

Racial imbalance in Nassau County public schools

Topics covered: Binomial random variable. Central Limit Theorem. Extra-Binomial variation. Standardizing sample proportions.

Key words: Binomial random variable. Histogram. Normal plot. Sample proportions. Scatter plot. Side-by-side boxplots. z -score.

Data file: `lischool.dat`

On May 17, 1954, the United States Supreme Court issued its landmark *Brown vs. Board of Education of Topeka, Kansas* ruling, which declared that racially segregated public schools were inherently unequal. This ruling invalidated the “separate but equal” justification for segregated schools.

As a result of the ruling, massive programs designed to integrate public schools were begun throughout the United States. Despite this, there is a common public perception that public schools in the 1990’s are still strongly segregated by race. Is this, in fact, true? Does the racial distribution in public schools reflect the general distribution in the community at large, or are there differences (reflecting *de facto* racial imbalance)?

The data examined here represent the proportion of white enrollment in the 56 school districts in Nassau County (Long Island, New York), for the 1992–1993 school year, as reported in the May 20, 1994, issue of *Newsday*. Given the total number of students enrolled in a given school district (n_i , for school district i), the number of white students (w_i) can be viewed as a **Binomial random variable**; that is,

$$w_i \sim \text{Binomial}(n_i, p_i),$$

where p_i is the true probability of a randomly selected student in school district i being white.

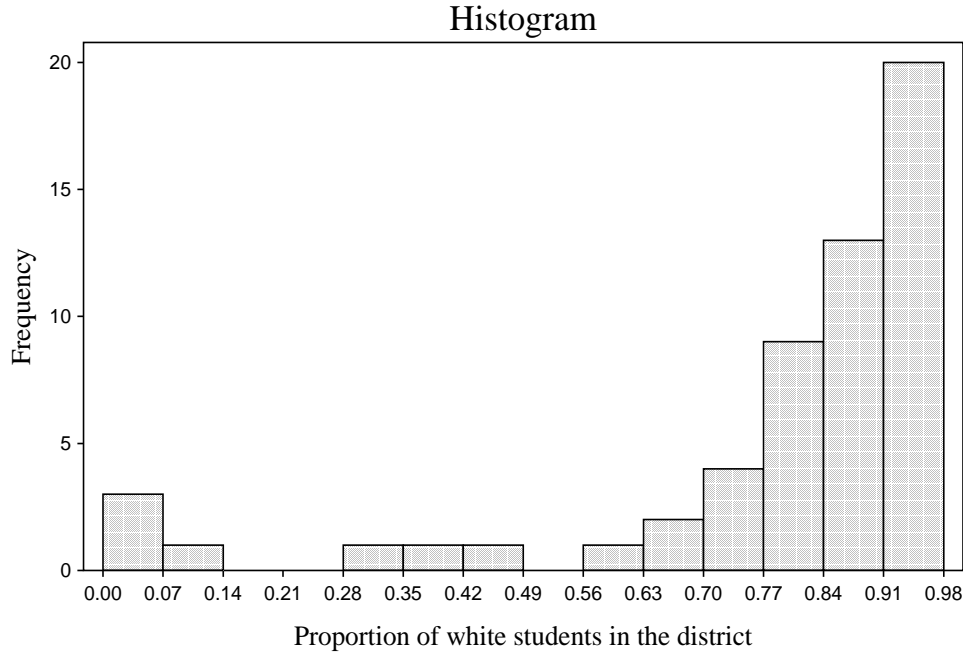
If school districts were perfectly integrated over the entire county, p_i would be the same for all school districts (say p_0). In 1992–1993, there were 174,556 students enrolled in Nassau County public schools, of whom 130,349 were white. Thus, a reasonable value for p_0 would be the overall ratio of white students, or

$$p_0 = \frac{130,349}{174,556} = .7467.$$

Note that since p_0 is based on the actual numbers for the entire county, it is not an estimate of the overall probability of a student being white, but rather the actual value itself.

