the questionnaires.

(viii) There was a discrepancy of a few hundred between the number of questionnaires that I analyzed and the number claimed to have been forwarded to me by the printer.

In response to these difficulties, I made a much more detailed study of the questionnaires, using the new data obtained from JIS in January. I was able to identify uniquely the juror id-number for all except a small handful of questionnaires. I was also able to eliminate a larger number of duplicate questionnaires and questionnaires from jurors who had filled out more than one questionnaire.

The first five difficulties can be overcome by analyzing the questionnaire data separately for each of the courthouses that participated.

The next table shows the distribution of questionnaires by courthouse and disqualification code. The meanings of most of the disqualification codes (the column headings of the table) are explained in Section 4. The blank code corresponds to jurors whose id-number I could not determine. Jurors 'excused by the court' have code 99. The OK code is my invention to denote jurors who were 'qualified', with OK.X in this table denoting either (i) qualified jurors whose service was cancelled by the court, but who turned up at the courthouse anyway, or (ii) jurors ("walkins") who turned up on the wrong date and were not disqualified. Also, only for the purposes of these tabulations, I have assigned to OK those jurors with 'unknown disqualification status' who signed questionnaires and were not disqualified, even though some small fraction might possibly have postponed and could later be disqualified in some way. (It would make little difference to the conclusions if I were to assign them to the OK.X category instead.) The row labels denote courthouses.

QUESTIONNAIRES BY COURTHOUSE

[HNB]		01	02	05	06	08	09	11	12	15	17	99	NS	OK	OK.X	total
COURT?	809															809
H12M												1			1	2
H13W		1												195	7	203
H17B									1					180	3	184
HHB								2	1			134		3745	35	3917
HHD			2	2	4	4	12	4	7	5	5	508	4	11915	361	12833
total	809	1	2	2	4	4	12	6	9	5	5	643	4	16035	407	17948

Clearly the main Hartford courthouse (HHD) and the New Britain courthouse (HHB) account for most of the questionnaires collected. The counts for Enfield (H13W) are small, and are also suspect on other grounds. For example, I have questionnaires from H13W for dates when the courthouse was sup-

posedly not distributing the questionnaires. If there is an important courthouse effect, only HHD and HHB have provided enough data for it to be found.

[HHD]	Hisp	NonHisp	??	total
1	11	1184	12	1207
1+2		6		6
1+2+3		2		2
1+3		13		13
1+5	2	5		7
2	126	10715	23	10864
2+2		1		1
2+3		24	1	25
2+3+5		2		2
2+4	1	5		6
2+5	6	14	1	21
3	6	22		28
4	3	111		114
4+5		1		1
5	219	66	6	291
5+1	1			1
??	207	15	22	244
total	582	12186	65	12833

[HHB]	Hisp	NonHisp	??	total
1	5	295	11	311
1+2		3		3
1+2+3		1		1
1+3		1		1
1+4		1		1
2	29	3338	29	3396
2+3		11	1	12
2+3+5		1		1
2+4		1		1
2+4+5			1	1
2+5	2	5	1	8
3	1	7	1	9
4		48		48
5	71	20		91
??	25	4	4	33
total	133	3736	48	3917

The raw counts by courthouse still suffer from the difficulties (vi) and (vii). As the next set of tables indicates, minorities were overrepresented amongst persons who turned up at HHD despite cancellation. The questionnaire responses do slightly overestimate the minority proportions amongst qualified jurors.

Supplemental questionnaires for HHD and HHB

[HHD]	02	05	06	08	09	11	12	15	17	99	NS	OK	OK.X	total
1					3		1		1	45	1	1093	63	1207
1+2												6		6
1+2+3												2		2
1+3										1		12		13
1+5												6	1	7
2	1	1		4	9	4	5	5	4	422	3	10149	257	10864
2+2												1		1
2+3												23	2	25
2+3+5												2		2
2+4												6		6
2+5										2		18	1	21
3	1									1		25	1	28
4										6		102	6	114
4+5												1		1
5			1				1			13		263	13	291
5+1										1				1
??		1	3							17		206	17	244
total	2	2	4	4	12	4	7	5	5	508	4	11915	361	12833

[HHB]	11	12	99	OK	OK.X	total
1		1	11	296	3	311
1+2			1	2		3
1+2+3			1			1
1+3				1		1
1+4				1		1
2	2		110	3259	25	3396
2+3			2	10		12
2+3+5				1		1
2+4			1			1
2+4+5			1			1
2+5				7	1	8
3			1	7	1	9
4			3	44	1	48
5			3	86	2	91
??				31	2	33
total	2	1	134	3745	35	3917

[HHD]	02	05	06	08	09	11	12	15	17	99	NS	OK	OK.X	total
Hisp		1	4				1	1		29	1	515	30	582
NonHisp	2	1		4	11	4	6	4	5	474	3	11351	321	12186
??					1					5		49	10	65
total	2	2	4	4	12	4	7	5	5	508	4	11915	361	12833

[HHB]	11	12	99	OK	OK.X	total
Hisp			3	125	5	133
NonHisp	2	1	129	3575	29	3736
??			2	45	1	48
total	2	1	134	3745	35	3917

To eliminate completely the effects of (vi) and (vii), I set aside all questionnaires except those for jurors in my OK disqualification category whose service was not cancelled. The Hispanic representation for both HHD and HHB are lower than for the entire set of questionnaires. (The counts from H13W and H17B are too small to be informative; the sampling fluctuation for

such a small sample would swamp the sytematic difference. I included the H13W and H17B counts merely for bookkeeping purposes.)

[OK]	Hisp	NonHisp	??	total
H13W	11	180	4	195
H17B	7	173		180
HHB	125	3575	45	3745
HHD	515	11351	49	11915
total	658	15279	98	16035

[OK]	Hisp	NonHisp	??	total
H13W	5.6	92.3	2.1	100.0
H17B	3.9	96.1	0.0	100.0
HHB	3.3	95.5	1.2	100.0
HHD	4.3	95.3	0.4	100.0
total	4.1	95.3	0.6	100.0

Only difficulty (viii) remains. I have evidence that refutes the suggestion that actual juror questionnaires were lost. I will submit the evidence to the court.

My bottom line—after all the extra work involved in matching juror questionnaires with JIS records, and after adjusting for the difficulties identified in (i) through (vii)—is the same as before. The questionnaire data do strongly suggest an underrepresentation of Hispanics, but the matter cannot be settled without further investigation into the effects of the disqualifications.

3. Measures of disparity

In legal jargon, the 4.13% proportion of Hispanics on the questionnaire would be called an ***absolute disparity*** of $6.56\% - 4.13\% = 2.43\%$, or a ***relative disparity*** of $(6.56 - 4.13)/6.56 \approx 37.06\%$. The relative disparity is also called comparative disparity. The 6.56% in these calculations refers to the fraction of the over-18 population of HNB judicial district counted as Hispanic in the 1990 Census. (The true relative disparity is probably close to 45%, because of the growth in Hispanic population since the 1990 Census.) Some documents refer to the calculation of the tiny probability

$$\mathbb{P}\{Binomial(22719, 0.0656) \le 938\} \approx 10^{-54}$$

as an application of 'Statistical Decision Theory' (SDT)[7], although it is really just a simple calculation of a *p-value*. The term 'statistical significance test', as in Kairys, Kadane & Lehoczky (1977, p. 792), would be more appropriate.

There are a number of other ways of expressing the disparity between observed proportions and various target proportions, which have been cited in the case law.[8] Some parties advocate a comparison with a target group more narrowly defined than the proportion in the over-18 population; some parties advocate comparison with the proportion in the total adult population. I will submit to the court a tabulation of a variety of disparity measures and comparisons in a separate document, after the questionnaire collection for the Rodriguez trial is completed.

There has been some misinterpretation of the p-value. It is calculated (using accepted methods of approximation) under an assumption of random sampling from a population of given size containing a given proportion of Hispanics. By carrying out the calculation I am not accepting the validity of the sampling assumption. Indeed, the whole point of the calculation is to show how implausibible the assumption is: the p-value shows how extremely unlikely it would be for a sample of size 22719 (from a population with 6.56% Hispanic) to generate so few Hispanics. It demolishes the explanation that the discrepancy is explicable as a chance fluctuation for ran-

[7] Terminology introduced by Finkelstein (1966). As understood for currently accepted statistical jargon, the terminology is misleading.

[8] See, for example, State vs. Castonguay, 194 Conn 416 September 1984.

dom sampling from a population with over 6.56% Hispanics. Nothing more is claimed. It suggests strongly that some other mechanism must be at work to generate the observed questionnaire responses.

According to an explanation proposed by the State at the King trial, the lower proportion of Hispanics might be due to two factors: jurors are disqualified from serving if they are not "citizens of the United States" or if they are "not able to speak and understand the English language". (See Section 4 for a listing of other causes for disqualification.) If a large enough fraction of Hispanics were being disqualified on those two grounds it might explain the apparent underrepresentation—those person would not appear at the courthouse to fill out the questionnaire. The questionnaire data themselves shed no light on this claim.

Ambiguity in the use of the word "random" has also caused some confusion. Sometimes it is used to refer to sampling whereby each individual in a population has an equal chance of being selected, or where each subset of the population of a given size has an equal probability of being chosen. Some authors use the words "uniform random" to refer to such an interpretation.

The word random can also be used legitimately in situations where not every individual has the same chance of being selected. For example, if I hold two tickets in a fair lottery and you hold only one, then we do not have the same chance of winning, even though the drawing of the winning ticket should be a random event. Sometimes randomness is understood in an even wider sense, to indicate unpredictability of a precise outcome. For example, the winner of the next 100 meter dash at the Olympic games is not predictable, but that is not to say that every sprinter has an equal probability of being the next gold medalist.

Randomness followed by nonrandom intervention can still result in randomness. For example, suppose each adult in a town holds two tickets in a fair lottery. Then the outcome is random, in the uniform sense. If, by some quirk of fate, every blue-eyed, blond man loses one of his tickets, the outcome is still random—in the sense of unpredictablity—but most Scandinavian males in the town will have only half the chance of winning compared with their brown-eyed neighbors. Even if all blue-eyed males have both their tickets confiscated, the outcome is still random, even though not every adult has the same chance of winning.

In short, if we know the mechanism that intervenes, we can sometimes still assign probabilities to the various outcomes, even if not every individual still has the same chance of success.

The following excerpt[9] illustrates some of the difficulties in interpretation created by the various meanings of the word random.

> (1) The intellectual core of SDT is **random** selection.
>
> (2) SDT measures the probability that the selection of a particular class of jurors (eg. blue-eyed, blond men) is **random**.
>
> (3) In the jury context, the greater the chance of **random**ness, the "better" the juror selection system.
>
> (4) But if the sample is not **random** (eg. all Scandinavians are excluded from the sample), SDT will produce a skewed probability prediction.
>
> (5) It is illogical to apply a theory based on **random** selection when assessing the constitutionality of a qualified wheel.

[9] US vs. Rioux (97 F.3d 648, *655); emphasis and sentence numbering added

(6) By definition, the qualified wheel is not the product of **random** selection; it entails reasoned disqualifications based on numerous factors.

(7) It is irrational to gauge the qualified wheel—an inherently non-**random** sample–by its potential for **random**ness.

The first sentence is correct, in the sense that SDT (and statistical inference in general) is based on the calculation of probabilities.

The second sentence is not quite accurate. The usual calculation does not give a probability that the process is random. Instead, it calculates probabilities of particular events under postulated mechanisms. The occurence of an outcome that should have been very rare under a particular mechanism casts doubt on any assertions that the data were generated by that mechanism. For example, if a large sample contains a very low fraction of Scandinavians relative to their proportion of the popululation that was sampled, then one begins to doubt any assertion that the sample was generated by a procedure that gave equal probability of selection to each member of the population.

The third sentence uses random in the sense of equal probabilities: it appears to be an an assertion that equal probability of selection is a good thing for a jury system.

The fourth sentence is correct only if SDT is turned around and used as a method for predicting what should have happened. If a probability calculation is based on a model that is known to be invalid in a particular setting, then the probability prediction has no relevance as a prediction of behavior under the known mechanism. However, it is still a valid calculation; it can still be used to destroy the credibility of anyone who asserts that the invalid model is the truth.

The fifth sentence points out that uniform randomness is no longer an interesting hypothesis to be testing.

The sixth sentence notes that the disqualifications have disturbed the uniform randomness.

In the last sentence, random in both cases refers to uniform randomness.

*****

Each of the measures of disparity provides evidence regarding one aspect of the jury selection process, namely, that the proportion of Hispanics in the final yield of qualified jurors is ‘significantly’ different from the proportion of Hispanics in the population from which the source lists are drawn. In fact, only the SDT calculation gives precise meaning to the term ‘significant’; the other calculations come with no mathematical calibration to aid the courts in their judgements of how large a disparity is ‘significant’.

None of the measures of disparity speaks directly to the fairness or representativeness of the selection process, because none of the calculations takes account of the mechanisms (such as statutory disqualifications) that control the process. The calculations can reveal existence of a disparity; and SDT can also provide overwhelming evidence that the disparity should not be interpreted as just some random fluctuation, due to sampling effects, around a population figure. But to decide whether a disparity implies a violation of legal rights, I believe one should enquire into the reasons behind the disparity.

4. The JIS data

Most of the data were transferred (using FTP) to a Statistics Department workstation by Mr. Lou Sapia, the programmer at JIS responsible for maintaining the juror database. Lou made the first transfer in March 1996. In addition, I received a number of printed reports and documents from JIS.

After some analysis—including a study of the work carried out by Jason Cross[10], a Yale Statistics doctoral candidate, for his practical project in the spring of 1996—and much discussion with Lou Sapia and Mr. Richard Gayer (the Jury Administrator for the State of Connecticut), and much study of Census data and other documents, I concluded that more data, in a slightly different form, would be helpful.

In August 1996 Lou Sapia made another transfer by FTP of JIS files for the whole state of Connecticut, for court years 1992-93, 1993-94, and 1994-95. In addition, Lou sent electronic versions of several reports summarizing various aspects of the jury selection process and other pieces of documentation. The summary file for 1995-96 was transferred[11] in October 1996.

The August-October transfer failed to capture all the information that Lou and I had expected. In particular, the records for possibly delinquent jurors (no-shows: see the explanation below concerning the interpretation of the sa_date and my NS classification) were incomplete for summons dates after November 1994. We therefore arranged for a further transfer of data after the scheduled purge of the summons files the following January. This transfer did not take place until after my testimony at the Rodriguez trial in January 1997; the Penultimate version (dated 5 January 1997) of the present report was based on August-October data.

The two main January 1997 files were similar in format to the previous files, except for the addition of more information about cancellations and actual appearance dates. One file contained records for all jurors summoned for the 1995-96 court year, including those who had postponed service into the 1996-97 court year, and all jurors scheduled to serve in the 1995-96 court year, including those who had postponed service from the 1994-95 court year into the 1995-96 court year. (I believe the same file was also transmitted by some means to the State's Attorney's Office.) The second file contained the corresponding data for 1996-97 court year, but with some information regarding delinquent jurors complete only up to January 1997. Unexpectedly, the problem with the incomplete records from 1994-95 was still unsolved.

Finally, in February 1997, Lou was able to locate, in a remnant of an unpurged summons file for 1994-95, the missing data. I received the final transfer of data in late February.

My Final report is based on a combination of data from the August, October, January, and February transfers. It uses both the August-October data (as used for the Penultimate version of the report) and the new data from

[10] Unpublished practical work project report, Yale University Department of Statistics. A preliminary version of the report, *The juror summons system of the Hartford-New Britain Judicial District and its effects on Hispanic representation* was accidently circulated more widely than intended. The final version of the report includes several warnings about the tentative nature of its conclusions.

[11] The file for 1995-96 was generated in response to a special request from the Public defender's Office. It did not contain all the information typically contained in a summary file, because the procedure for determination of final delinquents was not completed until early 1997.

January-February. The piecemeal nature of the transfers placed some constraints on the ways I could organize the data for analysis. It also had the unfortunate effect of fragmenting my draft reports, into sections of varying vintage. For the benefit of those who are already familiar with the Penultimate version of the report, I have not completely revised my original descriptions of the JIS data. Instead, I have added a few remarks and tables to to indicate changes necessitated by the new data.

The bulk of the August-October 96 data came in four large computer files (j_92_93.ex1, . . . , j_95_96.ex1), each about 40MB in size, with one record per juror. Lou had extracted these records from the ***juror summary files*** for the four court years, omitting confidential information (such as Social Security numbers, or whole records for jurors excused on medical grounds) and internal-bookkeeping codes. From these four files I set aside 982787 records for jurors summoned to courts outside the HNB judicial district, then divided the remaining 321815 records according to the court year of the original summons date: records for court year 1992-93 going to the file HNB9293, and so on.

	HNB9192	HNB9293	HNB9394	HNB9495	HNB9596	nonHNB	total
j_92_93.ex1	3278	81017				227278	311573
j_93_94.ex1		7967	62027			231105	301099
j_94_95.ex1			5787	77711		254350	337848
j_95_96.ex1				6409	77619	270054	354082
total	3278	88984	67814	84120	77619	982787	1304602

The January-February 97 data replaced some of the August-October 96 records, bringing them up-to-date regarding delinquency. The following table summarizes most of the reorganizations that I carried out.

