

Computing and Presenting Adjusted Rates for Health Utilization Research

William Kreuter¹

Senior Computer Specialist
Back Pain Outcome Assessment Team
Dept. of Health Services
University of Washington
1107 NE 45 St., Suite 427
Seattle, WA 98105-4631

billyk@u.washington.edu

May 31, 1996

¹The helpful suggestions of Brad McNeney and Prof. Paula Diehr, Biostatistics Dept., University of Washington, are gratefully acknowledged.

Abstract

Health services researchers frequently find it useful to compute rates of incidence of diseases or treatments and to compare these rates across time, geographic location, category of disease or treatment, or other factors of interest. Rates are usually "standardized" for age, sex and possibly other factors, in order to help avoid spurious inferences due to population anomalies or dissimilarities between subject populations. Standardization is accomplished by "direct adjustment" and "indirect adjustment". For many applications, the SAS data language and PROC tools do not provide a straightforward method for computation of direct or indirect rates. This paper illustrates some of these situations and the author's approaches to their solution. Results are given both as plain-text output (PROC TABULATE and PROC PRINT), and also as graphics produced by PROC GPLOT and PROC GMAP. To help avoid spurious inferences due to small population sizes, standard errors of indirect rates are also computed and presented.

1 Theory of Rate Adjustment

The rate of incidence of a disease, treatment, or other event or characteristic is often discussed in the epidemiological and health-services literature. In its simplest form, a rate is the number of occurrences of the event or characteristic to the at-risk population. The numerator, events, is typically a sum of occurrences during a defined time period (often one year). The denominator is usually given as the instantaneous at-risk population at an arbitrary point during the period of accumulation of the numerator—*e.g.*, July 1 of a calendar year, or some other date readily available from census or other public records.

This kind of ratio, which describes the overall disease incidence without comparison to other rates or other time periods, is called a *crude rate*. For example, suppose that a researcher is investigating back pain, finding that there are fifty hospitalizations for this disease in the year and region of interest, and that the population in the region for that year is 50,000. The crude rate in this region is 0.001, or 100 events per 100,000 persons.

An excellent article published by the South Carolina Center for Health Statistics describes the crude rate as “the relative frequency with which some event occurs in a study population, and it commonly takes a standard form such as a number per 100,000.” [1, p. 69] The latter number is chosen merely so that for typical diseases the rate, to three or four digits of precision, can be expressed conveniently as an integer.

When might the comparison of two crude rates be a misleading indicator of, say, the occurrence of disease or hospitalization? The text by Fleiss [2, p. 239] used by first-year epidemiology students at the University of Washington states,

If the populations are similar with respect to factors associated with the event under study—factors such as age, sex, race or marital status—there would be no problem in comparing the [crude] rates... as they stand. If the populations are not similarly constituted, however, the direct comparison of [crude] rates may be misleading.

A practical example is given in a recent University of Washington Biostatistics master's thesis by Parvin [3, p. 2], which used back-pain hospitalization data in discussing the adjustment of rates.

For instance, if a county has a large proportion of older people and surgical back pain admissions occur more often in older people, then that county will appear to have “too many” admissions even if its stratum-specific¹ surgical admission rates are equal to those of other counties.

To avoid spurious inferences due to anomalies in the numerator (the event) or to dissimilarities in the denominator (the population makeup) that may be present in subsets of the population of interest, crude rates are usually *standardized* or *adjusted* for age, sex, and sometimes race or other demographic factors. It is important to adjust crude rates whenever rates are to be compared across, for example, time, geographic location, or category of disease or treatment.

Standardization or adjustment is conducted in two possible forms: *direct* and *indirect*. Both methods require a numerator (*i.e.*, number of cases) and denominator (*i.e.*, number of persons) for each *stratum* of the population. A stratum is just the set of persons sharing a control characteristic—typically, gender and some range of age. (Five or ten years are typical age ranges.)

¹The “strata” are the population subsets for creating adjusted or standardized rates, as will be discussed shortly.

Both methods of standardization involve adjusting to a *standard population*. There are relatively few requirements for the standard population itself—only that it have some meaningful conceptual relationship to the rates being compared. A typical standard population might be the 1990 census data for the United States. Rates of disease in, say, Washington State in 1992 and 1994 could then be compared by adjusting the rates for those years to this 1990 national population. Another reasonable standard population might be the population of Washington State for 1994, when performing a *small-area analysis* [4, 5] to compare the rates in 1994 among the state's counties.