If p_i did equal p_0 for all i , we would expect that the observed proportion of white students for each district ($\bar{p}_i = w_i/n_i$) would be reasonably close to p_0 . Thus, the distribution of values of \bar{p}_i is informative about whether integration is suggested in the county. We must recognize, of course, that even if the actual probability of a randomly selected student being white *was* constant, we would expect some variation in the values of \bar{p}_i , because of random fluctuation.

Here is a histogram of the observed values of \bar{p}_i for the 56 Nassau school districts:



This is not the kind of picture we would expect to see under conditions of racial balance. The distribution is not at all centered on p_0 , as we would expect, and there is a pronounced left tail, with a small mode at a very small proportion of white students (these are the Hempstead, Roosevelt, Uniondale and Westbury school districts).

It might be thought that this histogram alone is enough to decide that racial imbalance is occurring in Nassau County, but that is not quite true. The problem is that we have not taken into account the inherent variability of the estimates \bar{p}_i . School districts with small enrollments will have values of \bar{p}_i that have much greater variability than school districts with large enrollments. For example, if the school districts whose values fall in the left tail all had very small enrollments, the large difference between \bar{p}_i and p_0 could be just because of random fluctuation. Let's make this more explicit. Since $w_i \sim \text{Binomial}(n_i, p_i)$, we know that

$$E(w_i) = n_i p_i$$

and

$$\text{Var}(w_i) = n_i p_i (1 - p_i).$$

This implies that

$$E(\bar{p}_i) = p_i \tag{1}$$

and

$$\text{Var}(\bar{p}_i) = p_i (1 - p_i) / n_i. \tag{2}$$

Thus, the smaller n_i is, the larger the variance of \bar{p}_i , meaning that it would be less unusual that it is farther away from p_0 (if p_i equaled p_0 for all i).

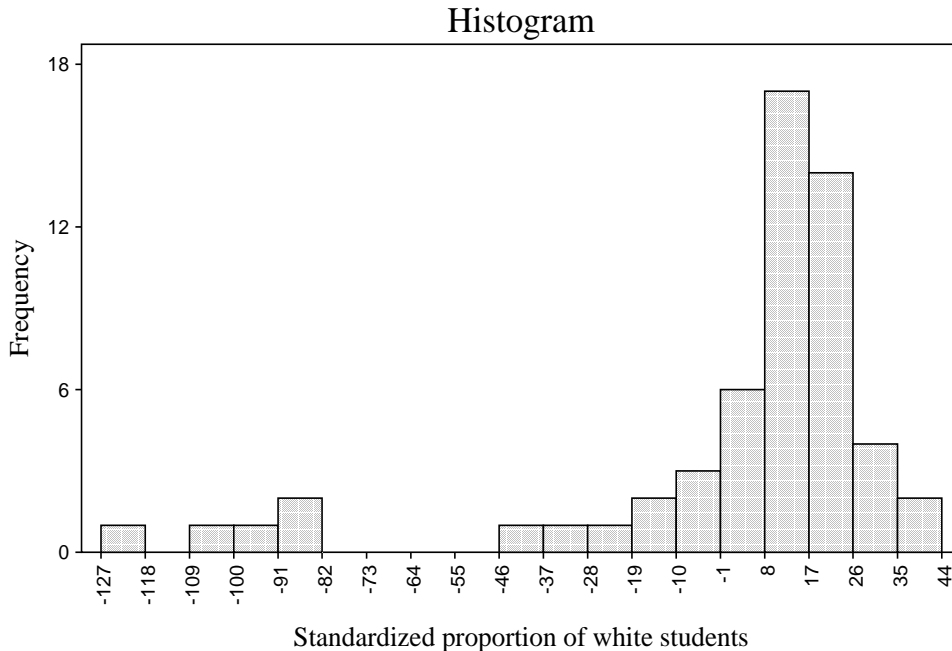
We can say even more about the distribution of \bar{p}_i . By the Central Limit Theorem, not only does \bar{p}_i have the mean and variance given in equations (1) and (2), respectively,

it also (roughly) follows a normal (Gaussian) distribution. Since the values of n_i are in the thousands for these data, we would expect that the normal approximation would be quite accurate. Thus, if racial balance did hold, the values

$$z_i = \frac{\bar{p}_i - p_0}{\sqrt{p_0(1 - p_0)/n_i}} \quad (3)$$

would follow a standard normal distribution. Thus, examination of these z -scores allows us to see if racial balance seems reasonable for these data.