	HNB9394	HNB9495	HNB9596	HNB9697	confid	updated	total
HNB9394.old	67814						67814
HNB9495.old		77711				6409	84120
HNB9596.old			2			77617	77619
NEW	9	2489	10359	52036	4973		69866
UPDATED		6409	77617				84026
total	67823	86609	87978	52036	4973	84026	

The first row shows that all 67814 records from the old (August-October vintage) HNB9394 were transferred intact to the new version of the file. In addition, 9 new records (from postponements) were added, bringing the total number of HNB9394 records to 67823. The second row shows the fate of the 84120 records in the old HNB9495: 77711 of them were transferred intact into a new version of the HNB9495 file, and 6409 were updated. In addition, 2489 new records were added, bringing the total number of records in the new HNB9495 to 86609. The third row tells a similar story for the old HNB9596 file: most records were updated, and more were added. The fourth row shows the distribution of new records, from either of the January 97 files, across the updated HNB files. The fifth row merely repeats the column headed 'updated", in order to keep the accounting straight.

The totals along the bottom row of the table show the new sizes of the HNB files. The column headed 'confid' shows that the January 97 data contained 4973 records for jurors whose counterparts were omitted from the August-October data on grounds of confidentiality. The names and addresses of the (disqualified) jurors were omitted from the new confidential records, and so those records were of no direct use to me for either geocoding or surname matching. I wrote them to a separate file, which I then omitted from most further analysis. The HNB9293 file was unaffected by the new data.

FORMAT OF THE JIS RECORDS

FIELD	characters	short description
year	2	court year
court	4	court codes
date	6	date of summons, in form yymmdd
id	8	unique juror id
towncode	3	code for one of 169 CT towns
sourcecode	1	1 = DMV list, 2 = voter list, 3 = both
name	35	L,*last name*,1,*first name*,I,*initial*,S,*suffix*,*?*,*?*
address	30	street number and street name
townname	16	name of city or town
state	2	usually CT
zip	5	zipcode . . .
zip4	4	. . . plus 4
disq	2	disqualification code
sa_date	6	date juror's name sent to State's Attorney
postpone	6	postponed until
walkin	6	unscheduled appearance at court
= record	= 136	= complete record for a juror

The August-October summary files contain a single record per juror. A juror can appear in only one summary file (unless he or she becomes re-eligible for jury service two years after serving, in which case a new juror id is created). Each record consists of 136 ascii characters (terminated by a newline), interpretable from left to right as 16 fields in the table. For example, here is the record for a summons sent to me. I have inserted extra colon (:) characters to indicate the breaks between the fields (these colons do not appear in the original file), and I have folded the record across three lines.

```
96:NNH :960104:96023531:101:1:L,POLLARD,1,DAVID,I,B,,          :
171 WAYLAND ST                :NORTH HAVEN     :
CT:06473:      :01:000000:000000:000000
```

I was summoned to the New Haven (NNH) courthouse for 4 January 1996. My juror id was 96023531. I live in North Haven, which has towncode 101. My name was drawn from the DMV list of licensed drivers. I was disqualified on the grounds that I am not a US citizen. I did not postpone or turn up unexpectedly at the courthouse. My name has not been sent (and should not be sent) to the State's Attorney's Office.

The earlier versions of the data (for only 1994-95 and 1995-96, transmitted by Lou Sapia in March 1996) were different. The earlier 1994-95 summary file did not contain the towncode or sourcecode information; the 1995-96 data were extracted from a ***summons file***, which could contain multiple records for each juror. Jason Cross worked on the March 96 version of the data, just for the Hartford-New Britain judicial district.

Description of the data fields

The names for the fields, as given in the table, are close to, but not identical with, the names used in the documentation from JIS.

Fields: court and towncode The state of Connecticut is divided into twelve judicial districts, each of which contains several courthouses identified by court code. For example, the New Haven district has a courthouse in the city of New Haven (court code NNH) and a courthouse in the city of Meridan (court code NNI). The Hartford-New Britain district has five courthouses:

in Hartford city (HHD); in Manchester (H12M); in Enfield (H13W); in Bristol (H17B); and in New Britain (HHB).

Each of the 169 Connecticut cities and towns[12] (identified by the *towncode*, which ranges from from 001 for Andover through 169 for Woodstock) is allocated to a judicial district. For example, the city of Hartford (towncode 064) belongs to the Hartford-New Britain district,

> Jurors are required to serve only at court houses located within the district where they live. For example, a resident of West Haven should not serve at the NNH or NNI courthouses, because West Haven is one of the twelve towns that make up the Ansonia-Milford judicial district. Residents of Bethany, Branford, Cheshire, East Haven, Guilford, Hamden, Madison, Meriden, New Haven, North Branford, North Haven, Wallingford, and Woodbridge—the thirteen towns that make up the New Haven judicial district—can be summoned to NNH or NNI.

Fields: year, date, sa_date, postpone, walkin The court *year* need not coincide with the year of summons (the yy part of the *date* field). When summoned, jurors are permitted to pospone their service for up to a year. For example, suppose juror Jane Doe was initially summoned to appear on 1 April 1993 (part of the 1992-93 court year, which ran from 1 September 1992 through 31 August 1993). If she served on that date her record would have appeared in the 1992-93 summary file j_92_93.ex1, with *date* equal to 930401 and the last three fields filled with 0's. If instead she had postponed to 1 December 1993, her record would have appeared in the 1993-94 summary file j_93_94.ex1 with *date* = 930401, and *postpone* = 931201. If she had gotten mixed up, and had actually turned up at the courthouse on 19 November 1993, then the *walkin* date would have been 931119, again in the 1993-94 summary file.

If, however, Jane Doe had failed to turn up within a year of the original summons date, then she would have become delinquent. Her name would have been sent to the State's Attorney's Office on the date listed in the *sa_date* field. In practice, Jane would have had a little more than a year's grace, because the program to purge delinquents from the summons file is usually only run in March, June, September, and December.[13] Her *sa_date* might have been something like 940630, if the first purge after 1 April 94 were carried out on June 30th.

Delinquent jurors are sometimes referred to as 'no-shows'. For the HNB data, I have tagged delinquents by inserting an 'NS' as the disqualification code. (See the description of the *disq* field, below.)

In short, one might have to look in two files to find the summary record for a juror first summoned in any particular court year. A good fraction (perhaps around 10%) of records for court year Y might refer to individuals summoned in court year $Y - 1$; and jurors from year Y who postpone or who become delinquent might have records in the year $Y + 1$ summary file.

Field: sourcecode, address, townname, state, zip, zip4 Under the current law, the names of prospective jurors come from two sources: samples from the lists of registered voters for each of the 169 Connecticut towns, and the

[12] In Census terminology: *minor civil divisions*

[13] JIS will move to more frequent reporting of delinquents to the State's Attorney's Office soon.

Department of Motor vehicles (DMV) list of licensed motor vehicle operators. JIS tries to eliminate duplicates from the lists before making a random selection of names from the combined lists. Names appearing on both lists are given *sourcecode* 3. For such a record, the DMV address is entered into the *address* field; for sourcecodes 1 and 2, there is only one address to use.

Only the voter list contains the towncode. The DMV list contains only a city or town name. Thus, for sourcecode 1, the JIS folks have to assign a towncode, a task that can be much more complicated than just looking up a list of unique translations. A juror might have a mailing address for the DMV that is different from the town in which he is registerd to vote. It would be most awkward if the two addresses were not in the same judicial district—I don't know how JIS would resolve that difficulty, if it occured.

Other address ambiguities have less unfortunate effects. For example, Spring Glen is not one of the 169 official towns, but it does have a towncode of 504 in the first column of a list sent to me by Lou Sapia, with the code 'SPGN093NEW HAVEN' in the second column. The city of New Haven itself has towncode 093, which might suggest that Spring Glen should be treated as a part of New Haven city for the purposes of juror selection. Unfortunately, Spring Glen is actually part of the town of Hamden, which has towncode 062. Likewise, the geographical overlap of Mystic (=364) with both the towns of Stonington (=137) and Groton (=059) is unfortunate only if one is trying to calculate statistics by town.

The *townname* field presumably comes from the juror's address as derived from voter or DMV sources. I do not know how JIS handles all the possible conflicts of town name versus towncode. When I worked on the March 96 version of the JIS data for Hartford-New Britain, I had attempted to match town names with 'official names' (a name on the list of 169) by poring through the Hagstrom "Hartford County Atlas".

The *address* field, which is vital to the efforts both Jason Cross and I have made at ***geocoding***, seems to be unambiguous for about 80%–90% of the records. We both had to correct for obvious misspellings (such as 'Farmongton Ave' instead of 'Farmington Ave'), or parsing difficulties caused by misplacement of street numbers within addresses, missing spaces, box and apartment numbers, and so on. (See Appendix C.) The presence of zipcode and town names often helped to eliminate ambiguity.

Only summons sent to jurors outside Connecticut fail to have 'CT' in the *state* field, and those jurors never serve because nonresidency in the judicial district is a disqualification for jury service (see below).

Most juror records have only a 5-digit zipcode, in the *zip* field; the four digits of the '+4' part of the code are usually missing. I found the US Postal Service publication "Connecticut and Rhode Island Zip + 4 State Directory" helpful for resolving some street-name/zipcode/townname puzzles.

Fields: id, name The first two digits of the juror *id* seems always to coincide with the year of inital summons. The whole *id* identifies a juror uniquely. For Hartford-New Britain, except in one case (Smith), I usually found the last five digits to be enough to identify a juror. I have been suspicious of some of the spellings of *names*, suspecting transcription errors.

Field: disq The Statute lays out various qualifications required of a juror. Disqualification codes 01 through 18 are assigned by JIS to records of jurors who are disqualified for one of the statutory reasons or who could not be delivered a summons or some other material. Any juror who is excused by the

court (not the same as waiting all day at the court but not being assigned to a jury) is assigned code 99. I generated the 'NS' and 'OK' codes, which do not appear in the original summary files. For the 1995-96 and 1996-97 court years, I have used a "??" to indicate an undetermined disqualification status: the juror was summoned, but either had not appeared at the courthouse or had postponed service beyond the date (22 January 1997) at which the summons files were generated.

A juror who has a blank *disq* field in a summary file must either have turned up at the court house ('OK') or have been declared delinquent. I assigned the 'NS' disqulification code to those records with a blank *disq* field and an *sa_date* not equal to 000000. With the earlier data, from the 1995-96 summons file, Jason and I were misled by blank *disq* fields for some jurors, whose names were sent to the State's Attorney after we received the file. We also initially underestimated the number of jurors summoned in 1994-95, because records were still unpurged.

Disqualification codes (abbreviated descriptions)

01 = not US citizen
02 = not CT resident
03 = under 18
04 = found by judge to be 'impaired'
05 = convicted felon
06 = can't speak/understand English
07 = member of general assembly while in session
08 = older than 70, chooses not to serve
09 = physical/mental disability
10 = elected state offficial
11 = served in last 2 years
12 = extreme hardship
13 = summons undeliverable
14 = deceased
15 = moved out of judicial district
16 = moved out of state
17 = standby notice/handbook notice or other undeliverable
18 = received summons for this court year
99 = juror excused by court

NS = no-show (blank disq code and nonzero sa_date)
OK = confirmed for jury service (blank disq code & showed up)
?? = disqualification status not yet determined

Actually, a juror might be notified that a case is cancelled, so that he or she need not appear at the courthouse, but I have counted that possibility as 'showing up'.

Changes in the January-February 97 records

All fields except the zip4 were present, but in a slighly different order. In addition the record contained fields giving: (i) the first date (if any) on which a juror served; (ii) an indication of whether the juror had responded to the initial summons, confirming an appearance date; and (iii) an indication of

whether the juror's court appearance was cancelled. The new information let me determine no-shows up to 22 January 97.

To minimize the need for changes to my computer programs, I rearranged the records to match the old format, except that the information from (i)—(iii) was coded into the record position previously occupied by zip4, and the new date from (i) was appended to the record. In the new HNB files, records drawn from the January-February 97 data are therefore 6 characters longer than the August-October records.

5. Federal data

I retain this Section in my final report only because it was the subject of some argument during my January 97 testimony. The information in the Section plays no role in my analysis of the State system. I have added a remark near the end of the Section, for the benefit of anyone who is trying to make sense of the January testimony. The rest of the material is unchanged, except for the correction of a typo that had added thirty years to the age of the Margolis reports.

The juror selection for the U.S. District Court (District of Connecticut) draws from the same source lists as the State, with a similar method for combining the lists. The Federal judicial districts are fewer; they combine larger numbers of counties into each of the three districts. For example, the Hartford district consists of the counties of Hartford, Litchfield, Tolland, and Windham.

The Federal courts have smaller needs for jurors. The number of summons for the whole District is tiny compared to the number of summons sent out for the State system.

The Federal summons procedure differs slightly from the State's procedure. Jurors are sent an initial questionnaire to determine whether they are qualified for juror service. The questionnaire asks the potential juror to indicate both race and ethnicity[14]. Jurors can be disqualified, excused or exempted for a wider class of reasons than covered by the State's disqualifications. The precise details are spelled out in the *Second Restated Plan*[15]. In paraphrase:

Qualifications for jury service Disqualified if:

(1) not US citizen, over 18, or a resident of the judicial district

(2) unable to read, write and understand English sufficiently to satisfactorily complete the juror qualification form

(3) unable to speak English

(4) incapable of serving because of physical or mental infirmity

(5) charged or convicted of crime . . .

Automatic exemptions

[14] The Federal questionnaire was used as the model for the questionnaire described in Section 2.

[15] United States District Court, District of Connecticut, 23 November 1992, Second restated plan for random selection of grand and petit jurors pursuant to jury selection and service act of 1968 (as amended); modified 27 June 1994

(1) active member of US armed forces

(2) active fire, police ...

(3) public officers of executive, legislative, or judicial branches ...

Excuses (on individual request) Requests to be excused are granted for:

(1) person over 70

(2) ministers, priests, ...

(3) attorneys, physicians, dentists, and registered and licensed practical nurses, actively so engaged

(4) jury service in last two years

(5) schoolteachers

(6) care of children under 12 ... care of aged and infirm

(7) sole proprietors of businesses

(8) volunteer safety personnel ... for a public agency

Clearly there is no exact correspondence between the Federal and State requirments for jury service, which complicates direct comparison of the two systems. Nevertheless, the Federal data (as summarized in a series of reports and letters from Magistrate Judge Margolis to the Chief Judge of the District) do give give some relevant information: for the entire district, Hispanics comprised only 2.2% of the '1993-96 qualified wheel' (the pool of qualified jurors), compared with 5.07% amongst the over-18 population, according to the 1990 Census. Also: the distribution of reasons for exclusion for different racial and ethnic groups suggests that the underrepresentation of Hispanics on the qualified wheel cannot be explained solely in terms of language or citizenship disqualifications.

REMARK:*The final paragraphs in this Section were of some interest to the State's attorneys during my January 97 testimony. As I explained at that time, the whole Section was no longer of great significance to me; with hindsight, I should probably have excised it from the Penultimate draft, to avoid unnecessary discussion.*

My original motivation for including a Section on the Federal system was an argument made by the State during the King trial. They produced one of the Margolis reports, with the suggestion that the Federal experience showed that the low percentage of Hispanics answering the questionnaires was explained by language and citizenship disqualifications. Also, I had initially thought that I might be able to use the Federal data as another cross-check on the JIS data, because the Federal administrators actually keep and analyze the information on race and ethnicity that they request of prospective jurors.

Again as I explained during my testimony, I quickly decided that the Federal data did not support the State's suggestion. A combination of factors led me to abandon my study of the Federal system: my concerns about the quality of the data; an opinion from Richard Gayer, to the effect that the Federal system was not comparable with the State system; the difficulties I had in obtaining Federal data; and the small numbers of jurors involved in the Federal system. The data were also less useful to me because, by the nature of the way in which they had been collected, they gave no informa-

tion about potential jurors who had not responded to the Federal summons questionnaire.

My original intention (regarding the table taken from one of the Margolis reports) was merely to note that Federal experience did not support the suggestion made by the State during the King trial. I was not suggesting that it was proper to exclude the undeliverables and 'no-shows' from the base for calculating percentages before making comparisons between Hispanic and nonHispanics who made it to the master list.

Final paragraphs from Penultimate report:

The following table is taken from the Margolis Report, May 1996. It cross-classifies the jurors who returned the initial Federal questionnaire, during the period 1 October 1993 through 16 April 1996. The table excludes the undeliverable summonses and the non-responses. I was unable to determine the precise method used to partition the white, black, and Hispanic populations into disjoint groups. (Indeed, some inconsistencies in a tabulation attached to the Margolis Report for November 1993 caused me some concerns, which I have not yet been able to resolve.)

Category	White	Black	Hispanic	Other/Unknown
# Qualified	3495 (57.5%)	185 (70.9%)	85 (59.9%)	71 (15.0%)
# Disqualified	439 (7.2%)	26 (10.0%)	41 (28.9%)	149 (31.4%)
# Exempted	75 (1.2%)	2 (0.8%)	3 (2.1%)	124 (26.2%)
# Excused	2070 (34.1%)	48 (18.4%)	13 (9.2%)	130 (27.4%)
Total	6079 (100.0%)	261 (100.1%)	142 (100.1%)	474 (100.0%)

The category '# Qualified' corresponds roughly to my 'OK' category for the State system. It is striking that the overall qualification rate for Whites, Blacks, and Hispanics *who responded to the questionnaire* are comparable. The higher rate of disqualification for Hispanics is balanced by a lower rate of excuse. The 59.9% yield of qualified Hispanics (amongst those Hispanics who respond) leads me to suspect that the underrepresentation of Hispanics on the qualified wheel must be caused in large part by either underrepresentation on the original source lists, or overrepresentation as undeliverables or 'no-shows'.