The method of direct adjustment is mathematically identical to computing the weighted average rate over the strata. That is, a rate is produced in each stratum, and then the average for all strata is found, weighted by the number of persons in the standard population's corresponding strata. The stratum-specific rates before adjustment are themselves the quotients of the number of persons in the "local" population, divided into the number of cases occurring among that population. Hence the directly adjusted rate is expressed as

$$\frac{\sum_{i=1}^{\# \text{ strata}} \frac{\text{local cases}_i \times \text{standard pop}_i}{\text{local pop}_i}}{\sum_{i=1}^{\# \text{ strata}} \text{standard pop}_i} \quad (1)$$

The indirect adjustment method is not identical to a weighted average, but it's similar. In other respects it's sort of the opposite of direct adjustment; indirect adjustment finds the rate (i.e., the number of cases divided by the population) in each stratum of the *standard* population, and then aggregates these rates, taking into account the local population in the strata. The indirectly adjusted rate is thus

$$\frac{\sum_{i=1}^{\# \text{ strata}} \text{standard cases}_i \times \sum_{i=1}^{\# \text{ strata}} \text{local cases}_i}{\sum_{i=1}^{\# \text{ strata}} \text{standard pop}_i \times \sum_{i=1}^{\# \text{ strata}} \frac{\text{standard cases}_i \times \text{local pop}_i}{\text{standard pop}_i}} \quad (2)$$

Equations (1) and (2) are expressed in terms of events and persons. It is probably more common in the biostatistics and epidemiology literature to take the quotients—that is, the proportions of events to population—as given, and hence to express the formulas for adjusted rates by referring to these proportions as variables. However, the data which I encounter as a SAS programmer working for health-services research faculty usually appear as counts of hospitalizations; I separately obtain census data, which are given as counts of persons. Therefore, I find it more practical with these kinds of data to work from the numerator and denominator constituents of the stratum-specific proportions.

Parvin [3, p. 3] notes, "A way to think of the indirect method is as the ratio of the observed number of occurrences to the expected number of occurrences in the study population multiplied by the crude rate for the standard population." Parvin also notes that indirect standardization is the same as the standardized mortality ratio multiplied by the probability of occurrence in the standard population [3, p. 6].

Both direct and indirect adjustment make use of each of these three data: cases in local area, population in local area and population in standard area. But while local cases need not be known or available by stratum when indirect adjustment is used (as long as the total number of cases for each area is given), they are individually multiplied by population separately in each stratum when computing direct rates. Therefore, if there is some small, perhaps statistically non-significant discrepancy or variation in the number of cases in some strata, the indirectly adjusted rate will be much more modestly affected than the directly adjusted rate.

Consider the example in table 1, adapted from reference [1] ("Example 1," p. 71, and "Example 2," p. 72). We are given the following hypothetical data, which for purposes of illustration are stratified by age but not by sex.

Table 1: Stratum-specific data

Age	Local Population			Standard Population		
	Persons	Having Disease	Rate (per 100,000) in this age stratum	Persons	Having Disease	Rate (per 100,000) in this age stratum
<40	10,000	10	100	1,000,000	1,000	100
40-59	20,000	60	300	3,000,000	9,000	300
≥60	10	9	90,000	500,000	5,000	1,000
Σ	30,010	79		4,500,000	15,000	

The crude rate obtained for the data in table 1 is $79 \div 30,010 = 263$ per 100,000. Applying equations (1) and (2) to the data in table 1, we obtain a directly adjusted rate of 10,222 and an indirectly adjusted rate of 376.

The direct rate in this example is absurd—the stratum of only ten persons completely skews the result for the entire county—but the data producing it were chosen to illustrate why indirect adjustment is sometimes preferred to direct. Suppose that the Table 1 data represent a locality in which, for some reason, the 60-and-over age stratum's population is disproportionately small compared to the comparison standard population. Direct adjustment may wildly exaggerate the (statistically and by common sense) insignificant contribution of this stratum, while indirect adjustment is not so prone to such skewed results.

However, even though indirect adjustment produces results that respond less to small variations in the cases, it is not always possible or even meaningful to use it. This is because indirect adjustment requires data, the number of cases in the standard population, which are not used in direct adjustment. An example in which using direct adjustment might be the only practically possible method would be adjusting Washington hospitalizations to regional experience, if hospitalizations in the rest of the region weren't available. Adjusting Washington admissions in a DRG² to some international standard would be an example of the indirect rate not even being defined, since DRGs aren't used in most other countries. On the other hand, indirect adjustment would be the only available method, and direct would be impossible, if disease incidence were known by stratum in the standard population, but only unstratified disease data were available for the local population.