Here is a histogram of the z -scores:

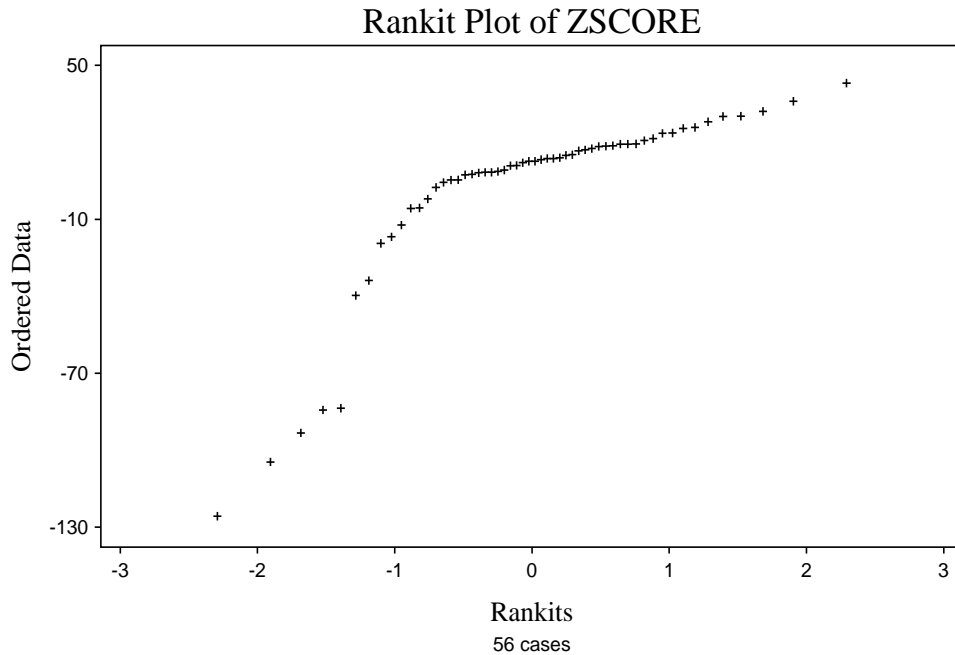


This picture gives strong evidence of racial imbalance, for two reasons. First, if the z_i followed a standard normal distribution, we would expect them to fall roughly in the interval $(-2.5, 2.5)$. This is not even remotely true here, with the observed values as high as 43, and as low as -125 . In addition, the appearance of the histogram is quite non-normal, with a pronounced left tail, corresponding to school districts with unusually low white student percentages.

A very useful plot that can be used to identify potential non-normality even more clearly is the **normal plot** (sometimes also called a *rankits* plot). In this plot, the ordered values of the variable being examined are plotted along the vertical axis, with the expected values of these values (assuming that they came from a normal distribution) being plotted along the horizontal axis. If the data values are close to normally distributed, the plot will roughly follow a straight line.

On the next page is the normal plot for the z -scores.

The non-normality of the z -scores is quite apparent. The long left tail of their distribution shows up in the plot as the “bend” in the straight line at about -0.7 (which corresponds to observed values being more negative than would have been expected under normality).



The long tails of the z -score distribution show that the actual variability of the z -scores is considerably greater than what would be expected based on the (normal approximation to the) Binomial distributions with constant p_i (each school district's count has a different n_i , of course). This greater-than-expected variation is often called **extra-Binomial variation**. One common cause of extra-Binomial variation is when a set of data is modeled as being based on Binomial distributions with constant p (as is being done here) when it actually represents a mixture of several different Binomial distributions (with different values of p).