6. Connecticut population trends

The main source of data about the population of Connecticut is the 1990 decennial Census of population and housing, much of which is available on CD-ROM or through online lookup services of the US Census Bureau on the World Wide Web[16]. The data are collected into various ***summary tape files***, identifiable by codes such as STF1A or STF3B. The different STF's have different levels of coverage (state, county, tract,...), different data tabulations, and are based on different census coverage (some tabulations are derived from the 'complete counts' from the 'short' Census form, and others are derived from sample data based on the 'long' Census form, which was filled out by only a sample of households).

The Census Bureau has also collected data on a less extensive scale since the decennial census, and has a program to provide estimates and projections that supplement the 1990 data. Much of the new data is also available via the World Wide Web.

[16] at the URL http://www.census.gov/cdrom/lookup

I have drawn most of my Census data from CD-rom versions of STF1A and STF3A, sometimes via the Census lookup service. As a crosscheck, I have compared various totals from the CD-roms with the corresponding figures in printed reports.[17]

County data

The population of Connecticut has decreased slowly since the 1990 Census. The Hartford County population has also decreased slowly. The fraction of Hispanic population has increased steadily, as clearly shown by the following table, which expresses estimated Hispanic populations (on 1 July of each year) as a percentage of the corresponding estimates of total population, for counties and the whole state.

ESTIMATED PERCENT HISPANIC POPULATION

% Hisp	total					over 20				
	90	91	92	93	94	90	91	92	93	94
Fairfield	8.61	8.89	9.13	9.42	9.63	7.25	7.53	7.74	7.98	8.19
Hartford	8.45	8.73	8.98	9.27	9.48	6.45	6.70	6.88	7.10	7.29
Litchfield	1.10	1.14	1.18	1.22	1.26	0.96	1.00	1.03	1.07	1.10
Middlesex	2.02	2.09	2.16	2.24	2.29	1.54	1.61	1.66	1.72	1.77
New Haven	6.38	6.59	6.78	7.01	7.16	5.03	5.23	5.38	5.55	5.70
New London	3.34	3.47	3.58	3.72	3.81	2.71	2.82	2.91	3.02	3.11
Tolland	1.73	1.79	1.85	1.91	1.96	1.52	1.58	1.63	1.69	1.73
Windham	4.18	4.33	4.45	4.61	4.70	3.15	3.27	3.36	3.48	3.57
CT	6.52	6.74	6.92	7.15	7.30	5.22	5.43	5.58	5.75	5.90

The percentages are taken from Section 1 of Appendix A, which is based on Census Bureau estimates downloaded from the World Wide Web[18]. Professor Thomas Steahr, a professional demographer from the University of Connecticut, has extended the estimation forward for Hartford County: 7.56% for July 1995 and 7.80% for July 1996.

The percentage over 20 would need to be increased very slightly to account for those persons 18 or 19 years of age. For Hartford County in 1990, the 1990 population counts were[19]:

	all	over 18	over 20
Hispanic	71575	43725	40891
whole pop	851783	659440	635829

We could inflate the percentage Hispanic in the over-20 populations by a factor of $(43725/40891)/(659440/635829) \approx 1.03$ to estimate the percentage Hispanic in the over-18 population for each year, which would add even more weight to the conclusion that the Hispanic proportion must be well above the 1990 figure.

As is true at the national level, the Hispanics are younger (lower median age; greater fractions of the population in the lower age brackets) than the general population, and the birth rates are higher[20]. Hispanics are expected[21] to make up over 10% of the US population by the year 2000.

[17] The most relevant report has been 1990 CPH-3-172B, *Population and Housing Characteristics for Census Tracts and Block Numbering Areas: Hartford-New Britain-Middletown, CT CMSA (Part); Hartford, CT PMSA.*

[18] http://www.census.gov/population/estimates/county/casrh/9094ct.dat

[19] Source: STF1A, tables P11, P13

[20] Bureau of the Census Statistical Brief SB/95-25

[21] Statistical Abstracts of the United States 1993, Table 20

The rapid growth in Hispanic population since the 1990 Census creates some difficulties in the interpretation of the jury data. One might attempt to develop a demographic estimate of population through to the present, or one might crudely treat the over-12 age groups for 1990 as a surrogate for the over-18 age groups in 1996. Or—as I will do—one might merely regard all assertions about the current underrepresentation of Hispanics as understatements of the problem.

Projections are complicated because little directly relevant data below the county level has been collected since the 1990 Census. The situation for the HNB judicial district is further complicated by the fact that is does not exactly coincide with Hartford County—the judicial district includes the Litchfield County town of Plymouth and it excludes Hartland. The differences have only a small effect on most calculations, because neither town has a large population and only very small fractions of those populations are minorities.

Town data

The 1990 population counts for the towns of the HNB judicial district are easier to digest when most of the outlying towns are grouped into a single category 'otherHNB'. (See Section 2 in Appendix A for the complete counts.) The first table makes clear the concentration of the minority population of the whole HNB judicial district in a few towns. The extreme concentration is even more obvious when the minority populations are expressed as percentages of the total population for each town, as in the second table.

1990 POPULATION DISTRIBUTION ACROSS HNB

Town	all		white		black		hispanic	
	over18	under18	over18	under18	over18	under18	over18	under18
Bloomfield	2.4	1.9	1.7	0.9	9.7	8.3	0.9	0.7
East Hartford	6.1	5.1	6.3	5.2	4.7	5.2	4.6	3.5
Hartford	15.2	19.7	8.1	6.4	61.6	63.8	59.8	64.3
New Britain	8.9	8.2	8.9	7.3	6.5	6.8	16.5	18.2
otherHNB	67.4	65.1	75.1	80.1	17.5	16.0	18.1	13.3
all HNB	100	100	100	100	100	100	100	100

1990 PERCENTAGE MINORITY (OVER 18) FOR SELECTED TOWNS

% hisp	hisp	nonhisp	total
Bloomfield	2.5	97.5	100
East Hartford	5.0	95.0	100
Hartford	25.9	74.1	100
New Britain	12.1	87.9	100
otherHNB	1.8	98.2	100
all HNB	6.6	93.4	100

%black	black	nonblack	total
Bloomfield	37.3	62.7	100
East Hartford	7.0	93.0	100
Hartford	36.9	63.1	100
New Britain	6.6	93.4	100
otherHNB	2.4	97.6	100
all HNB	9.1	90.9	100

The black and Hispanic populations were highly concentrated near Hartford town, and, to a lesser extent, in the town of New Britain. Averages taken over the whole judicial district tend to disguise any effects on the minority populations. If there are any factors that systematically disadvantage minorities, they should become apparent from a closer examination of the juror data for the towns of Hartford and New Britain (and maybe East Hartford and Bloomfield as well).

It would be incorrect to simply add the percentages for Hispanics and blacks from the previous pair of tables in order to determine percentage minority for the towns. Hispanic origin is not currently a racial category; per-

sons of Hispanic origin can be of any race. For example, for the total population of Hartford County, the next table shows the 1990 distribution of Hispanics and non-Hispanics across the five standard racial categories.

	White	Black	Amerind	Asian	Other	total
Hispanic	39.6	7.5	0.3	0.6	52.1	100
non-Hispanic	87.5	10.5	0.2	1.7	0.2	100

If we merely added the numbers of Hispanics to the numbers of blacks in Hartford County we would be double-counting the 7.5% of Hispanics who were black.

7. Source lists

There are two key steps involved in creating a pool of potential jurors that is a "fair cross section" of the community. First, lists of names need to be obtained whose combined coverage of the eligible population is as complete as realistically possible. Then the required number of names must be selected at random from the combined list, with each name on that list having an equal probability of being chosen. Of course it is important that duplicate names be weeded out from the combined list, as far as possible, for otherwise a person whose name appeared more than once would have a higher chance of being selected than a person whose name appeared only once.

Currently JIS constructs its master list of potential jurors for a court year from only two sources: voter registrations and motor vehicle licenses (obtained from DMV, the Department of Motor vehicles). The projected needs of the courts determine the size of the master list. During the court year, jurors are summoned in random order—based on a juror id assigned at random to the names on the master list—in response to requests from the courts for jurors.

8. Hartford-New Britain judicial district

I decided to exclude the HNB9192 file from all my analysis of the data for HNB, because it contains data for only a frgment of a court year. Also, I can present only partial analysis for HNB9596 and HNB9697, because those files contain records for jurors whose disqualification status is not yet settled (the '??' code).

The first table gives the overall breakdown of disqualifications by court-tyear of summons. The figures for 1995-96 and 1996-97 are subject to change, because of the '??'.

DISQUALIFICATIONS BY YEAR OF SUMMONS

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
HNB9293	2113	3100	1		333	1787	1	9751		17	1066	3758	10523	1090	1282	413	912	233	2586		4145	45873	88984
HNB9394	1598	2688			230	1320		7483		11	2004	3013	9072	861	957	69	932	123	1573		2938	32951	67823
HNB9495	1988	3339			278	1788	4	9832		13	2831	5603	12297	1109	1244	290	1469	165	1585		3678	39096	86609
HNB9596	1922	3314	1		291	1981		10159		16	3042	5102	11154	1326	1151	401	1661	149	1093	4483	1865	38867	87978
HNB9697	979	1325	2		110	934	3	5117		4	1998	1208	4629	497	492	201	698	35	286	20754		12764	52036
collapse		xjd	rest		rest		rest			rest	rest			rest	xjd	xjd		rest	rest				

There were no code 04 (found by judge to be 'impaired') or code 09 (physical/mental disability) contained in the earlier summons files, because

records for those individuals were removed by JIS for confidentiality reasons. As explained in Section 4, I also excluded the code 04 and 09 from the later summons files. In theory persons under 18 are screened out of the DMV lists at a preliminary stage of the JIS selection procedure, and they should not even be on the voter lists, but a few code 03 slipped through.

The large proportion of '??' in the HNB9697 file confuses the interpretation of the NS and OK codes. If my experience with HNB9495 is any guide, many of the '??' will become NS or OK, with only a sprinkling of other types of disqualifications. I will therefore omit the 1996-97 data from subsequent tabulations (but the full counts do appear in Appendix A).

To conserve on space in tabulations within the body of this report, henceforth I will collapse the counts for the three codes

02 = not CT resident
15 = moved out of judicial district,
16 = moved out of state

into a single category 'xjd', and for the eight codes

03 = under 18
05 = convicted felon
07 = member of general assembly while in session
10 = elected state offficial
11 = served in last 2 years
14 = deceased
18 = received summons for this court year
99 = juror excused by court

into a single category 'rest'. Codes 04 and 09 will be omitted altogether. The last line of the table indicates the two collapsed categories. Section 4 of Appendix A gives the full counts for the seventeen disqualification codes plus the NS and OK.

Patterns across time

Cross-tabulations of the counts of disqualifications by month are given in Section 3 of Appendix A, from which the following table is derived. The months run from 9209 (= September 1993) for the first table through 9608 (= August 1996) for the last table. The numbers in the bodies of the tables give the percentages of summonses for each month for selected disqualification codes.

HNB: PERCENTAGE DISQUALIFICATIONS BY MONTH

[HNB9293]	01	06	08	12	13	17	NS	OK	xjd	rest	total
9209	2	2	11	4	10	1	5	53	5	6	100
9210	2	2	11	5	10	1	4	53	5	6	100
9211	2	2	11	4	11	1	5	52	5	6	100
9212	2	2	11	5	11	1	4	53	5	6	100
9301	3	2	11	4	11	1	5	52	6	6	100
9302	2	2	11	4	11	1	4	52	5	6	100
9303	2	2	11	4	12	1	4	52	5	6	100
9304	2	2	12	4	13	1	5	49	6	6	100
9305	3	2	11	4	13	1	5	49	5	7	100
9306	2	2	11	4	14	1	4	50	6	6	100
9307	2	2	10	4	14	1	4	50	5	5	100
9308	2	2	11	4	15	2	4	50	5	5	100
total	2	2	11	4	12	1	5	52	5	6	100

[HNB9394]	01	06	08	12	13	17	NS	OK	xjd	rest	total
9309	2	2	10	4	10	1	4	53	6	8	100
9310	3	2	11	5	11	1	5	48	6	8	100
9311	2	2	11	4	12	1	5	48	6	8	100
9312	3	2	11	5	12	1	5	49	5	7	100
9401	3	2	11	4	12	1	4	50	5	7	100
9402	2	2	11	4	12	1	4	51	5	7	100
9403	2	2	11	4	14	1	4	50	5	6	100
9404	3	2	11	4	14	1	5	47	6	8	100
9405	2	2	11	4	15	1	5	47	5	7	100
9406	2	2	11	5	16	1	4	47	5	6	100
9407	2	2	12	4	16	2	4	46	6	6	100
9408	2	2	11	4	16	2	4	46	6	7	100
total	2	2	11	4	13	1	4	49	5	7	100

[HNB9495]	01	06	08	12	13	17	NS	OK	xjd	rest	total
9409	3	2	11	7	12	1	4	48	6	7	100
9410	3	2	11	7	12	2	5	44	5	8	100
9411	2	2	11	7	12	1	5	45	5	8	100
9412	2	2	11	7	13	2	5	45	5	7	100
9501	2	2	12	7	13	2	4	45	5	7	100
9502	2	2	12	6	14	1	4	47	5	6	100
9503	2	2	12	6	14	2	4	44	6	7	100
9504	2	2	11	7	15	2	4	44	6	7	100
9505	2	2	11	6	15	2	4	43	6	7	100
9506	2	2	11	6	17	2	3	45	6	6	100
9507	2	2	11	6	16	2	4	44	5	7	100
9508	2	2	12	6	16	2	3	47	6	6	100
total	2	2	11	6	14	2	4	45	6	7	100

[HNB9596]	01	06	08	12	13	17	??	NS	OK	xjd	rest	total
9509	3	2	12	6	10	2		6	46	6	8	100
9510	2	3	12	6	9	3		5	46	6	8	100
9511	2	2	11	6	11	2		5	46	6	8	100
9512	2	2	12	6	11	2		5	46	6	7	100
9601	2	2	12	6	13	2	3	3	45	5	7	100
9602	2	2	11	6	13	2	7		45	6	7	100
9603	2	2	11	6	13	2	7		44	6	6	100
9604	2	2	12	6	14	2	8		43	5	6	100
9605	2	2	12	6	14	2	9		42	5	7	100
9606	2	2	11	6	15	2	9		43	5	6	100
9607	2	2	12	5	15	2	12		39	5	6	100
9608	2	2	12	5	16	2	9		43	5	5	100
total	2	2	12	6	13	2	5	2	44	6	7	100

Notice that the undeliverable rate increases fairly steadily through each court year. Addresses go 'stale', making it harder for the Postal Service to deliver a summons to the addressee. The no-show rate seems to stay fairly constant, at around 5% of all summonses sent out.

Unless a summons is returned as undeliverable, there is no way of distinguishing a no-show from an undeliverable, a fact acknowledge by Richard Gayer of Jury Administration. As attested[22] by Attorney Angela Macchiarulo, who has responsiblility at the Office of the Chief State's Attorney for the collection of fines from the no-shows, the Postal Service is unable to deliver follow-up letters to many of the no-show addresses. Admittedly this failure occurs at least a year after the original summons date, but it does cast further doubt on the distinction between undeliverables and no-shows.

Patterns across towns

The pattern of disqualifications is not uniform across all the towns in the judicial district, as shown by the next four tables. The first pair of tables gives the percentage breakdown within selected towns. For both years, Hartford town has a very low yield of qualified jurors—just over 30% of mailed summonses return a qualified juror—and a very high rate of undeliverable summonses. And the Hartford no-show rate runs at about three times the 'other HNB' group. New Britain is slightly less extreme, but still rather different from the 'other HNB' group.

[HNB9293]	01	06	08	12	13	17	NS	OK	xjd	rest	total
HARTFORD	4	5	8	2	27	3	12	35		5	100
NEW BRITAIN	5	5	15	3	15	1	6	44		5	100
nonHNB									97	2	100
otherHNB	2	1	12	5	9	1	3	60	1	7	100
total	2	2	11	4	12	1	5	52	5	6	100

[HNB9394]	01	06	08	12	13	17	NS	OK	xjd	rest	total
HARTFORD	3	5	8	2	31	3	10	33		4	100
NEW BRITAIN	4	5	15	3	17	2	6	42		7	100
nonHNB									98	2	100
otherHNB	2	1	12	5	10	1	3	57		8	100
total	2	2	11	4	13	1	4	49	5	7	100

[HNB9495]	01	06	08	12	13	17	NS	OK	xjd	rest	total
HARTFORD	3	5	7	3	36	4	10	27		4	100
NEW BRITAIN	4	6	17	5	16	2	5	37		7	100
nonHNB									98	2	100
otherHNB	2	1	13	8	10	1	3	53		8	100
total	2	2	11	6	14	2	4	45	6	7	100

[HNB9596]	01	06	08	12	13	17	??	NS	OK	xjd	rest	total
HARTFORD	3	5	8	2	28	5	8	5	31		4	100
NEW BRITAIN	4	6	17	5	15	2	6	2	37		6	100
nonHNB										98	2	100
otherHNB	2	1	13	7	10	1	5	2	51	1	8	100
total	2	2	12	6	13	2	5	2	44	6	7	100

9. Hispanic surname matching

Surnames have been used by the Census Bureau since 1950 to identify His-

[22] Page 120 of the transcript of the proceedings of State vs. Ortiz, October 1995, CR 14 448783

panics[23]. The method of estimation was refined by Word and Perkins[24] by means of data derived from the the 1990 Post-enumeration sample (PES)[25] leading to the production of an extensive list of names classified as heavily (category 01··), generally (category 02··), moderately (category 03··), occasionally (category 04··), and rarely (category 5· · ·) Hispanic. In addition, they listed both the number of persons in the PES with each surname and the number of those persons identifying themselves as Hispanic.