²DRG $\equiv$ Diagnosis Related Group, a system to classify diseases promulgated for cost-containment purposes by the Medicare system's Health Care Financing Agency

Both direct and indirect adjustment are criticized for other reasons. For example, Fleiss [2, p. 239] states, "If the specific rates vary in different ways across the various strata, then no single method of standardization will indicate that these differences exist. Standardization will, on the contrary, tend to mask these differences." Several practical examples are cited. And for a particular study, my research unit at the University of Washington chose to use a regression model to pursue the same objective as rate adjustment. This was because, when phrased as a statistical prediction problem, a parsimonious regression model is known to perform better in this situation than the regression model corresponding to direct standardization.

2 Statistical Significance of Adjusted Rates

As just discussed, rates are adjusted to obtain a "level playing field" when they are compared. However, for a variety of reasons including the sort of anomalies just described, differences between rates might not be statistically significant. The kinds of significance tests which are performed for statistical measures such as categorical tables are desirable for adjusted rates as well.

Significance measures are available for both direct and indirect rates. The test for a direct rate is outside the scope of this paper. For an indirect rate, provided that the events or diseases are fairly uncommon compared to the size of the population (*i.e.*, the rate is small), an approximation to the standard error of the rate is

$$se \approx \frac{rate}{\sqrt{cases}} \quad (3)$$

where *rate* is the computed indirect rate, and *cases* is the count of events over all strata. This expression needs further modification if more than one case may be attributable to an individual—*e.g.*, if some of the cases represent repeat hospital admissions for a patient [4, 5]. Omitting this modification will not result in serious change if readmissions are relatively rare.

Having computed equation (3), confidence intervals or *p*-values to test the significance of the null hypothesis that some rate is no different from some nominally fixed standard (*e.g.*, a national or regional rate, or the rate in some decennial census year) can be computed similarly to conducting a one-sample test on a binomial proportion.³[6, pp. 219–222] For example, the two-sided *p*-value to test that the rate of a disease in, say, King County, Washington, is different from the statewide Washington rate is

$$2 \times Z_{\left(\frac{King\ County\ rate - Washington\ rate}{se} \right)} \quad (4)$$

In equation (4), *Z* is a normal-table lookup.⁴ For example, *Z*_{1.96} is .025.

If equation (4) produces a *p*-value greater than .05, then the 95% confidence interval for the rate will overlap the statewide rate (or whatever comparison rate you choose). Therefore, the two-sided 95% confidence interval for a rate is

$$rate \pm 1.96 \times se \quad (5)$$

³Closely related would be a χ^2 test on the standardized mortality ratio [6, pp. 226–228], which, as noted, is a component of an indirectly adjusted rate.

⁴Equation (4) can be implemented in a SAS data step as `2*PROBNORM(-ABS(RATE-COMPARE)/SE)` where *RATE* and *COMPARE* are respectively shown as "King County rate" and "Washington rate" in equation (4).

For simultaneous comparison of several rates—for example, a small-area analysis comparing Washington counties to the state experience—it may be appropriate to modify the confidence intervals with a multiple-comparison adjustment, such as the Bonferroni adjustment, so as not to falsely conclude that significant differences exist.

3 SAS Implementation

SAS does not provide any PROC for the purpose of obtaining an adjusted rate. This is perhaps just as well, since the numerator, denominator and stratum data needed for rate adjustment are often maintained in datasets whose structure may vary considerably from problem to problem. Hence even setting up a macro might not be convenient, as the programmer might need to customize the code for every situation.

There are often common elements between rate standardization problems. I frequently find that the event (numerator) data originate in the form of individual hospitalization records, while the population (denominator) data usually are census records for which an observation represents a stratum for an area.

Let's suppose that the data in table 1 were provided in that fashion. Imagine that the table represents County #1 in some state. The permanent dataset DATASETS.ADMITS has one observation for each occurrence of disease. So there are 79 observations for County #1. For County #2, let's have the disease cases be the same as #1, but let's imagine a more reasonable population for the ≥ 60 stratum. The standard population is, naturally, not county-specific.

Since a directly adjusted rate is identical to a weighted average, we can obtain direct standardization from the weighted mean offered by, *e.g.*, PROC TABULATE. However, PROC TABULATE works over one dataset. We first need to aggregate the hospitalization data up to the stratum level. We also need properly aggregated population data, both for the local areas and for the standard population.