We also need to recognize an implication of the large enrollments for the school districts, which is a result of equation (3). Consider the Herricks school district, which has an enrollment of 3432. Based on the form of the z -score, this large sample size implies that even small differences from p_0 in practical terms could lead to large z -scores. The observed proportion of white students in Herricks is 70.4%, which is not very far from the overall proportion of 74.7%; however, the z -score for Herricks is approximately

$$z = \frac{.704 - .747}{\sqrt{.747 \times .253 / 3432}} = -5.79,$$

which would be very unlikely under a normal distribution. For this reason, we need to decide if the observed racial imbalance is important from a practical point of view, even though we are already convinced (based on the observed z -scores) that it exists.

We can do this by examining the differences between \bar{p}_i and p_0 for the 56 school districts (called PDIFF here), and the absolute values of those differences (called ABSDIFF).

Here are some summary statistics:

DESCRIPTIVE STATISTICS

	PDIFF	ABSDIFF
MEAN	0.0316	0.1875
SD	0.2495	0.1657
MINIMUM	-0.7440	0.0149
1ST QUARTILE	0.0237	0.0894
MEDIAN	0.1208	0.1545
3RD QUARTILE	0.1810	0.2027
MAXIMUM	0.2292	0.7440

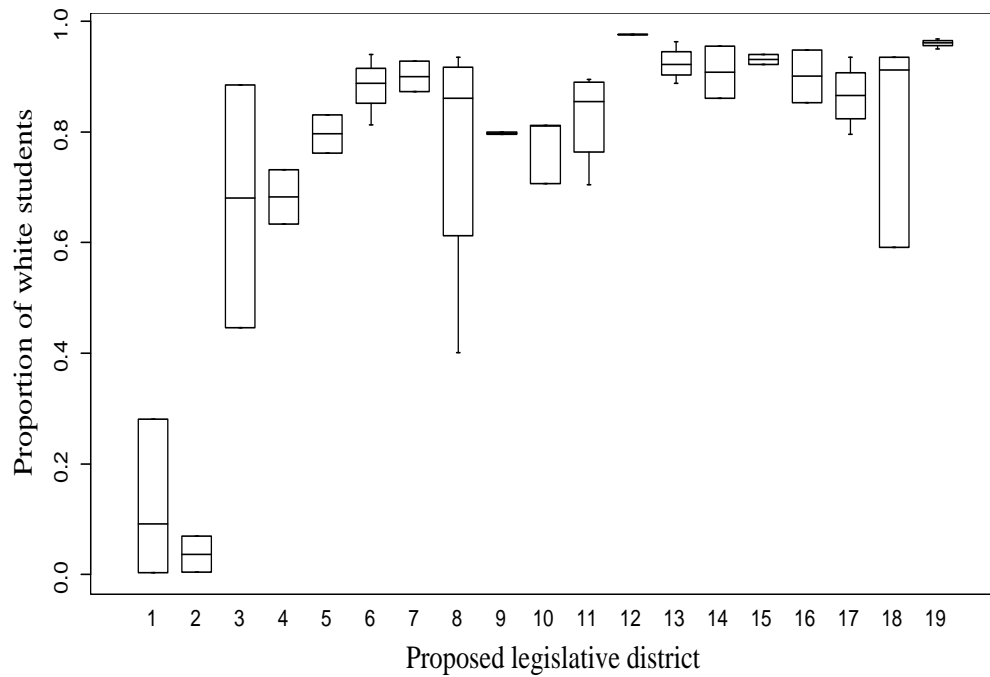
It is apparent that the observed imbalance is of practical importance. Observed proportions of white students range from .744 less than the overall proportion to .2292 greater than the overall proportion. Even more worrying is that about 75% of the districts have an absolute difference from the overall proportion of at least .09, indicating strong racial imbalance.

From where could this imbalance be coming? The most natural explanation is that Nassau County itself is not characterized by a uniform racial distribution, but rather by individual neighborhoods that are predominantly white or nonwhite. If regions of racial homogeneity could be identified, then the extra-Binomial variation could be at least partially accounted for by estimating p_i by the proportion of white students for the appropriate region, rather than for the county as a whole (p_0).