From the Word/Perkins list I was able to estimate the probability that a person with a given surname is Hispanic. For example, Word and Perkins rated the surname "Garcia" as heavily Hispanic. For the whole PES 94.5% (3881 out of 4106) of the "housholders" surnamed Garcia identified themselves as Hispanic, and the figure rose to 95.5% (3379 out of 3541) when calculated for householders in one of the 11 states identified as having large numbers of Hispanics.[26]For any surname in Word and Perkins' "Rarely Hispanic Surname" category (any of their category codes that start with a 5) I took the Hispanic probability as zero. For example, even though about 1% of persons surnames 'Smith' in the PES sample from the 11 states identified themselves as Hispanic, I take all Smiths as nonHispanic because of the Word and Perkins category code of 5500 for SMITH.

Several methods have been suggested for estimating numbers of Hispanics on any list using surname matching. One method, which on statistical grounds should be less accurate, attempts to classify every name as either 'Hispanic' or 'nonHispanic'. One then counts the number of 'Hispanic' names on the list. The method has the disadvantage that it treats a name as completely Hispanic or completely nonHispanic; it would give equal weight to a Garcia and a Silva (68.9% of householders in the 11 states surnamed Silva identified themselves as Hispanic). Another method, which I have adopted, would count each Garcia on the list as contibuting 0.955 to the Hispanic counts, and each Silva as contributing 0.689.[27]

As an example, consider a very hypothetical population made up of 50 persons named Garcia, 30 persons named Silva and 100 persons named Smith. I would estimate the number of Hispanics in that population as

$$(0.955 \times 50) + (0.689 \times 30) + (0 \times 100) \approx 68.4$$

By contrast, if I had counted every Garcia and Silva as Hispanic, and every Smith as nonHispanic, my estimate would have been 80.

---

[23] Passel and Word, "Constructing the list of Spanish surnames for the 1980 census: an application of Bayes' theorem", Technical report from the US Bureau of the Census, April 1980.

[24] "Building a Spanish surname list for the 1990's—a new approach to an old problem", US Bureau of the Census Population Division Technical working paper #13, March 1996. Available for download from the Bureau of the Census WWW site, http://www.census.gov

[25] More precisely, they worked from a list of 5,609,592 records taken from the Spanish Origin sample, which was larger than the PES sample.

[26] Connecticut is one of 11 states. The others are: Arizona, California, Colorado, Florida, Illinois, New Jersey, New Mexico, New York, Pennsylvania, and Texas. David Word advised me to use the figures from those 11 states for surname matching in HNB judicial district.

[27] My understanding of some of the pitfalls in surname matching benefitted from discussions with Laura McKinney, a graduate student in the Yale Statistics Department. I also relied on advice from David Word.

Females who change their surnames after marriage create some difficulties for any method of identification based on surnames. An Hispanic Ms. Garcia who married a Mr. Smith would not be counted as Hispanic; a nonHispanic Ms. Smith who married a Mr. Garcia would be miscounted as Hispanic. Word and Perkins recognized the problem in their definition of "housholder" by limiting it to "male or never married female householders plus any other male or never married female in the household not related to the householder". Tom Steahr has pointed out to me that this approach creates a potential systematic error, because Hispanic householders are not exactly the same population as Hispanic adults. My sampling experiments with the questionnaire data from Section 2 suggest that the systematic error is not large.

My estimates of the proportions of Hispanics in the JIS files appear in the Appendix C, where they are compared with estimates based on a completely different method, and in the first Section of the report.

10. Geocoding

In principle, to geocode an address list one merely has to locate each address from the list on a detailed street map. Indeed, for small lists, geocoding can actually be done with a wall-map and some push-pins.

In principle, if we had enough pushpins, and patience, we could create a gigantic wall-map showing the address to which each juror summons was sent. We could also use different colored pins for each disqualification code to get a representation of the distribution of summonses and disqualifications across the whole judicial district. Of course we would have some difficulty with some addresses (such as the mail rooms of large housing complexes or student dorms) to which multiple summons were sent, or misspelled addresses, or inconsistent addresses (such as a zipcode incompatible with a town name).

In practice, it would be impossible to carry out the a pin-pushing project for all the summons in the HNB judicicial district (nearly 90000 pins would be needed for the 1992-93 court year alone) or even for only the City of Hartford (over 10000 pins). Something equivalent must be done by computer.

If we could geocode all the addresses in the JIS files, we could identify regions of the judicial district where various types of disqualifications were overrepresented compared to the population for the region. If we were looking for racial or ethnic patterns in the disqualifications, we would need to choose regions that are small enough to capture the variation of race/ethnicity across the district, but not so small that the patterns in the data were dominated by random fluctations.

I chose to work with two types of region: individual towns, and Census tracts. The 1990 Census tabulations contain very detailed information about both towns and tracts. At the town level, I could work directly with codes in the JIS files to allocate juror records to towns. At the tract level I had to geocode using the address, towncode, and zipcode fields.

Instead of the giant wall-map, the computer uses an electronic TIGER[28] database of street segments, constructed by the Census Bureau. TIGER approximates a street (or river, or town boundary, or ...) as a chain of straight line segements. The latitude and longitude of the endpoints of each segment is recorded to great accuracy. For each street segment, TIGER also records

[28] Topologically Integrated Geographic Encoding and Referencing system

information such as: the street name, the range of address numbers on both sides of the street, zipcodes and tract numbers for both sides of the street (where defined), and many other identifying codes. In principle, one can match house addresses to points on individual street segments, and thereby determine the latitude and longitude of the house with great accuracy. In particular, each correctly matched address is then located within a uniquely determined Census tract.

The Census Bureau has simplified the geocoding task slightly by producing, from the TIGER database, a more concise CTSI[29] database, on CD-ROM. Roughly speaking, the CTSI records correspond to chains of TIGER segments that share the same street name, zipcode, and Census tract (for both sides of the street).

For my first attempt at geocoding the JIS addresses to tracts I used a commercial mapping program, MapInfo[30], which works from a slightly enhanced version of the TIGER database. Unfortunately, MapInfo was unable to handle satisfactorily the JIS data. The large number of records kept the program grinding away for a long time. Also, the algorithms used by MapInfo for automatic correction of small imperfections of addresses (many variations on spelling errors, addresses in nonstandard form, ...) gave it a very low matching rate; and, of course, the MapInfo interactive mode—the recommended method for handling ambiguous cases—was completely out of the question for the huge numbers of records in the JIS data sets.

I had much more geocoding success with an improved form of the matching algorithm that I wrote myself, based on the ideas documented in the MapInfo manual. I tried to match juror records—using the *towncode*, *address*, *and zipcode* fields—to addresses in the TIGER/CTSI database. (The details of the method and more complete listings are given in Appendix C.) In short, my method attempts to match five components of the address (house number, street name, street type, prefix, and direction suffix).

With summonses geocoded to tracts, I was able to estimate the proportion of disqualifications for each minority group, using the Census data for each tract. In essence, if geocoding implies that a particular tract receives N summonses, and if the 1990 Census data lists a fraction h of the over-18 population of that tract as Hispanic, then one could estimate the number of summonses sent to Hispanics in that tract as $h \times N$. One sums over all tracts in a particular region to estimate the total number of summonses sent to Hispanics in the region.

Appendix C describes in more detail how I constructed the geocoding estimates, based on more refined estimates of the minority proportions in each tract. My results are summarized in the first Section of the report and in Appendix C.

The geocoding method suffers from the disadvantage that it must work with estimates of minority populations derived from the 1990 Cenus. I would expect the estimates of total counts to increase over time if up-to-date minority proportions could be used. Geocoding would also suffer from the undercounts of minority populations that are known to have occurred with the 1990 Census.[31] In contrast, the SSL estimates (based on surname matching)

[29] TIGER/Census Tract Street Index version 2, issued December 1994; covering states CT, MA, ...; CD-CTSI-V2-01.

[30] sold by the Mapinfo Corporation, Troy New York

[31] See the special section on Census undercount in the September 1993 volume of the Journal of the American Statistical Association.

increase over time, as would be expected if the Hispanic population were increasing.

Several other small points to be aware of when interpreting my geocoding estimates are explained in Appendix C. For example, I have, deliberately, slightly underestimated the Hispanic OK count in order to get an upper bound for the language disqualifications.

References

Finkelstein, M. O. (1966), 'The application of statistical decision theory to the jury discrimination cases', *Harvard Law Review* **80**, 338–376.

Gerber, E. & de la Puente, M. (1996), The development and cognitive testing of race and ethnic origin questions for the year 2000 decennial census, Technical report, US Bureau of the Census. (Paper presented at the Bureau of the Census 1996 Annual Research Conference, March 17–19, Arlington, Virginia. Published in the proceedings of the Annual Research Conference.).

Hansen, M. L., Hurwitz, W. N. & Madow, W. G. (1953), *Sample Survey Methods and Theory*, Wiley. (In two volumes).

Kairys, D., Kadane, J. B. & Lehoczky, J. P. (1977), 'Jury representativeness: a mandate for multiple source lists', *California Law Review* **65**, 776–827.

Munsterman, G. T. (1996), *Jury System Management*, Court Management Library Series, National Center for State Courts, Williamsburg Virginia.

Appendix

Detailed listings

1. Connecticut population estimates and projections

The first three tables are derived from Bureau of the Census table PE-48: Estimates of the Population of Counties by Age, Sex and Race/Hispanic Origin. (Estimated populations for 1 July 1990, 1 July 1991, 1 July 1992, 1 July 1993; and 1 July 1994.)[32]

all population	total					20+				
	90	91	92	93	94	90	91	92	93	94
CT	3289105	3290747	3279331	3278038	3275276	2443622	2444059	2431134	2421461	2412357
Fairfield	827925	828874	827929	828816	829791	619415	620002	618404	617088	616325
Hartford	851885	851624	846947	843766	839616	635258	634698	629979	625229	620314
Litchfield	174489	175718	176731	177797	178528	129515	130453	130984	131383	131574
Middlesex	143465	144033	144770	145667	146689	107794	108245	108619	108953	109412
New Haven	804599	803990	801996	799499	796477	597935	597219	594663	590641	586673
New London	255176	253931	247887	248838	249587	186952	185924	181023	180966	180842
Tolland	128905	129349	129849	130209	130899	93756	94145	94293	94187	94311
Windham	102661	103228	103222	103446	103689	72997	73373	73169	73014	72906
CT	3289105	3290747	3279331	3278038	3275276	2443622	2444059	2431134	2421461	2412357

Hispanic population	total					20+				
	90	91	92	93	94	90	91	92	93	94
Fairfield	71251	73710	75614	78083	79887	44904	46678	47850	49245	50472
Hartford	71969	74387	76027	78259	79598	40984	42526	43371	44415	45221
Litchfield	1923	2008	2081	2176	2246	1239	1301	1350	1404	1452
Middlesex	2900	3016	3124	3259	3364	1665	1743	1803	1873	1937
New Haven	51306	52998	54356	56008	57035	30088	31220	31968	32786	33419
New London	8512	8807	8870	9253	9520	5063	5251	5271	5460	5618
Tolland	2231	2319	2396	2493	2568	1427	1491	1538	1591	1635
Windham	4290	4466	4591	4764	4878	2297	2402	2462	2541	2601
CT	214382	221711	227059	234295	239096	127667	132612	135613	139315	142355

pct Hisp	total					20+				
	90	91	92	93	94	90	91	92	93	94
Fairfield	8.61	8.89	9.13	9.42	9.63	7.25	7.53	7.74	7.98	8.19
Hartford	8.45	8.73	8.98	9.27	9.48	6.45	6.70	6.88	7.10	7.29
Litchfield	1.10	1.14	1.18	1.22	1.26	0.96	1.00	1.03	1.07	1.10
Middlesex	2.02	2.09	2.16	2.24	2.29	1.54	1.61	1.66	1.72	1.77
New Haven	6.38	6.59	6.78	7.01	7.16	5.03	5.23	5.38	5.55	5.70
New London	3.34	3.47	3.58	3.72	3.81	2.71	2.82	2.91	3.02	3.11
Tolland	1.73	1.79	1.85	1.91	1.96	1.52	1.58	1.63	1.69	1.73
Windham	4.18	4.33	4.45	4.61	4.70	3.15	3.27	3.36	3.48	3.57
CT	6.52	6.74	6.92	7.15	7.30	5.22	5.43	5.58	5.75	5.90

The changes in the Hispanic population since 1990 are partly explained by the differences is population distributions across ages: the Hispanic population is more concentrated in the younger age groups.

[32] Source: http://www.census.gov/population/estimates/county/casrh/9094ct.dat

CONNECTICUT PROJECTIONS FOR 1996 ELECTION

[Nov 1996]	all over 18	Male					Female				
		18+	18–24	25–44	45–64	65+	18+	18–24	25–44	45–64	65+
White	2,223	1,066	111	462	308	185	1,157	106	465	323	261
Black	199	91	16	46	22	8	108	16	52	28	13
Other	46	23	4	12	6	1	23	4	13	6	1
Hispanic	173	84	16	45	17	6	89	16	47	19	8
Not Hispanic	2,295	1,096	114	475	319	188	1,199	109	482	338	267
CT	2,468	1,180	130	520	335	195	1,288	126	529	357	275

Projections of the Population (000's) of Voting Age by Sex, Race, and Selected Ages[33]

2. HNB population by town, 1990

Town	all		white		black		hispanic		1990 pop	% HNB
	over18	under18	over18	under18	over18	under18	over18	under18		
Avon	10916	3021	10646	2906	89	40	92	26	13937	1.62
Berlin	12963	3824	12772	3720	59	25	167	57	16787	1.95
Bloomfield	15775	3708	9496	1339	5886	2198	396	194	19483	2.26
Bristol	47239	13401	45642	12600	833	430	1042	610	60640	7.04
Burlington	5029	1997	4966	1971	29	11	32	24	7026	0.82
Canton	6369	1899	6285	1860	34	15	61	28	8268	0.96
East Granby	3297	1005	3228	986	50	13	34	17	4302	0.50
East Hartford	40578	9874	36035	7756	2856	1379	2022	984	50452	5.85
East Windsor	7930	2151	7545	1965	248	94	98	66	10081	1.17
Enfield	35200	10332	33622	9960	1050	158	833	206	45532	5.28
Farmington	16238	4370	15694	4136	192	78	168	72	20608	2.39
Glastonbury	21417	6484	20687	6092	174	85	350	212	27901	3.24
Granby	6993	2376	6898	2326	37	11	55	33	9369	1.09
Hartford	101349	38390	46382	9487	37360	16978	26207	17930	139739	16.22
Manchester	40500	11118	38302	9960	1297	708	804	425	51618	5.99
Marlborough	3969	1566	3901	1540	40	12	40	28	5535	0.64
New Britain	59553	15938	50818	10787	3920	1803	7223	5061	75491	8.76
Newington	23571	5637	22873	5324	296	117	444	168	29208	3.39
Plainville	13779	3613	13279	3407	299	118	272	99	17392	2.02
Plymouth	8909	2913	8821	2875	30	14	78	33	11822	1.37
Rocky Hill	13636	2918	13020	2742	351	80	243	83	16554	1.92
Simsbury	16386	5637	15990	5442	136	48	170	84	22023	2.56
South Windsor	16650	5440	15811	5079	393	116	249	121	22090	2.56
Southington	29392	9126	28876	8860	253	97	347	161	38518	4.47
Suffield	8642	2785	8409	2664	130	47	66	32	11427	1.33
West Hartford	48391	11719	45893	10600	910	400	1276	615	60110	6.98
Wethersfield	21043	4608	20619	4402	208	85	288	134	25651	2.98
Windsor	21378	6439	17321	4660	3364	1428	621	332	27817	3.23
Windsor Locks	9922	2436	9579	2305	141	44	119	44	12358	1.43
total	667014	194725	573410	147751	60665	26632	43797	27879	861739	100

Source: Data on CD-ROM from US Census Bureau, STF1A. Counts at summary level 060 summed over tables P011, P012, P013 for age categories 17 years or less and 18 years or more.

Hispanics were 6.57% of the over-18 population, and 8.32% of the total population. Blacks were 9.09% of the over-18 population, and 10.1% of the total population.