A first attempt at identifying such regions can be motivated by a recent court case. On June 30, 1993, Federal District Court Judge Arthur D. Spatt determined that the structure of the Nassau County government was illegal, due to violation of the principle of "one person, one vote." He ordered that the government be reorganized away from the current Board of Supervisors system. A Commission on Government Revision was set up to oversee this reorganization. On March 29, 1994, the Commission presented its proposal to form a 19-district County Legislature, which it subsequently formally recommended on May 24, 1994. The districts for this legislature were created to be consistent with the Voting Rights Act, which requires that minorities have a reasonable opportunity to elect representatives of their choice.

These 19 districts can be used as a classification of the districts into (perhaps) more racially homogeneous regions. If that is the case, then the z -scores that are calculated based on equation (3), using the proportion of white students for the entire district as the true p_i (in place of p_0), should be more closely normally distributed. This would then confirm the geographical source of the racial imbalance.

Is this a reasonable hypothesis to pursue? Here are side-by-side boxplots of the proportion of white students for the 56 districts, separated by the proposed legislative district number. Note that most proposed legislative districts have only 2 – 4 school districts in them, so we can only get general impressions from the boxplots:



Obviously there is a good deal of variation between districts. The school districts in proposed legislative districts 1 and 2 apparently have 0 – 20% of the students being white; those in proposed legislative districts 3, 4 and 5 have roughly 60 – 80% students being white; while the other proposed legislative districts often have over 80% of the students being white.

Descriptive statistics separated by proposed legislative district make this more explicit:

Proposed legislative district	Total enrollment	White enrollment	Proportion white
1	13885	2158	0.1554
2	8345	222	0.0266
3	11266	7210	0.6400
4	7829	5319	0.6794
5	7828	6183	0.7899
6	11092	9790	0.8826
7	8857	8094	0.9139
8	8522	6696	0.7857
9	3535	2818	0.7972
10	11527	8969	0.7781
11	8907	7112	0.7985
12	6678	6517	0.9759
13	13261	12087	0.9115
14	8220	7342	0.8932
15	8686	8125	0.9354
16	9415	8433	0.8957
17	10278	8763	0.8526
18	6830	5333	0.7808
19	9595	9178	0.9565

The entries given under “Proportion white” in the table above can now be used as the true p_i for school districts in the given proposed legislative district.