3. HNB disqualifications by month: 1992-93 through March 1996-97

The counts in the bodies of the tables give the total number of juror summons for each particular combination of disqualification code and month of summons, for September 1992 through

[33] Source: http://www.census.gov/population/socdemo/voting/proj/votepg2.asc

August 1996. For example, in the first row of the first table, for 1992-93, out of the total of 9980 jurors who were summoned in month 9209 (= September 1992), there were: 1066 disqualified under code 08 (= older than 70, chooses not to serve); 507 who were eventually classified as delinquent (NS); and 5318 who turned up at the court.[34]

[HNB9293]	01	02	03	05	06	07	08	10	11	12	13	14	15	16	17	18	99	NS	OK	total
9209	249	293		32	213		1066	2	162	433	962	97	130	104	78	11	323	507	5318	9980
9210	238	305		52	208		1151	1	149	497	1079	121	153	115	83	15	319	445	5556	10487
9211	205	246	1	33	195		886	1	92	364	894	78	115	73	60	10	295	420	4335	8303
9212	178	228		26	140		805	1	76	339	818	85	101	46	66	20	228	323	3884	7364
9301	157	229		18	120		699	2	55	245	652	69	79	34	67	12	182	325	3178	6123
9302	154	246		21	127	1	764	4	76	281	754	84	111	6	73	24	204	283	3474	6687
9303	202	349		38	192		1035	1	111	379	1129	121	132	10	81	23	261	417	4796	9277
9304	149	271		34	126		788	1	85	277	865	89	110	2	101	23	179	356	3382	6838
9305	163	236		26	112		693		73	253	840	95	91	7	85	20	201	322	3134	6351
9306	138	254		24	134		678		81	253	869	79	100	5	72	25	165	265	3133	6275
9307	146	226		11	115		601	2	62	236	848	95	90	5	62	24	125	256	2957	5861
9308	134	217		18	105		585	2	44	201	813	77	70	6	84	26	104	226	2726	5438
total	2113	3100	1	333	1787	1	9751	17	1066	3758	10523	1090	1282	413	912	233	2586	4145	45873	88984

[HNB9394]	01	02	05	06	08	10	11	12	13	14	15	16	17	18	99	NS	OK	total
9309	110	216	17	104	518		195	216	480	38	56	2	71	5	126	177	2634	4965
9310	163	243	23	111	588	2	197	261	625	58	70	7	72	3	140	276	2644	5483
9311	118	211	20	110	564		183	227	586	62	70	5	75	6	147	241	2441	5066
9312	112	155	19	92	473	1	156	217	511	40	69	2	62	4	110	216	2188	4427
9401	132	206	7	101	595	1	169	224	645	49	75	4	71	7	128	214	2649	5277
9402	117	211	21	104	594	1	152	224	665	72	67	3	63	7	123	209	2744	5377
9403	175	262	20	139	816	1	213	329	1027	93	118	6	80	17	138	291	3731	7456
9404	167	266	31	118	692	2	184	270	915	101	92	8	87	19	159	323	2989	6423
9405	151	224	21	118	695	1	153	253	903	97	90	3	83	13	127	276	2857	6065
9406	134	226	26	114	695	1	142	301	967	69	90	7	84	9	128	232	2916	6141
9407	107	247	11	106	665	1	126	245	880	90	81	11	90	18	104	250	2593	5625
9408	112	221	14	103	588		134	246	868	92	79	11	94	15	143	233	2565	5518
total	1598	2688	230	1320	7483	11	2004	3013	9072	861	957	69	932	123	1573	2938	32951	67823

[HNB9495]	01	02	05	06	07	08	10	11	12	13	14	15	16	17	18	99	NS	OK	total
9409	163	267	22	125		699		263	417	775	48	94	8	80	2	130	226	3074	6393
9410	175	257	21	153		719	1	274	472	776	58	87	11	98	3	178	358	2884	6525
9411	179	302	23	160		824		322	512	917	85	95	12	105	6	156	408	3343	7449
9412	165	266	25	153		742	1	251	472	925	82	107	9	116	16	147	356	3174	7007
9501	156	257	22	169		831		245	488	963	81	100	9	122	12	147	316	3231	7149
9502	202	273	22	171	1	949		253	508	1171	86	120	11	118	8	130	292	3795	8110
9503	255	432	37	203	1	1244	4	322	671	1502	126	150	38	204	22	174	462	4685	10532
9504	139	252	16	125	1	720	1	165	427	951	86	92	42	122	20	120	265	2729	6273
9505	147	278	22	140		804	2	224	456	1066	101	115	40	121	19	126	314	3053	7028
9506	139	247	20	132	1	773	2	173	381	1130	124	100	47	137	19	98	238	3041	6802
9507	133	230	22	112		667	1	168	373	991	108	70	28	113	15	100	234	2696	6061
9508	135	278	26	145		860	1	171	426	1130	124	114	35	133	23	79	209	3391	7280
total	1988	3339	278	1788	4	9832	13	2831	5603	12297	1109	1244	290	1469	165	1585	3678	39096	86609

[34] The OK's might not have actually appeared at the courthouse. They might have been notified by telephone that they were not needed. Nevertheless, they had done their duty.

[HNB9596]	01	02	03	05	06	08	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
9509	194	275		25	166	820	3	325	419	687	81	96	40	138	3	118		397	3241	7028
9510	183	321		31	208	934	2	369	497	750	122	117	41	214	3	111		428	3729	8060
9511	198	341		25	179	894	1	340	481	862	103	119	44	140	16	148		439	3701	8031
9512	157	273		24	163	806	5	264	421	766	109	110	32	134	12	73		351	3101	6801
9601	194	300		27	177	1046	1	291	497	1077	126	118	31	147	8	118	220	250	3833	8461
9602	215	329		37	204	999		277	512	1152	132	136	36	153	16	141	630		4015	8984
9603	169	350		26	189	1030		305	596	1252	123	135	46	186	16	102	680		4130	9335
9604	152	283		24	183	870	2	223	431	1049	135	77	30	123	13	82	633		3237	7547
9605	145	257	1	21	156	844	2	235	403	994	109	84	24	123	22	80	630		2983	7113
9606	98	168		21	108	571		142	277	737	75	30	28	85	11	35	470		2114	4970
9607	117	211		14	128	688		139	292	916	103	75	26	131	19	58	685		2337	5939
9608	100	206		16	120	657		132	276	912	108	54	23	87	10	27	535		2446	5709
total	1922	3314	1	291	1981	10159	16	3042	5102	11154	1326	1151	401	1661	149	1093	4483	1865	38867	87978

[HNB9697]	01	02	03	05	06	07	08	10	11	12	13	14	15	16	17	18	99	??	OK	total
9609	148	228		22	156		747		326	222	676	72	87	26	103	9	65	1349	2509	6745
9610	223	298	1	24	212		1090	2	479	289	933	104	111	41	184	4	65	1733	3446	9239
9611	213	265	1	30	202		926		393	261	803	108	97	47	219	5	68	1700	2971	8309
9612	153	238		16	154		871		329	188	748	91	84	41	126	5	54	1794	2453	7345
9701	154	215		12	132	2	838	2	288	162	777	68	86	17	65	8	34	2834	1385	7079
9702	88	81		6	78	1	641		172	86	679	53	27	29	1	3		5981		7926
9703							4		11		13	1				1		5363		5393
total	979	1325	2	110	934	3	5117	4	1998	1208	4629	497	492	201	698	35	286	20754	12764	52036

4. HNB disqualifications by town: 1992-93 through March 1996-97

[HNB9293]	01	02	03	05	06	07	08	10	11	12	13	14	15	16	17	18	99	NS	OK	total
AVON	31			2	5		133	1	18	69	135	18		8	8	4	59	38	782	1311
BERLIN	20			6	25		255		29	86	75	31		7	8	2	64	42	1040	1690
BLOOMFIELD	79			11	9		269	1	22	52	159	29		10	25	6	58	126	1074	1930
BRISTOL	113			33	77		717		84	345	569	75		12	72	13	168	284	3437	5999
BURLINGTON	6			2	2		36		9	44	50	2		5	1	8	28	16	498	707
CANTON	9			2	4		46		10	43	76	8		6	6	3	36	27	511	787
EAST GRANBY	7			1			33		9	28	28	4			1		9	15	292	427
EAST HARTFORD	132			33	97		572		77	183	588	72		28	44	7	138	212	2763	4946
EAST WINDSOR	12			2	9		101		11	47	113	16		1	1	3	29	34	607	986
ENFIELD	57			13	16	1	392		48	281	454	77	1	21	37	7	101	148	2665	4319
FARMINGTON	54			4	12		245		25	100	248	32		16	17	6	73	51	1161	2044
GLASTONBURY	36			9	18		267	1	22	140	209	35		18	16	1	99	67	1793	2731
GRANBY	10			2	1		71		9	67	75	21		5	6	5	45	20	605	942
HARTFORD	500		1	58	700		1072	5	144	280	3663	99		48	351	33	298	1560	4725	13537
MANCHESTER	81			18	40		659	1	68	210	618	53		19	49	15	150	152	2945	5078
MARLBOROUGH	9			3	1		27		7	30	34	4		3	2	3	19	20	373	535
NEW BRITAIN	362			35	379		1093	1	109	246	1135	110		24	87	13	135	454	3223	7406
NEWINGTON	75			7	77		399	1	30	123	211	49		12	17	5	77	85	1712	2880
PLAINVILLE	33			7	25		198		22	96	165	15		1	19	5	48	57	1051	1742
PLYMOUTH	15			4	6		135		17	81	77	17		2	7	3	28	46	712	1150
ROCKY HILL	39			4	35		146		23	53	230	23		13	11	2	45	55	951	1630
SIMSBURY	31			10	1		178		37	137	181	19	1	28	13	4	80	54	1319	2093
SOUTH WINDSOR	45			9	24		132	1	22	125	174	23		9	9	3	80	65	1494	2215
SOUTHINGTON	46			22	37		387		49	236	245	49		6	28	10	132	100	2306	3653
SUFFIELD	20			5	3		127		13	57	81	11		11	6	2	37	35	701	1109
WEST HARTFORD	137			8	91		1176	4	74	272	463	91		47	29	49	268	158	3232	6099
WETHERSFIELD	52			3	58		444		29	113	156	52		7	10	4	71	48	1472	2519
WINDSOR	83			12	24		282	1	35	130	223	34		16	24	12	85	136	1712	2809
WINDSOR LOCKS	15			8	11		157		14	73	88	21		6	6	5	33	36	700	1173
nonHNB	4	3100					2			11			1280	24	2		93	4	17	4537
total	2113	3100	1	333	1787	1	9751	17	1066	3758	10523	1090	1282	413	912	233	2586	4145	45873	88984

[HNB9394]	01	02	05	06	08	10	11	12	13	14	15	16	17	18	99	NS	OK	total
AVON	16		1	4	106	1	41	71	111	12		1	5	2	30	20	635	1056
BERLIN	25		6	20	177		43	73	60	16		1	5	8	32	37	769	1272
BLOOMFIELD	63		4	8	196	1	45	54	137	21		1	15	4	33	116	763	1461
BRISTOL	81		28	59	563		148	283	449	63		6	55	13	82	186	2468	4484
BURLINGTON	8		4	1	25		27	37	38	4		1	8	3	20	15	354	545
CANTON	12		4		49	2	23	36	50	7		1	4	3	18	12	364	585
EAST GRANBY	6		2		23		19	17	20	1					14	11	195	308
EAST HARTFORD	116		20	81	510	1	108	157	478	70		7	58	7	69	130	1979	3791
EAST WINDSOR	6		4	1	70		22	28	98	15	1		10		16	29	430	730
ENFIELD	41		9	20	315		127	221	394	49		3	36	4	56	110	1902	3287
FARMINGTON	33		3	17	177		40	76	165	23		3	8	1	52	46	893	1537
GLASTONBURY	40		3	7	199	1	92	115	181	23		2	10	2	81	39	1245	2040
GRANBY	5		2	2	59		23	45	61	5			8	3	23	14	422	672
HARTFORD	363		30	488	865		176	220	3289	78		6	359	23	169	1094	3460	10620
MANCHESTER	51		8	37	469		121	165	508	57		4	57	5	108	138	2109	3837
MARLBOROUGH	3		2		13		31	23	29	4			2		11	11	278	407
NEW BRITAIN	247		35	300	845	1	142	183	930	98		6	97	5	88	310	2348	5635
NEWINGTON	65		5	50	295		77	102	186	39			16	6	42	50	1242	2175
PLAINVILLE	26		11	23	161		45	64	112	22			9	2	31	53	771	1330
PLYMOUTH	14		7	2	119		28	71	55	13		1	11	1	23	31	490	866
ROCKY HILL	32			17	134		36	55	216	17		3	11	1	30	38	695	1285
SIMSBURY	25		1	3	129		59	111	149	15		3	16	2	52	34	965	1564
SOUTH WINDSOR	34		4	12	116		67	120	124	17		2	7	2	45	51	1100	1701
SOUTHINGTON	36		9	27	292		114	175	251	39		2	24	7	79	52	1697	2804
SUFFIELD	6		6	3	109		32	53	69	6		1	7	1	22	32	473	820
WEST HARTFORD	129		6	63	814	4	177	186	507	68		5	49	4	142	132	2158	4444
WETHERSFIELD	45		4	52	336		53	110	134	28		3	13	7	56	46	1035	1922
WINDSOR	64		6	18	222		72	93	186	31	1	7	24	6	41	78	1188	2037
WINDSOR LOCKS	6		6	5	95		16	69	85	20	1		8	1	23	22	521	878
nonHNB		2688									954				85	1	2	3730
total	1598	2688	230	1320	7483	11	2004	3013	9072	861	957	69	932	123	1573	2938	32951	67823

[HNB9495]	01	02	05	06	07	08	10	11	12	13	14	15	16	17	18	99	NS	OK	total
AVON	27		1	10		131		46	113	151	14		13	12	5	36	31	703	1293
BERLIN	20		7	26		256		62	147	86	23		4	7	5	41	43	869	1596
BLOOMFIELD	58		2	11		276	1	44	86	174	31		4	38	4	38	104	981	1852
BRISTOL	114		25	69		736		251	510	605	88		15	85	9	107	190	3018	5822
BURLINGTON	9		1	2		41		39	61	38	5			3	5	18	25	419	666
CANTON	7		1	1		58		26	68	72	4	1	4	6	2	26	23	462	761
EAST GRANBY	6					41		23	44	45	6		3	4		7	11	234	424
EAST HARTFORD	147		18	107		653	1	171	263	639	71		11	70	6	78	183	2364	4782
EAST WINDSOR	7		4	3		93		33	94	117	14		5	15	1	19	48	536	989
ENFIELD	46		15	23	1	422		177	376	571	63		23	53	5	59	137	2247	4218
FARMINGTON	45		5	18		214		73	169	240	21		10	27	3	47	41	1055	1968
GLASTONBURY	46		5	13		286	1	110	240	220	36		18	32	4	53	60	1528	2652
GRANBY	17		3	1		60		44	88	96	16		3	12	2	21	19	555	937
HARTFORD	398		48	625		984	4	197	368	4896	126		30	572	29	144	1419	3715	13555
MANCHESTER	85		25	38		650	1	182	329	608	75		15	74	10	96	164	2458	4810
MARLBOROUGH	6			4		28		20	69	37	3		3	8		19	9	310	516
NEW BRITAIN	317		29	438		1190	1	224	364	1185	124		19	143	15	91	387	2658	7185
NEWINGTON	72		8	78	2	397		101	175	227	41	1	6	24	9	50	62	1575	2828
PLAINVILLE	36		5	16		200		81	129	138	25		5	23	4	29	66	881	1638
PLYMOUTH	13		5	8		117		57	112	87	18	1		19	5	24	34	562	1062
ROCKY HILL	30		7	34		175		45	96	267	21		5	13	5	40	29	841	1608
SIMSBURY	36		1	4		185		114	186	215	19		20	19	3	45	43	1154	2044
SOUTH WINDSOR	44		10	17		142		103	217	146	28		10	17	7	48	50	1298	2137
SOUTHINGTON	56		16	38		392		139	341	263	41		11	45	8	80	81	2021	3532
SUFFIELD	11		2	2		136		51	81	82	16		2	4	1	27	37	602	1054
WEST HARTFORD	177		10	94	1	1051	4	198	427	632	89		27	69	7	140	176	2601	5703
WETHERSFIELD	50		9	71		471		82	182	144	44		6	23	6	38	53	1248	2427
WINDSOR	92		11	31		285		84	169	219	31		15	41	3	49	109	1539	2678
WINDSOR LOCKS	16		5	6		162		54	99	97	16		3	11	2	21	44	660	1196
nonHNB		3339										1241				94		2	4676
total	1988	3339	278	1788	4	9832	13	2831	5603	12297	1109	1244	290	1469	165	1585	3678	39096	86609