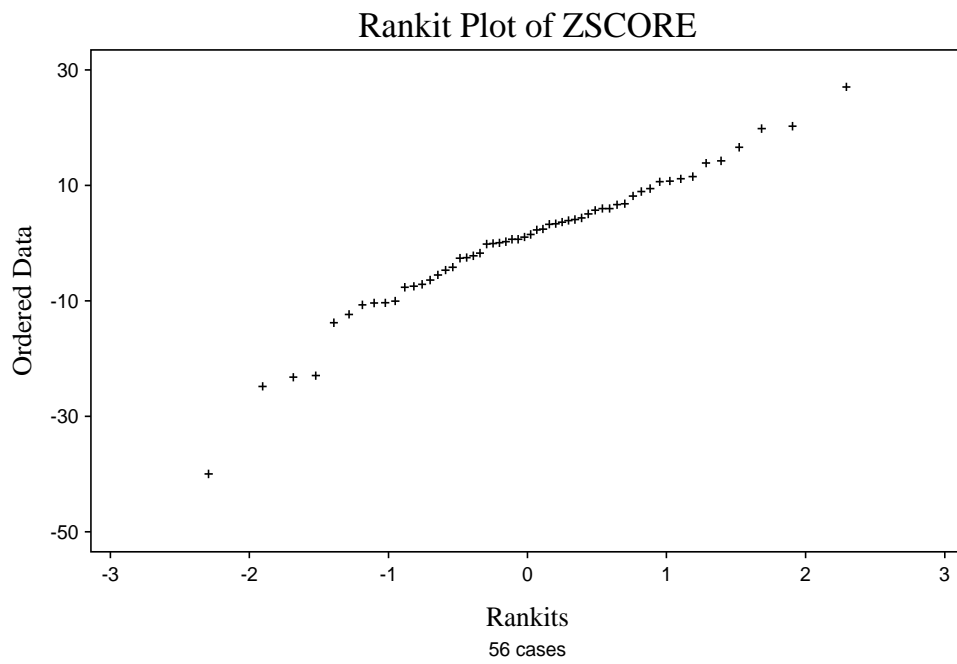
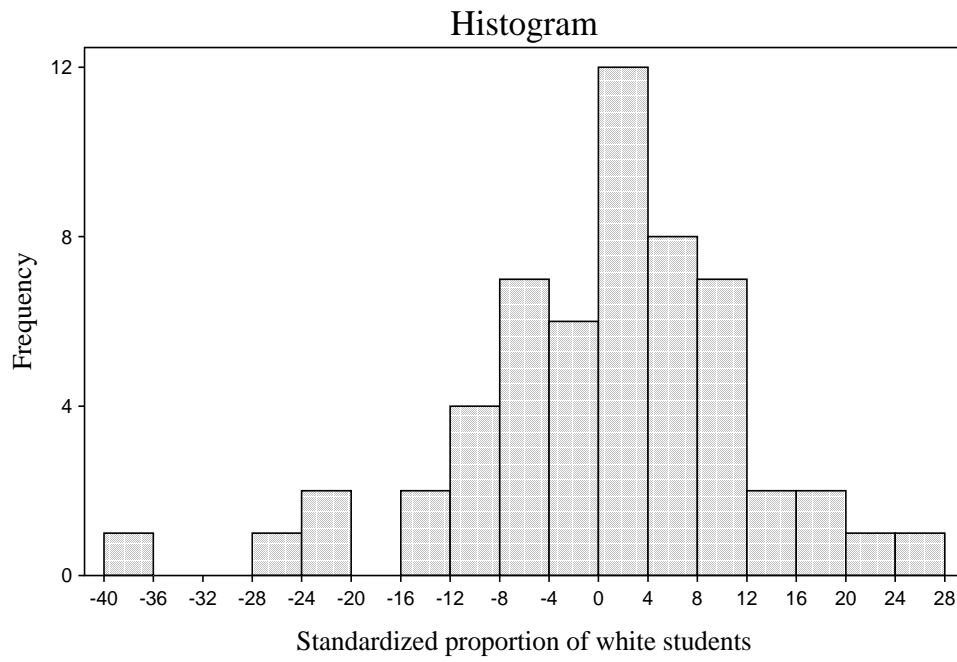
On the next page are a histogram and normal plot of the new z -scores for the school districts.

The values look much more normally distributed, but the scale is still too wide for a standard normal (by a factor of about ten or so). Once again, this could possibly reflect only that the sample sizes are so large that even small differences between a school district’s proportion of white students and the proposed legislative district’s proportion lead to large z -scores. Descriptive statistics can help determine that:

DESCRIPTIVE STATISTICS

	PDIFF	ABSDIFF
MEAN	4.510E-03	0.0618
SD	0.0937	0.0701
MINIMUM	-0.3851	0.0000
1ST QUARTILE	-0.0300	0.0164
MEDIAN	4.936E-03	0.0405
3RD QUARTILE	0.0489	0.0796
MAXIMUM	0.2446	0.3851

The observed differences are considerably smaller than before, but many are still too large. The school district’s proportion is as much as .3851 less than the proposed legislative district’s proportion, and as much as .2446 greater. Still, about 75% of the differences are less than .08 in absolute value, as compared with about 75% being *greater* than .09 previously. Thus, this regionalization of the districts has accounted for much of the racial imbalance, but not all of it.



Where does this geographical model fail? Here are the PDIFF values for the 56 school districts, along with the associated proposed legislative district (LEGISLAT):