[HNB9596]	01	02	03	05	06	08	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
AVON	20				5	172	1	51	116	124	13		18	11	2	24	63	11	705	1336
BERLIN	25			3	23	249		81	140	81	34	1	6	16	5	26	83	19	848	1640
BLOOMFIELD	55			4	4	304	1	62	74	172	33		9	49	3	27	140	60	941	1938
BRISTOL	100			25	78	744		247	488	623	97		18	100	11	65	305	101	2873	5875
BURLINGTON	10			3	4	41		35	66	39	9		4	4	4	8	34	7	409	677
CANTON	8			1	4	71		47	75	83	13		4	2		11	31	10	429	789
EAST GRANBY	5			1		36		15	32	23	12	1	2	2		5	26	4	253	417
EAST HARTFORD	129			26	117	653		163	227	641	88		28	103	9	72	222	107	2345	4930
EAST WINDSOR	9			7	8	95		37	67	115	16	1	4	19	2	15	40	24	546	1005
ENFIELD	62			17	23	439		198	361	461	71	2	25	59	9	46	200	68	2262	4303
FARMINGTON	55			6	27	233		87	163	203	20		13	20	3	44	97	26	964	1961
GLASTONBURY	42			7	16	291	3	142	206	214	35		14	24	4	42	128	30	1506	2704
GRANBY	9			2	1	85		44	82	85	13		5	10	1	8	37	17	533	932
HARTFORD	416			52	734	1105	5	233	307	3857	143		47	639	33	123	1090	709	4192	13685
MANCHESTER	84		1	21	44	630	1	188	318	695	85		39	103	5	48	219	93	2386	4960
MARLBOROUGH	8			1	2	27		23	58	42	7		4	5	1	10	35	7	306	536
NEW BRITAIN	285			34	468	1221		211	334	1137	124		22	154	5	67	421	179	2690	7352
NEWINGTON	61			5	79	462		110	171	231	47		6	22	2	46	133	41	1500	2916
PLAINVILLE	29			8	29	197		67	118	134	36		6	18	4	21	83	24	909	1683
PLYMOUTH	11			9	7	152		50	101	87	18		5	14	1	12	54	25	559	1105
ROCKY HILL	30			5	29	157	1	57	92	253	34		7	20	3	25	67	27	787	1594
SIMSBURY	39			2	7	171		105	175	192	25		14	21	5	26	102	25	1150	2059
SOUTH WINDSOR	53			8	20	146		121	193	129	29		13	22	4	37	87	23	1260	2145
SOUTHINGTON	54			15	36	391		147	297	285	58		13	34	11	44	165	52	1932	3534
SUFFIELD	12			1	6	147		55	77	81	32		4	15	2	15	61	13	608	1129
WEST HARTFORD	157			13	116	1024	3	215	341	629	112	1	28	82	13	78	272	56	2604	5744
WETHERSFIELD	58			6	63	500	1	92	186	166	47		16	19	4	29	98	32	1256	2573
WINDSOR	81			7	26	257		99	151	281	49		21	50	3	37	136	53	1459	2710
WINDSOR LOCKS	15			2	5	159		60	86	91	26		6	24		9	53	22	640	1198
nonHNB		3314										1145				73	1		15	4548
total	1922	3314	1	291	1981	10159	16	3042	5102	11154	1326	1151	401	1661	149	1093	4483	1865	38867	87978

[HNB9697]	01	02	03	05	06	07	08	10	11	12	13	14	15	16	17	18	99	??	OK	total
AVON	11				3		82		53	28	60	5		4	3	1	10	356	213	829
BERLIN	8			1	14		130		51	31	23	16		6	3	1	6	419	294	1003
BLOOMFIELD	32		1	4	5		133		49	17	71	15		5	16		4	511	347	1210
BRISTOL	61			10	24		363		138	122	219	38		13	30		15	1576	942	3551
BURLINGTON	3				1		22		26	13	13	1		3	6		2	214	122	426
CANTON	5				1		36		23	14	20	4		1	4	1	4	191	145	449
EAST GRANBY	1			1			19		12	8	12	2		2	1	2	1	111	77	249
EAST HARTFORD	68			10	59	1	336		110	71	225	34		9	46		18	1211	745	2943
EAST WINDSOR	9			1	1		53		27	13	38	3		2	8		2	244	158	559
ENFIELD	25			8	8		210		119	76	170	31		5	32	4	11	1150	740	2589
FARMINGTON	27				11		107		41	35	101	15		8	10	1	9	502	337	1204
GLASTONBURY	21			2	11		143		80	43	94	13		10	10	1	12	688	497	1625
GRANBY	4				2		39		31	16	45	2		4	4	2	3	230	187	569
HARTFORD	214		1	17	367		520	2	145	65	1798	38		18	259	7	38	3208	1510	8207
MANCHESTER	38			4	17		334		114	77	253	23		17	31		10	1257	792	2967
MARLBOROUGH	3			1	1		17	1	13	14	27	1		5	3		1	151	90	328
NEW BRITAIN	164			11	221	1	570		154	75	477	54		14	77	2	21	1659	947	4447
NEWINGTON	44			3	29		227		84	45	82	20		6	18	2	8	664	440	1672
PLAINVILLE	20			3	19		95		42	33	49	16		4	10	2	6	437	286	1022
PLYMOUTH	10			2	3		72		31	33	22	12	1	1	7		3	276	187	660
ROCKY HILL	14			2	18		80		37	22	98	12		4	13		6	383	259	948
SIMSBURY	13			2	3	1	95		63	47	80	16		10	11		6	552	356	1255
SOUTH WINDSOR	17			3	10		93		63	48	45	9		6	9		7	588	402	1300
SOUTHINGTON	25			8	14		211		116	80	109	24		4	19	1	7	1008	611	2237
SUFFIELD	4			2	1		76		37	19	31	3		5	2		4	277	193	654
WEST HARTFORD	68			4	46		537	1	154	70	255	42		21	35	4	22	1365	852	3476
WETHERSFIELD	31			3	30		283		65	42	64	27		3	10		7	562	378	1505
WINDSOR	34			5	10		149		75	28	106	13		7	16	2	10	671	445	1571
WINDSOR LOCKS	5			3	5		85		45	23	42	8		4	5	2	1	293	212	733
nonHNB		1325											491				32			1848
total	979	1325	2	110	934	3	5117	4	1998	1208	4629	497	492	201	698	35	286	20754	12764	52036

Appendix

The geocoding algorithm

Perfect geocoding to tracts would correctly match each juror record in the JIS summary files with a unique Census tract. I am not able to achieve perfection, but I can get quite a high success rate.

It has taken me many months to arrive at the current form of the algorithm, by a process of repeated error-checking and modification. What follows in the first Section is an outline of the main technical features of the algorithm I use to generate the geocoding estimates in this report. I hope it will aid anyone wishing to read the Perl source code of my implementation of the algorithm.

1. Outline of the method

For each record in the JIS summary files, I generate a return-code (giving information about the reliability of the match) and a list of possible matching tracts.

I use the TIGER/CTSI database, which maps streets to Census tracts. I preprocess the CTSI entries into a hashtable of possible matching addresses indexed by street name/zip code pairs. The table is augmented by entries identified by means of a laborious procedure involving a street atlas, a tract map, a zipcode directory, and (occasionally) a telephone directory: As I found street JIS address that were not being matched by the CTSI data, I added hash entries. For example, I added

MAHL,06120 = “5013:2:72:E::St::Hartford” + “ 5018:1:71:O::St::Hartford”
MAHL,06112 = “5018:1:11:O::St::Hartford”

to solve a problem involving a street in Hartford that crosses a zipcode boundary and has had a name change not recorded in the TIGER file. The list records tract number, range of street numbers, parity (odd or even or both sides of the street), various prefixes and suffixes, and town name.

The matching procedure for a given the juror record begins by extracting the towncode, address, zip, and disq fields.

[1] For disq equal to 02, 15, or 16 put return-code equal to xjd and give an empty list of matching tracts; then move on the next record.

[2] Look for addresses starting with ‘POB’ or PO Box” and so on. Put return-code equal to xjd and give an empty list of matching tracts; then move on the next record.

[3] Attempt to parse the address into components
(house-number:streetname-prefix:streetname:street-type:direction-suffix)
For example,

original address	parsed form
24 Hillhouse Avenue	(24::Hillhouse:Ave:)
Apt23A 199 East Main Street	(199:E:Main:St:)
17 Euclid Strt West	(17::Euclid:St:W)
24 E. Euclid St	(24:E:Euclid:St:)

There are many subtle cases that require delicate handling. For example, should "12E Grove Hill" be interpreted as "12 East Grove Hill", or apartment 12 at East Grove Hill"? And should it be "Grove Hill Road" or "Grove Hill" with the "Hill" playing the role of a street type?

The parsing step has to contend with strange abbreviations, missing spaces (between house-number and streetname, for example), and various other ways in which an address can get mangled.

The algorithm makes up to five attempts (labelled A, B, C, D, and E as the first character of the result code) at finding a match.

(A) Use the address from the JIS file.

(B) Try again with an address prefix (NSEW) as part of the street name.

(C) Attempt spelling corrections then try again. (Apply a substitution defined by a lookup into a hash table indexed by a compressed form of the street name plus towncode. Some zipcode errors are also corrected by the lookup.) The improved matching rates for Hartford and New Britain towns are mostly due to hard labor expended in construction of the hash tables.

(D) Strip off trailing characters (such as a suite number in a strange form) that might be misinterpreted as an 'fetype' (road, avenue, etc) then try again.

(E) Replace zipcode by adjacent zipcode (for example, 06106 instead of 06105), then try again.

The algorithm makes another pass only if all previous passes have found no possible streetname/zipcode matches in the CTSI hashtable.

The algorithm checks the parsed form of the JIS address against the CTSI hashtable, using the streetname/zipcode as a lookup key. If it finds a nonempty list of possible matches, the match-level is set at 1. If any of the level-1 matches has a range of street numbers and parity consistent with the JIS street number, the match-level is increased to 2. If any of the level-2 matches has the same 'fetype' as the JIS address, the match-level is increased to 3. If any of the level-3 matches has the same direction-suffix (NSEW) as the JIS address, the match-level is increased to 4. If any of the level-4 matches has the same prefix (NSEW) as the JIS address, the match-level is increased to 5. The list of possible tract numbers for the highest level match is written to a file. The return code is made up of the pass-letter together with the highest level of match. For example, a return code of A5 indicates that the first pass found at least one match at level 5. If the list of tracts contains more than one distinct tract number, the match is recorded as 'multi' (multiple matches); otherwise it is recorded as 'unique'.

The tables at the end of the Section summarize the results of the geocoding with the JIS for HNB judicial district, for each of the five court years.

In addition to recognizing the varying degrees of certainty in a match, I performed many consistency checks on the geocoding results. For example, I used a Census tract map to determine which tracts lie in which towns, and then I compared the towncode listed in the JIS summary file with the towncode corresponding to the matched tract. The comparison for the 1992-93 court year—the first full year of operation for the current JIS system—is the most interesting. It shows that 921 out of 6280 records that are geocoded to a tract in West Hartford correspond a JIS address giving the town as Hartford. A handful of the misallocated addresses lie along Propect Avenue, the town boundary, but the rest are squarely in the 06119 and 06110 zipcode regions of West Hartford.[35]

[35] Karna Bryan, a graduate student in the Yale Statistics Department, found similar inconsistencies between JIS towncodes and zipcodes in the 1992-93 JIS data for the New Haven judicial district. Several jurors from West Haven appear to have served in the wrong district because their address gave New Haven as their town.

I believe this misallocation reflects one of the teething problems that the JIS system had to overcome after its first year of operation. The JIS problem seems largely to have disappeared after the first year. The towncode assigned via geocoding agrees with the JIS towncode for most records.

MATCHING RATES FOR GEOCODING

HNB9293	unique	multi	—	total
A1	2431	870		3301
A2	20931	360		21291
A3	859	163		1022
A4	298	33		331
A5	47698	24		47722
B1	9	16		25
B2	123			123
B3	5			5
B5	286			286
C1	3	20		23
C2	151			151
C3	21	1		22
C4	2			2
C5	191			191
D1	6			6
D2	7	2		9
E1	52	21		73
E2	18			18
E3	1			1
E5	157			157
nomatch			8072	8072
pbox			1358	1358
xjd			4795	4795
total	73249	1510	14225	88984

HNB9394	unique	multi	—	total
A1	2061	686		2747
A2	15465	243		15708
A3	677	120		797
A4	229	25		254
A5	36699	17		36716
B1	14	10		24
B2	88			88
B3	3			3
B5	201			201
C1	4	16		20
C2	134			134
C3	4	1		5
C4	3			3
C5	165			165
D1	5			5
D2	13	1		14
E1	29	13		42
E2	12	1		13
E3				
E5	93			93
nomatch			6094	6094
pbox			983	983
xjd			3714	3714
total	55899	1133	10791	67823

HNB9495	unique	multi	—	total
A1	2464	753		3217
A2	19843	273		20116
A3	846	157		1003
A4	275	24		299
A5	46911	27		46938
B1	8	12		20
B2	114			114
B3	3			3
B5	279			279
C1	4	5		9
C2	138			138
C3	4			4
C4	4			4
C5	184			184
D1	7			7
D2	9	1		10
E1	41	17		58
E2	21	1		22
E4	1			1
E5	117			117
nomatch			7990	7990
pbox			1203	1203
xjd			4873	4873
total	71273	1270	14066	86609

HNB9596	unique	multi	—	total
A1	2527	836		3363
A2	20124	325		20449
A3	891	182		1073
A4	251	30		281
A5	47232	22		47254
B1	12	11		23
B2	109			109
B3	2			2
B5	225			225
C1	12	9		21
C2	135			135
C3	3	2		5
C4	4			4
C5	214			214
D1	9			9
D2	8	2		10
E1	57	26		83
E2	20			20
E3	7			7
E4	1			1
E5	116			116
nomatch			8452	8452
pbox			1256	1256
xjd			4866	4866
total	71959	1445	14574	87978

HNB9697	unique	multi	—	total
A1	1535	517		2052
A2	12156	196		12352
A3	543	92		635
A4	136	10		146
A5	28335	15		28350
B1	9	9		18
B2	70			70
B3	3			3
B5	154			154
C1	7	11		18
C2	78			78
C3	3			3
C5	121			121
D1	8			8
D2	9			9
E1	36	18		54
E2	21			21
E3	3			3
E5	105			105
nomatch			5049	5049
pbox			769	769
xjd			2018	2018
total	43332	868	7836	52036

2. Estimation via geocoding

For each disqualification code (including NS and OK), I have counts of the number of JIS records uniquely geocoded to each Census tract, the number of records geocoded to multiple tracts, and the number of records that could not be matched to any tract. From these counts I am able to form estimates of the numbers of blacks and Hispanics disqualified in various ways.

For any particular court year, let me write $N(\text{TRACT}, \text{DISQ})$ for the number of records with disqualification code 'DISQ' (=01,...,OK) that are uniquely geocoded to tract number 'TRACT' (mostly in the range 4001 ... 5241). From the STF3A census tables P14C and P14D, I determine the fraction $b(\text{TRACT})$ of the over-18 population of the tract that was black according to the 1990 census. The numbers listed in the rows labelled 'Bgeo' in the tables at the end of the Section are calculated as

$$\underset{\text{TRACT}}{\text{Sum over}} \big(N(\text{TRACT}, \text{DISQ}) \times b(\text{TRACT}) \big) \qquad \text{for each DISQ code.}$$

The sum over all disqualification codes (including NS and OK) appears in the last column of each table. The 'Bgeo' rows give the estimates of the numbers of blacks disqualified (or qualified) in various ways, *amongst all the JIS records that could be uniquely geocoded*. The total counts are less important than the percentage breakdowns, which follow the tables of counts.

It is important to realize that the geocoding estimates for the disqualification codes 02, 15, and 16 are meaningless because those codes cannot be assigned to tracts in the HNB judicial district, by definition. If the counts for those disqualifications could be added to the table, the percentages for the other disqualifications would decrease slightly.

I also attempted to adjust the estimates for black disqualifications by removing the proportions that could be allocated to black Hispanics. I do not tabulate the results, because the figures are almost identical with the 'Bgeo' values.

For the 'Hgeo' rows—the estimates of the Hispanic disqualifications based on the geocoding to tracts—I refine the method of estimation by drawing on other types of Census data. I attempt to apportion the various disqualifications between the 'eligible populations' within a tract. For code 13 it seems reasonable to use the raw fractions calculated from STF3A, because the disqualification mechanisms cannot operate if a person does not even receive the summons. For the language and citizenship disqualifications I attempt to narrow the eligible populations down to those who (at least on the basis of 1990 Census counts) could be expected to be eligble for the disqualifications. I adjust the OK eligible population only for the language. (More precise mathematical descriptions of my geocoding estimates are given in the last Section of this Appendix.)

For disqualification codes 01, 06, 08, and OK I replace the proportions $h(\text{TRACT})$ of Hispanics in tract number 'TRACT' by proportions $h_{01}(\text{TRACT})$, $h_{06}(\text{TRACT})$, $h_{08}(\text{TRACT})$, and $h_{OK}(\text{TRACT})$, calculated as follows.

(i) From PUMS[36] I estimate the proportion of noncitizens amongst each of the classified Hispanic subgroups. I combine those proportions with the counts from STF3A table P11 to estimate the proportion of Hispanics who are noncitizens for each tract. I apply those proportions to the ratio of Hispanics over 18 (from STF3A tables P15A and P15B) to all noncitizens over 18 (from STF3A table P37) to estimate $h_{01}(\text{TRACT})$, the proportion of noncitizens over 18 who are Hispanic.[37]

[36] Public Use Microdata Samples, the 5% sample for Connecticut. The region covered by PUMA's 00200, 00300, 00400, 00500, 00600, 00700, and 00800 almost coincides with Hartford County.

[37] When I calculated the citizenship disqualifications in this way, I got figures very close to (but always slightly smaller than) what I got by applying the over-18 tract proportions of Hispanics to disqualification 01 counts. The effect on the percentage breakdown of Hispanic disqualifications was barely noticeable. In the interests of simplicity of method, I have therefore not used the PUMS calculation for the tabulations in this report.

Probably this method overestimates the citizenship disqualifications, because it does not exclude from the noncitizen pool those Hispanics who would be eligible for other disqualifications.

(ii) From STF3A table P28 I calculate the total number of persons over 18 in each TRACT who spoke English "well", "not well or not at all". (That is, I exclude persons who identified themselves as speaking English "very well" from the pool of persons who might be disqualified on language grounds.) I calculate the similar figure for persons with Spanish listed as the "language spoken at home", which I use as a surrogate for Hispanics. The ratio of the two counts estimates the proportion $h_{06}(\text{TRACT})$ of Hispanics amongst the over 18 population of a tract who might be candidates for a code 06 disqualification.

I expect the inclusion of the "well"" and "not well or not at all" categories in the pool of those who could claim a language disqualification will lead to an overestimate of code 06 Hispanic disqualifications. The pool undoubtedly includes persons who would have been eligible for other disqualifications. Accordingly, I suggest that the code 06 estimates should be regarded as conservative upper bounds for the language disqualifications.

(iii) From STF3A tables P13, P15A, and P15B I calculate $h_{08}(\text{TRACT})$ as the proportion of persons over 70 in the TRACT (in 1990) who were Hispanic.

(iv) In order to avoid an overestimate of the total number of Hispanic disqualifications, I also adjust the eligible population for the "OK" category by subtracting out the language pool already accounted for in (ii). This correction probably leads to an underestimate of the OK and totals by geocoding.

The estimates in the rows labelled 'Hgeo' are calculated in a similar fashion to the 'Bgeo' rows, but with the Hispanic proportions substituted for the black proportions.

The 'SSLgeo' rows are calculated by applying the Hispanic proportions by surname, calculated as in Section 9, to those JIS records that could be uniquely geocoded, then summing over tracts. I include the 'SSLgeo' row for the sake of comparison with the 'Hgeo' and 'SSL' (calculated by applying the surname proportions to all records in the JIS files) rows. The calculation for the 'SSL' rows draw from records for jurors in my 'xjd' (= 02,15, and 16) disqualification grouping.

The 'ALLgeo' rows merely count up the numbers for each disqualification code amongst the JIS records that can be uniquely geocoded to a tract. The 'ALL' calculates similarly for all JIS records, and not just those that can be geocoded.

I obtain the estimates for nonHispanics (row 'nonH') by subtracting the estimates in the 'SSL' rows from the counts in the 'ALL' rows.

The tabulations

The first five tables contain estimates and counts for the whole judicial district. The next five tables give corresponding estimates and counts for Hartford town. I extracted from the JIS summary files those records with the towncode for Hartford (064), then applied the same methods as before. The next five tables give corresponding estimates and counts for New Britain town, using records with the towncode for New Britain (089).

The last fifteen tables merely express the estimates and counts as percentages by row.

Estimated counts for the whole HNB judicial district

HNB9293	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	NS	OK	total
Bgeo	249				34	181		521		1	65	198	1330	58			147	12	147	867	2898	6708
Hgeo	143				21	592		111		1	46	122	1226	39			127	8	98	484	1143	4161
SSLgeo	120				30	606		80		1	23	95	1448	14			155	4	71	578	1487	4712
ALLgeo	1839				285	1609	1	8626		17	939	3284	9057	967			797	205	2130	3711	39781	73248
SSL	137	77			35	669		88		1	26	106	1572	14	28	15	168	6	86	617	1692	5337
nonH	1976	3023	1		298	1118	1	9663		16	1040	3652	8951	1076	1254	398	744	227	2500	3528	44181	83647
ALL	2113	3100	1		333	1787	1	9751		17	1066	3758	10523	1090	1282	413	912	233	2586	4145	45873	88984

HNB9394	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	NS	OK	total
Bgeo	198				21	139		442			117	160	1265	43			163	13	94	654	2312	5621
Hgeo	112				13	447		94			65	100	1095	35			115	6	62	339	879	3362
SSLgeo	104				14	439		71			38	93	1250	13			168	2	55	406	1170	3823
ALLgeo	1385				206	1194		6581		9	1769	2628	7890	745			823	102	1290	2640	28636	55898
SSL	119	55			16	485		78		1	44	101	1370	15	20	3	178	3	60	429	1318	4295
nonH	1479	2633			214	835		7405		10	1960	2912	7702	846	937	66	754	120	1513	2509	31633	63528
ALL	1598	2688			230	1320		7483		11	2004	3013	9072	861	957	69	932	123	1573	2938	32951	67823

HNB9495	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	NS	OK	total
Bgeo	228				26	180		573		1	135	293	1812	72			258	16	94	808	2620	7116
Hgeo	123				20	589		113		1	84	184	1674	49			199	7	56	446	976	4521
SSLgeo	121				30	611		91			46	168	2073	13			283	7	49	553	1423	5468
ALLgeo	1754				243	1614	4	8752		11	2459	4865	10792	986			1303	146	1265	3311	33767	71272
SSL	140	95			32	668		99			52	188	2199	15	38	16	298	8	54	585	1598	6085
nonH	1848	3244			246	1120	4	9733		13	2779	5415	10098	1094	1206	274	1171	157	1531	3093	37498	80524
ALL	1988	3339			278	1788	4	9832		13	2831	5603	12297	1109	1244	290	1469	165	1585	3678	39096	86609

HNB9596	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
Bgeo	219				27	193		602		1	144	236	1504	77			271	14	68	646	404	2791	7197
Hgeo	131				19	653		124		2	91	155	1280	58			227	8	48	345	208	1031	4380
SSLgeo	116				24	720		106			70	170	1798	22			321	4	50	400	247	1596	5644
ALLgeo	1669		1		241	1772		8930		13	2619	4375	9544	1152			1483	135	875	3916	1663	33571	71959
SSL	134	104			32	801		114			82	188	1967	23	33	20	343	5	58	435	265	1795	6399
nonH	1788	3210	1		259	1180		10045		16	2960	4914	9187	1303	1118	381	1318	144	1035	4048	1600	3707 2	81579
ALL	1922	3314	1		291	1981		10159		16	3042	5102	11154	1326	1151	401	1661	149	1093	4483	1865	3886 7	87978

HNB9697	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
Bgeo	106		1		9	107		285			103	54	681	27			111	4	21	1811		948	4268
Hgeo	68				8	319		62			59	34	596	19			87	1	12	1131		374	2770
SSLgeo	66				11	375		53			50	27	833	7			128	1	15	1336		599	3501
ALLgeo	867		2		96	835	3	4520		3	1722	1046	4022	438			609	31	214	17929		10995	43332
SSL	75	54			13	420		62			57	29	898	8	16	10	141	1	18	1479		672	3953
nonH	904	1271	2		97	514	3	5055		4	1941	1179	3731	489	476	191	557	34	268	19275		12092	48083
ALL	979	1325	2		110	934	3	5117		4	1998	1208	4629	497	492	201	698	35	286	20754		12764	52036

Estimated counts for Hartford town

HAR9293	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
Bgeo	162				21	134		243		1	37	105	1065	25			118	6	87		722	1588	4314
Hgeo	81				14	396		51			23	48	979	15			105	3	57		384	552	2708
SSLgeo	53				20	398		30		1	10	34	1136	3			112		35		406	615	2853
ALLgeo	452				54	636		1036		5	135	271	3402	94			332	31	275		1484	4379	12586
SSL	59				20	443		30		1	11	34	1213	3		9	118	1	41		427	680	3090
nonH	441		1		38	257		1042		4	133	246	2450	96		39	233	32	257		1133	4045	10447
ALL	500		1		58	700		1072		5	144	280	3663	99		48	351	33	298		1560	4725	13537

HAR9394	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
Bgeo	135				12	102		234			61	81	1045	17			139	8	59		545	1347	3785
Hgeo	67				6	300		48			27	46	894	15			92	4	36		271	450	2256
SSLgeo	35				3	277		21			8	32	958	2			117	1	31		274	525	2284
ALLgeo	344				30	459		856			169	211	3141	74			338	20	159		1057	3264	10122
SSL	36				3	301		21			9	32	1017	3			125	1	32		284	560	2424
nonH	327				27	187		844			167	188	2272	75		6	234	22	137		810	2900	8196
ALL	363				30	488		865			176	220	3289	78		6	359	23	169		1094	3460	10620

HAR9495	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
Bgeo	149				16	130		281		1	67	150	1526	40			212	12	53		685	1476	4798
Hgeo	67				13	385		49		1	32	76	1405	24			166	4	29		360	483	3094
SSLgeo	49				15	400		24			14	65	1669	5			225	2	19		412	611	3510
ALLgeo	378				43	592		961		4	191	356	4703	126			558	26	134		1371	3512	12955
SSL	53				16	422		25			14	69	1736	5		4	232	3	21		429	651	3680
nonH	345				32	203		959		4	183	299	3160	121		26	340	26	123		990	3064	9875
ALL	398				48	625		984		4	197	368	4896	126		30	572	29	144		1419	3715	13555

HAR9596	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
Bgeo	145				18	140		315		1	73	117	1229	39			221	11	40	500	341	1670	4860
Hgeo	79				11	427		58		2	36	63	1036	29			189	6	29	255	170	536	2926
SSLgeo	45				7	455		31			19	57	1391	7			259	2	26	247	178	741	3465
ALLgeo	392				44	681		1081		5	222	291	3622	141			616	31	114	1041	677	3963	12921
SSL	47				11	499		31			21	60	1483	7		8	273	3	28	260	185	790	3706
nonH	369				41	235		1074		5	212	247	2374	136		39	366	30	95	830	524	3402	9979
ALL	416				52	734		1105		5	233	307	3857	143		47	639	33	123	1090	709	4192	13685

HAR9697	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
Bgeo	69		1		5	82		136			48	25	570	10			89	4	14	1216		568	2837
Hgeo	42				6	217		32			23	12	499	7			70	1	7	715		209	1840
SSLgeo	22				8	245		21			17	11	651	3			94	1	6	795		293	2167
ALLgeo	205		1		17	342		514		2	137	61	1720	36			242	6	35	3029		1406	7753
SSL	26				8	265		22			17	12	677	3		7	100	1	7	841		317	2303
nonH	188		1		9	102		498		2	128	53	1121	35		11	159	6	31	2367		1193	5904
ALL	214		1		17	367		520		2	145	65	1798	38		18	259	7	38	3208		1510	8207

Estimated counts for New Britain town

NB9293	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
Bgeo	20				1	24		62			6	16	69	6			6	1	7		31	193	442
Hgeo	39				2	109		24			10	27	148	11			13	1	12		61	197	654
SSLgeo	10				3	108		13			2	19	174	2			20	1	11		92	239	694
ALLgeo	332				25	355		1044		1	106	238	1036	107			78	13	124		418	3035	6912
SSL	11				7	116		13			2	20	190	2		1	22	1	11		99	260	755
nonH	351				28	263		1080		1	107	226	945	108		23	65	12	124		355	2963	6651
ALL	362				35	379		1093		1	109	246	1135	110		24	87	13	135		454	3223	7406

NB9394	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
Bgeo	14				2	19		48			8	11	58	5			7		5		22	140	339
Hgeo	27				4	85		20			12	17	116	9			14	1	8		40	148	501
SSLgeo	9				4	88		14			7	16	157	2			30		4		65	184	580
ALLgeo	232				31	282		807		1	136	179	853	92			89	4	83		293	2207	5289
SSL	10				4	96		15			8	17	183	2			31		5		68	204	643
nonH	237				31	204		830		1	134	166	747	96		6	66	5	83		242	2144	4992
ALL	247				35	300		845		1	142	183	930	98		6	97	5	88		310	2348	5635

NB9495	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
Bgeo	18				2	28		69			12	22	79	8			9	1	6		28	154	436
Hgeo	32				3	123		27			19	38	157	13			18	2	9		53	155	649
SSLgeo	14				8	116		19			8	30	216	2			32	3	8		75	205	736
ALLgeo	298				27	408		1139		1	214	352	1080	121			133	15	84		367	2490	6729
SSL	16				8	130		19			9	34	232	2		2	35	3	9		79	222	800
nonH	301				21	308		1171		1	215	330	953	122		17	108	12	82		308	2436	6385
ALL	317				29	438		1190		1	224	364	1185	124		19	143	15	91		387	2658	7185

NB9596	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
Bgeo	15				2	31		71			11	20	69	8			10		4	26	12	161	440
Hgeo	30				4	133		30			18	32	136	14			22		7	44	20	167	657
SSLgeo	12				6	143		14			7	35	208	3			33	1	6	64	33	243	808
ALLgeo	264				31	434		1169			197	323	1035	119			144	5	63	388	163	2515	6850
SSL	14				7	158		15			10	37	229	3		1	35	1	6	73	37	272	898
nonH	271				27	310		1206			201	297	908	121		21	119	4	61	348	142	2418	6454
ALL	285				34	468		1221			211	334	1137	124		22	154	5	67	421	179	2690	7352

NB9697	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK	total
Bgeo	9				1	14		33			8	4	30	3			4		1	100		57	264
Hgeo	15				1	60		13			12	6	55	6			9		2	172		58	409
SSLgeo	13				2	65		6			8	2	89				19		2	199		97	502
ALLgeo	153				10	205	1	543			142	73	425	53			68	2	19	1546		879	4119
SSL	14				2	75		6			8	2	104			1	22		2	221		107	564
nonH	150				9	146	1	564			146	73	373	54		13	55	2	19	1438		840	3883
ALL	164				11	221	1	570			154	75	477	54		14	77	2	21	1659		947	4447

Estimated percentage disqualifications for the whole HNB judicial district

HNB9293	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	4	3	8	3	20	2	13	43	5		100
Hgeo	3	14	3	3	29	3	12	27	5		100
SSLgeo	3	13	2	2	31	3	12	32	3		100
ALLgeo	3	2	12	4	12	1	5	54	6		100
SSL	3	13	2	2	29	3	12	32	3	2	100
nonH	2	1	12	4	11	1	4	53	6	6	100
ALL	2	2	11	4	12	1	5	52	6	5	100

HNB9394	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	4	2	8	3	22	3	12	41	5		100
Hgeo	3	13	3	3	33	3	10	26	5		100
SSLgeo	3	11	2	2	33	4	11	31	3		100
ALLgeo	2	2	12	5	14	1	5	51	7		100
SSL	3	11	2	2	32	4	10	31	3	2	100
nonH	2	1	12	5	12	1	4	50	7	6	100
ALL	2	2	11	4	13	1	4	49	7	5	100

HNB9495	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	3	3	8	4	25	4	11	37	5		100
Hgeo	3	13	3	4	37	4	10	22	5		100
SSLgeo	2	11	2	3	38	5	10	26	3		100
ALLgeo	2	2	12	7	15	2	5	47	7		100
SSL	2	11	2	3	36	5	10	26	3	2	100
nonH	2	1	12	7	13	1	4	47	7	6	100
ALL	2	2	11	6	14	2	4	45	7	6	100

HNB9596	01	06	08	12	13	17	??	NS	OK	rest	xjd	total
Bgeo	3	3	8	3	21	4	9	6	39	5		100
Hgeo	3	15	3	4	29	5	8	5	24	5		100
SSLgeo	2	13	2	3	32	6	7	4	28	3		100
ALLgeo	2	2	12	6	13	2	5	2	47	7		100
SSL	2	13	2	3	31	5	7	4	28	3	2	100
nonH	2	1	12	6	11	2	5	2	45	7	6	100
ALL	2	2	12	6	13	2	5	2	44	7	6	100

HNB9697	01	06	08	12	13	17	??	NS	OK	rest	xjd	total
Bgeo	2	2	7	1	16	3	42		22	4		100
Hgeo	2	12	2	1	21	3	41		13	4		100
SSLgeo	2	11	2	1	24	4	38		17	2		100
ALLgeo	2	2	10	2	9	1	41		25	6		100
SSL	2	11	2	1	23	4	37		17	2	2	100
nonH	2	1	11	2	8	1	40		25	6	4	100
ALL	2	2	10	2	9	1	40		25	6	4	100

Estimated percentage disqualifications for Hartford town

HAR9293	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	4	3	6	2	25	3	17	37	4		100
Hgeo	3	15	2	2	36	4	14	20	4		100
SSLgeo	2	14	1	1	40	4	14	22	2		100
ALLgeo	4	5	8	2	27	3	12	35	5		100
SSL	2	14	1	1	39	4	14	22	2		100
nonH	4	2	10	2	23	2	11	39	5		100
ALL	4	5	8	2	27	3	12	35	5		100

HAR9394	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	4	3	6	2	28	4	14	36	4		100
Hgeo	3	13	2	2	40	4	12	20	4		100
SSLgeo	2	12	1	1	42	5	12	23	2		100
ALLgeo	3	5	8	2	31	3	10	32	4		100
SSL	1	12	1	1	42	5	12	23	2		100
nonH	4	2	10	2	28	3	10	35	5		100
ALL	3	5	8	2	31	3	10	33	4		100

HAR9495	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	3	3	6	3	32	4	14	31	4		100
Hgeo	2	12	2	2	45	5	12	16	3		100
SSLgeo	1	11	1	2	48	6	12	17	2		100
ALLgeo	3	5	7	3	36	4	11	27	4		100
SSL	1	11	1	2	47	6	12	18	2		100
nonH	3	2	10	3	32	3	10	31	5		100
ALL	3	5	7	3	36	4	10	27	4		100

HAR9596	01	06	08	12	13	17	??	NS	OK	rest	xjd	total
Bgeo	3	3	6	2	25	5	10	7	34	4		100
Hgeo	3	15	2	2	35	6	9	6	18	4		100
SSLgeo	1	13	1	2	40	7	7	5	21	2		100
ALLgeo	3	5	8	2	28	5	8	5	31	4		100
SSL	1	13	1	2	40	7	7	5	21	2		100
nonH	4	2	11	2	24	4	8	5	34	5		100
ALL	3	5	8	2	28	5	8	5	31	4		100

HAR9697	01	06	08	12	13	17	??	NS	OK	rest	xjd	total
Bgeo	2	3	5	1	20	3	43		20	3		100
Hgeo	2	12	2	1	27	4	39		11	2		100
SSLgeo	1	11	1	1	30	4	37		14	2		100
ALLgeo	3	4	7	1	22	3	39		18	3		100
SSL	1	11	1	1	29	4	36		14	2		100
nonH	3	2	8	1	19	3	40		20	4		100
ALL	3	4	6	1	22	3	39		18	3		100