CASE	DISTRICT	LEGISLAT	PDIFF
1	Baldwin	5	-0.028
2	Bellmore	19	0.003
3	Bellmore - Merrick	19	-0.007
4	Bethpage	17	0.082
5	Carle Place	11	0.086
6	East Meadow	13	-0.023
7	East Rockaway	6	0.057
8	East Williston	11	0.096
9	Elmont	3	-0.194
10	Farmingdale	14	-0.032
11	Floral Park	8	0.112
12	Franklin Square	3	0.245
13	Freeport	1	0.125
14	Garden City	8	0.149
15	Glen Cove	18	-0.190
16	Great Neck	10	0.034
17	Hempstead	2	-0.023
18	Herricks	11	-0.095
19	Hewlett Woodmere	7	-0.014
20	Hicksville	17	-0.057
21	Island Park	7	-0.042
22	Island Trees	15	-0.014
23	Jericho	17	0.026
24	Lawrence	4	0.052
25	Levittown	15	0.004
26	Locust Valley	18	0.131
27	Long Beach	4	-0.046
28	Lynbrook	6	0.032
29	Malverne	8	-0.385
30	Manhasset	9	-0.001
31	Massapequa	12	0.000
32	Merrick	19	0.011
33	Mineola	11	0.026
34	New Hyde Park	9	0.002
35	North Bellmore	13	0.015
36	North Merrick	13	0.005
37	North Shore	18	0.154
38	Oceanside	7	0.014
39	Oyster Bay - East Norwich	17	-0.001
40	Plainedge	14	0.062
41	Plainview - Old Bethpage	16	0.052
42	Port Washington	10	-0.071
43	Rockville Centre	5	0.041
44	Roosevelt	1	-0.153
45	Roslyn	10	0.033
46	Seaford	19	0.004
47	Sewanhaka	3	0.040
48	Syosset	16	-0.043
49	Uniondale	1	-0.065
50	Valley Stream 30	6	-0.070
51	Valley Stream 13	6	0.029
52	Valley Stream 24	6	-0.019
53	Valley Stream Central	6	-0.031
54	Wantagh	13	0.051
55	West Hempstead	8	0.038
56	Westbury	2	0.042

Three school districts with particularly large (absolute) values of PDIFF are Elmont (unusually low proportion of white students), Franklin Square (unusually high

proportion of white students) and Malverne (unusually low proportion of white students). Each of these anomalies can be explained geographically. Elmont is a community on the county border with New York City, near neighborhoods with relatively low white populations. Franklin Square is in proposed legislative district 3, but borders district 8, which has considerably higher proportion of white students. Malverne is in proposed legislative district 8, but is wedged between districts 2 and 3, which have considerably lower proportions of white students. Glen Cove, in proposed legislative district 18, also has a large (absolute) value of PDIFF, which is probably because of the large size of the district (it is roughly 10 miles wide from east to west, when the entire county is only about 15 miles wide from east to west), and the resultant heterogeneity. Thus, if the proposed legislative districts had been drawn slightly differently, the observed deviations from the mixture of Binomials model would be even less.

Summary

Examination of the distribution of the proportion of white students in the 56 school districts of Nassau County indicates strong racial imbalance. This is supported by observed extra-Binomial variation, which is often the result of trying to model a set of sample proportions as being from Binomial random variables with constant probability of success, rather than as a mixture of Binomials with different probabilities of success. This possibility is investigated for these data, based on separating the districts according to proposed county legislative districts. This geographical effect accounts for much of the observed extra-Binomial variation, but not all of it. Some of the largest observed errors based on the effect can be accounted for by the specific ways the proposed legislative districts were drawn, since slight changes in the lines would reduce the errors.

Technical terms

Binomial random variable: a discrete random variable that is often used to model data that come as a set of binary trials. If one of the two outcomes is arbitrarily termed a “success,” then the random variable is appropriate as a representation of the number of successes in a given number of trials if the outcomes are independent from trial-to-trial, and the probability of success is constant from trial-to-trial.

Extra-Binomial variation: an observed pattern where counts that could be reasonably modeled as coming from a Binomial random variable (or a set of Binomial random variables with constant probability of success) exhibit more variation than would be expected under such a model. One common source of extra-Binomial variation is when the counts come from a mixture of several different Binomial random variables (with different probabilities of success), rather than Binomials with the same probability of success.

Normal plot: a graphical procedure for examining the normality of a variable, it is also known as a **rankits plot**. It also can be used for identifying outliers. The plot is obtained by plotting the ordered values of the variable against their expected values if they came from a normal population. The plot should appear to be more or less a straight line. Outliers are identified as points far removed from this approximate straight line, particularly at the extremes. Long tails can be identified by a pattern of points deviating off a straight line at either end.