ESTIMATED PERCENTAGE DISQUALIFICATIONS FOR NEW BRITAIN TOWN

NB9293	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	4	5	14	4	16	1	7	44	5		100
Hgeo	6	17	4	4	23	2	9	30	6		100
SSLgeo	2	16	2	3	25	3	13	35	3		100
ALLgeo	5	5	15	3	15	1	6	44	5		100
SSL	2	15	2	3	25	3	13	34	3		100
nonH	5	4	16	3	14	1	5	45	6		100
ALL	5	5	15	3	15	1	6	44	5		100

NB9394	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	4	6	14	3	17	2	7	41	6		100
Hgeo	5	17	4	3	23	3	8	30	7		100
SSLgeo	1	15	2	3	27	5	11	32	3		100
ALLgeo	4	5	15	3	16	2	6	42	7		100
SSL	1	15	2	3	28	5	11	32	3		100
nonH	5	4	17	3	15	1	5	43	7		100
ALL	4	5	15	3	17	2	6	42	7		100

NB9495	01	06	08	12	13	17	NS	OK	rest	xjd	total
Bgeo	4	6	16	5	18	2	6	35	7		100
Hgeo	5	19	4	6	24	3	8	24	7		100
SSLgeo	2	16	3	4	29	4	10	28	4		100
ALLgeo	4	6	17	5	16	2	5	37	7		100
SSL	2	16	2	4	29	4	10	28	4		100
nonH	5	5	18	5	15	2	5	38	7		100
ALL	4	6	17	5	16	2	5	37	7		100

NB9596	01	06	08	12	13	17	??	NS	OK	rest	xjd	total
Bgeo	4	7	16	5	16	2	6	3	36	6		100
Hgeo	5	20	5	5	21	3	7	3	25	7		100
SSLgeo	2	18	2	4	26	4	8	4	30	3		100
ALLgeo	4	6	17	5	15	2	6	2	37	6		100
SSL	2	18	2	4	26	4	8	4	30	3		100
nonH	4	5	19	5	14	2	5	2	37	6		100
ALL	4	6	17	5	15	2	6	2	37	6		100

NB9697	01	06	08	12	13	17	??	NS	OK	rest	xjd	total
Bgeo	4	5	12	2	11	2	38		22	5		100
Hgeo	4	15	3	2	13	2	42		14	5		100
SSLgeo	3	13	1		18	4	40		19	2		100
ALLgeo	4	5	13	2	10	2	38		21	6		100
SSL	3	13	1		18	4	39		19	2		100
nonH	4	4	15	2	10	1	37		22	6		100
ALL	4	5	13	2	11	2	37		21	5		100

3. Precise mathematical description of the geocoding calculations

Hispanic subtypes, for P11 and PUMS:

$$\mathcal{H} = \{\text{Mexican, Puerto Rican, Cuban, Dominican,}$$
$$\text{Central American, South American, Other Hispanic}\}$$

For STF3A table P28, let *english?* denote the categories 'Speak English well, not well or not at all'.

From PUMS (for 1990 Census), Hartford County:

$$\begin{aligned}\widehat{\theta}_\alpha &= \text{proportion of Hispanic noncitizens, subtype } \alpha \in \mathcal{H} \\ &= \text{PUMS}(\alpha, noncit)/\text{PUMS}(\alpha),\end{aligned}$$

where

$$\begin{aligned}\text{PUMS}(\alpha, noncit) &= \text{number Hispanic subtype } \alpha\text{, noncitizens} \\ \text{PUMS}(\alpha) &= \text{number Hispanic subtype } \alpha\end{aligned}$$

from the PUMS data. The numbers are calculated by summing the PWGT1 weights for the 5% samples for PUMAs covering Hartford County.

From STF3A tables (for 1990 Census), tract t:

$$\begin{aligned} H(t) &= \text{all Hispanics (from P15)} \\ H(t,18) &= \text{ over-18 Hispanics (from P15)} \\ H(t,70) &= \text{Hispanic over-70 (from P15)} \\ H_\alpha(t) &= \text{all Hispanics, subtype } \alpha \in \mathcal{H} \text{ (from P11)} \\ H(t,18,06) &= \text{over-18; Spanish at home; english? (from P28)} \end{aligned}$$

and

$$\begin{aligned} N(t) &= \text{total population (from P13)} \\ N(t,18) &= \text{total over-18 population (from P13)} \\ N(t,70) &= \text{all over-70; from P13} \\ N(t,18,01) &= \text{all over-18 noncitizens (from P37)} \\ N(t,18,06) &= \text{all over-18; english? (from P28)} \end{aligned}$$

From geocoding:

$$n(t,d) = \text{number of persons with disq = } d\text{, tract = } t$$

Unobserved:

$$\begin{aligned} h(t,d) &= \text{number of Hispanics disq code } d\text{, tract } t \\ h(d) &= \sum_t h(t,d) = \text{number of Hispanics with disq code } d \end{aligned}$$

Estimates of proportions for tract t:

$$\widehat{p}(t,d) = \begin{cases} H(t,18)/N(t,18) & \text{for } d \neq 01,\ 06,\ 08,\ \text{OK} \\ H(t,70)/N(t,70) & \text{for } d = 08 \\ H(t,18,06)/N(t,18,06) & \text{for } d = 06 \end{cases}$$

$$\widehat{p}(t,\text{OK}) = \frac{H(t,18) - H(t,18,06)}{N(t,18) - N(t,18,06)}$$

$$\widehat{p}(t,01) = \sum_{\alpha\in\mathcal{H}} \frac{\widehat{\theta}_\alpha H_\alpha(t) H(t,18)}{H(t) N(t,18,01)}$$

All proportions are truncated to lie between 0 and 1.

Estimate for expected number of Hispanics with disqualification code d:

$$\widehat{h}(d) = \sum_t n(t,d)\widehat{p}(t,d).$$

In particular,

$$\widehat{h}(01) = \sum_{\alpha\in\mathcal{H}} \widehat{\theta}_\alpha \widehat{h}_\alpha(01) \qquad \text{with} \qquad \widehat{h}_\alpha(01) = \textstyle\sum_t X_\alpha(t) Y(t) n(t,01)/Z(t),$$

where

$$\begin{aligned} X_\alpha(t) &= H_\alpha(t)/H(t) \\ Y(t) &= H(t,18)/N(t,18) \\ Z(t) &= N(t,18,01)/N(t,18). \end{aligned}$$

As noted above, when I calculated the $\widehat{h}(01)$ in this way, using PUMS data, I got figures very close to (but always slightly smaller than) what I got by merely applying the over-18 tract proportions of Hispanics to disqualification 01 counts, so I decided to use the simpler method for the tabulations in this report. I have left the description of the more involved calculation in the Report for the benefit of anyone who wishes to adapt my methods to other situations.

Appendix

Error analysis

1. Systematic error and sampling error

The Bureau of the Census gives very clear explanations of the sorts of errors that can arise in survey sampling. Almost the same explanations apply to the types of data and estimation treated in my report.

From Appendix III of the *Statistical Abstracts of the United States, 1993 edition* (a standard sourcebook from the Bureau of the Census):

> Wherever the quantities in a table refer to an entire universe, but are constructed from data collected in a sample survey, the table quantities are referred to as *sample estimates*. In constructing a sample estimate, an attempt is made to come as close as is feasible to the corresponding universe quantity that would be obtained from a complete census of the universe. Estimates based on a sample will, however, generally differ from the hypothetical census figures. Two classifications of errors are associated with estimates based on sample surveys: (1) *sampling error*—the error arising from the use of a sample, rather than a census, to estimate population quantities and (2) *nonsampling error*—those errors arising from nonsampling sources. As discussed below, the magnitude of the sampling error for an estimate can usually be estimated from the sample data. However, the magnitude of the nonsampling error for the estimate can rarely be estimated. Consequently, actual error in an estimate exceeds the estimated error in the estimate.
>
> The particular sample used in a survey is only one of a large number of possible samples of the same size which could have been selected using the same sampling procedure. Estimates derived from the different samples would, in general, differ from each other. The *standard error* (SE) is a measure of the variation among the estimates derived from all possible samples. The standard error is the most commonly used measure of the sampling error of an estimate. . . .

Later in the same section:

> All surveys and censuses are subject to nonsampling errors. Nonsampling errors are two kinds—*random and nonrandom.* Random nonsampling errors arise because . . . Random nonresponse errors usually, but not always, result in an understatement of sampling errors and thus an overstatement of the precision of survey estimates. Estimating the magnitude of nonsampling errors would require special experiments or access to independent data and, consequently, the magnitudes are seldom available.
>
> Nearly all types of nonsampling errors that affect surveys also occur in complete censuses. Since surveys can be conducted on a smaller scale than censuses, nonsampling errors can presumably be controlled more tightly. Relatively more funds and effort can perhaps be expended towards eliciting responses, detecting and correcting response error, and reducing processing errors. As a result, survey results can sometimes be more accurate than census results.

And later:

> For an estimate calculated from a sample survey, the total error in the estimate is composed of the sampling error, which can usually be estimated from the sample, and the nonsampling error, which usually cannot be estimated from the sample. The total error present in a population quantity obtained from a complete census is composed of only nonsampling error.

The bottom line is that

(i) a sample can be better than a large census

(ii) there are two sources of error that need to be considered.

If the sampling error is much smaller than known systematic errors, it can be misleading to quote just standard errors (or variances, which are squares of standard errors) without mention of the systematic errors.

The two methods that I have used, surname matching and geocoding, are based on independent Census data, and they work with different fields from the JIS records. (Also I have data for more than four years of the jury selection process, but the changes from year to year are definitely subject to systematic effects due to aging of the Census data and also effects related to quality of the source lists.) Together the two methods provide a check on nonsampling errors, as mentioned in the second quote.

From this point onwards, the Appendix is written for Statisticians.

2. Geocoding

I will estimate standard errors, as measures of variability, only for the Hgeo estimates. Similar measures could be calculated for Bgeo, but I will omit the tabulations because the legal challenge for the Rodriguez trial is centered on Hispanic representation.

Use the notation introduced in Section 4 of Appendix C.

There are two sources of random error that need to be considered for each table of estimates. One is the variability of $h(d)$, the number of Hispanics disqualified with code d, about its expectation $\mathbb{E}h(d)$. The second is the variability of $\widehat{h}(d)$, the estimate of $\mathbb{E}h(d)$, about $\mathbb{E}\widehat{h}(d)$. The bias $\mathbb{E}h(d) - \mathbb{E}\widehat{h}(d)$ is systematic error. The tables in the next Section give estimates for the standard error of the Hgeo estimated counts for each disqualification code.

Both standard errors are small enough that they have little effect on the interpretation of the percentage breakdowns of Hispanic disqualifications. For example, a standard error of about 30 (as for code 13 in HNB) for a total count of about 4000 represents about three-quarters of a percentage point. Thus random fluctuations might contribute at most a one or two percentage point change for the larger disqualification categories. We are in the situation where the random fluctuations due to sampling are less important than any systematic errors.

Modelling assumptions

For tract t and disqualification code d, the $h(t, d)$ are a simple random sample of size $n(t, d)$ from an "eligible population" containing a proportion $p(t, d)$ of Hispanics.

Systematic errors

(i) changes in underlying population proportions since 1990 Census

(ii) smoothing over tracts ignores within-tract variability

(iii) modelling approximations

(iv) possible selection-bias due to geocoding (SSL shows this effect is small)

(v) changes over time; bias $\sum_t n(t, d)\,(p(t, d) - \mathbb{E}\widehat{p}(t, d))$

(vi) approximation of HNB by Hartford County PUMAs

Estimates of variance

Condition on geocoded values. Estimates for the variance (conservative upper bounds):

$$\widehat{\text{var}}(h(d)) = \sum_t n(t,d)\,\widehat{p}(t,d)\,(1-\widehat{p}(t,d))$$

The square roots of the $\widehat{\text{var}}(h(d))$ are the estimated standard errors for the variation of the $h(d)$ about their expected values under the model.

The $\widehat{p}(t,d)$ are proportions. In general, if $\widehat{p} = \text{num}/\text{denom}$, where num counts the number of individuals with some property out of a total of denom, and for sampling with replacement,

$$\widehat{\text{var}}(\widehat{p}) = \widehat{p}\,(1-\widehat{p})\,/\text{denom}$$

For sampling without replacement, the variance estimator is smaller; the righthand side gets multiplied by a correction factor that is smaller than 1. (See Appendix C of most Census tabulations—such as the tract tables cited near the start of Section 6—or a standard book on sampling, such as Hansen, Hurwitz & Madow 1953.) The usual estimates are

$$\begin{aligned}\widehat{\text{var}}(\widehat{h}(d)) &= \sum_t n(t,d)^2\,\widehat{\text{var}}(\widehat{p}(t,d)) \\ &= \sum_t n(t,d)^2\,\widehat{p}(t,d)\,(1-\widehat{p}(t,d))\,/\text{denom}(d,t),\end{aligned}$$

where $\text{denom}(d,t)$ stands for whichever of the $N(\cdot)$ counts appears as the denominator defining the proportion.

ESTIMATED STANDARD ERRORS FOR THE WHOLE HNB JUDICIAL DISTRICT

HNB9293	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	NS	OK
h(disq)	10				4	15		10		1	6	10	28	6			9	3	9	17	31
$\widehat{h}$(disq)	1					3		3				1	5				1			2	8

HNB9394	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	NS	OK
h(disq)	9				3	13		9			7	9	26	5			8	2	7	15	27
$\widehat{h}$(disq)	1					2		3				1	5				1			2	6

HNB9495	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	NS	OK
h(disq)	10				4	15		10		1	8	12	32	6			11	2	7	17	29
$\widehat{h}$(disq)	1					3		3			1	1	7				1			2	7

HNB9596	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK
h(disq)	10				4	16		10		1	9	11	29	7			12	2	6	15	12	29
$\widehat{h}$(disq)	1					3		4			1	1	5				1			2	1	7

HNB9697	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK
h(disq)	7				2	11		7			7	5	19	4			7	1	3	28		18
$\widehat{h}$(disq)						2		2					3							5		2

ESTIMATED STANDARD ERRORS FOR THE HARTFORD TOWN

HAR9293	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK
h(disq)	7				3	10		6		1	4	6	24	3			8	2	6		15	20
$\widehat{h}$(disq)	1					2		2					5				1				2	5

HAR9394	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK
h(disq)	7				2	8		6			4	5	23	3			7	2	5		13	18
$\widehat{h}$(disq)						1		2					4								1	4

HAR9495	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK
h(disq)	7				3	9		6		1	5	7	28	4			10	2	4		14	19
$\widehat{h}$(disq)						2		2					7				1				2	4

HAR9596	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK
h(disq)	7				3	10		7		1	5	6	25	4			10	2	4	12	10	20
$\widehat{h}$(disq)						2		2					5				1			1	1	4

HAR9697	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK
h(disq)	5				2	7		5			4	3	17	2			6	1	2	21		12
$\widehat{h}$(disq)						1		1					2							4		2

ESTIMATED STANDARD ERRORS FOR THE NEW BRITAIN TOWN

NB9293	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK
h(disq)	6				1	8		5			3	5	11	3			3	1	3		7	13
$\widehat{h}$(disq)						1		2					2								1	3

NB9394	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK
h(disq)	5				2	7		4			3	4	10	3			3	1	3		6	11
$\widehat{h}$(disq)						1		1					1									2

NB9495	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK
h(disq)	5				2	8		5			4	6	11	3			4	1	3		6	12
$\widehat{h}$(disq)						2		2					2								1	3

NB9596	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK
h(disq)	5				2	9		5			4	5	10	3			4	1	2	6	4	12
$\widehat{h}$(disq)						2		2					1							1		3

NB9697	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	99	??	NS	OK
h(disq)	4				1	6		3			3	2	7	2			3		1	12		7
$\widehat{h}$(disq)						1		1					1							2		1

3. Surname matching

A similar analysis can be carried out for surname matching (SSL). As with geocoding, it is the possible systematic error that is more troublesome. Accordingly, it is rather misleading to quote estimated standard errors if one wishes to give a reasonable idea of the variability that should be expected in SSL estimates.

Luckily I have an alternative way to assess the variability: I have a large sample (the questionnaire data) for which I can test the SSL estimates against Hispanic self-identification.

I carried out a resampling experiment (Monte Carlo) to determine the variability in

ratio = (true Hispanic count)/(SSL estimate).

For each of replication, I took a sample of size 1000 (without replacement) from the jurors who had answered the Hispanic question on the questionnaires. With 5000 replications the marginal distributions had similar spreads and location:

summary stats	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
SSL estimates	24.74	39.04	42.94	43.05	46.94	66.74
actual numbers	22	39	43	43.48	48	69

If one is using the SSL as an estimate of the actual Hispanic proportions, it is more useful to have an idea of the variability in the ratio. Here are the percentiles from the sampling experiment:

Pct	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95
Percentile	0.86	0.89	0.91	0.93	0.95	0.96	0.97	0.99	1.00	1.01	1.02	1.03	1.05	1.06	1.08	1.09	1.11	1.14	1.18

The ratio had a distribution centered slighly above 1, with an interquartile range (which corresponds to about 1.35 standard deviations for the normal) about 0.17. The distribution is slightly skewed to the right. For a population similar to the questionnaire sample, the SSL estimate has only slight systematic error. Apparently there is a cancellation effect that balances out false positives with false negatives.

HISTOGRAM OF RATIOS FROM SAMPLING EXPERIMENT