Let's assume that we have census data for the local and standard areas already set up in permanent datasets. Then a complete program to obtain a direct rate might be the following.

```
data;
  set datasets.admits;
  select;
    when (age<40) age=0;
    when (age<60) age=40;
    otherwise age=60;
  end;
  keep county age;
proc freq;
  tables county*age / out=cases sparse;
  title 'Example of Directly Adjusted Rates';
  title2 'Admissions data, aggregated to the stratum level';
proc print data=datasets.localpop;
  by county;
  sum localpop;
  format localpop comma6.;
```

```

    title2 'Local population dataset';
proc print data=datasets.stdrdpop;
    sum stdrdpop;
    format stdrdpop comma10.;
    title2 'Standard population dataset';
data;
    merge cases(in=in_cases) datasets.localpop;
    by county age;
    if not in_cases then count=0;
    rate=1e5*count/localpop;
    keep county age rate;
proc sort;
    by age;
data;
    merge _last_ datasets.stdrdpop;
    by age;
    keep county rate stdrdpop;
proc tabulate;
    class county;
    var rate;
    weight stdrdpop;
    tables county,rate*mean='direct'*f=comma6. / rts=8;
    title2 'Results of proc tabulate';
run;

```

Note that the variable AGE changes into a mere label for the stratum. Numerically, it then represents the youngest age in a range. However, the numerical value has no meaning for computational purposes; it's present just so that the aggregated case dataset CASES can be merged with DATASETS.LOCALPOP and DATASETS.STDRDPOP, using AGE as one of the merge variables.

In order to obtain the required single dataset which can then be processed by PROC TABULATE to obtain the weighted-mean average rate over the strata, two merges need to be performed. This is because the local cases and population need to be merged on two variables, COUNTY and AGE, while the standard population does not reference county. The second merge, therefore, copies the standard population variable identically into each stratum.

The user should be aware of the common pitfalls inherent in SAS datastep merges. The examples in this paper would give correct results even if PROC SUMMARY (which lacks a SPARSE keyword) had been used instead of PROC FREQ, and there were no patients (or events) in some county and age stratum. However, possibly difficult-to-detect erroneous results can arise if rates are to be disaggregated by some variable in the underlying hospitalization data which is not common to the population data—for instance, if the adjusted rates are to be disaggregated by a disease class in addition to (or instead of) the county. The potential problem will occur when there are no patients in some disease class, county and age stratum; this is a perfectly ordinary possibility. In such a case, a "sparse" aggregation and merge is required; it is obtained in the preceding program by use of the SPARSE keyword in PROC FREQ and the test on the IN= variable in the subsequent datastep.

The program produces a .LST file akin to the following:

Example of Directly Adjusted Rates
Cases aggregated to the stratum level

TABLE OF COUNTY BY AGE

COUNTY	AGE			
Frequency				
Percent				
Row Pct				
Col Pct	0	40	60	Total
1	10	60	9	79
	6.33	37.97	5.70	50.00
	12.66	75.95	11.39	
	50.00	50.00	50.00	
2	10	60	9	79
	6.33	37.97	5.70	50.00
	12.66	75.95	11.39	
	50.00	50.00	50.00	
Total	20	120	18	158
	12.66	75.95	11.39	100.00

Example of Directly Adjusted Rates
Local population dataset

COUNTY=1

OBS	AGE	LOCALPOP
1	0	10,000
2	40	20,000
3	60	10

COUNTY		30,010

COUNTY=2

OBS	AGE	LOCALPOP
4	0	10,000
5	40	20,000
6	60	7,500

COUNTY		37,500

Standard population dataset

OBS	AGE	STDRDPOP
1	0	1,000,000
2	40	3,000,000
3	60	500,000
		=====
		4,500,000

Results of proc tabulate

	RATE

	direct
-----+-----	
COUNTY	

1	10,222
-----+-----	
2	236

Note that with the more "reasonable" population of 7,500 for the oldest age stratum in County #2, the direct rate is no longer an outlier.

The program to produce the indirect rate requires somewhat more complicated merging, and cannot be adapted to a weighted-mean feature in quite as straightforward a manner. It also requires the additional data to represent the disease cases in the standard population—that is, the "Having Disease" column on the right-hand side of Table 1. The program follows; datasets not displayed by it are the same as those displayed by the direct-adjustment program.

```
proc summary data=datasets.admits nway;
  class county;
  output out=lclcases;
proc print data=datasets.stdcases;
  sum stdcases;
  format stdcases comma6.;
  title 'Example of Indirectly Adjusted Rates';
  title2 'Cases in standard population';
data standard;
  merge datasets.stdrpop datasets.stdcases;
  by age;
proc summary data=standard;
  var stdrpop stdcases;
  output out=totals sum=totpop totcases;
proc sort data=datasets.localpop out=localpop;
  by age;
data;
  merge standard localpop;
  by age;
  expected=localpop*stdcases/stdrdpop;
  keep county expected;
proc summary nway;
  class county;
  var expected;
  output sum=;
data;
  merge _last_(keep=county expected) lclcases(keep=county _freq_);
  by county;
data;
  if _n_=1 then set totals(keep=totcases totpop);
  set;
  indirect=1e5*totcases*_freq_/(totpop*expected);
  keep county indirect;
proc print;
  format indirect 3.;
  title2 'Indirect rate computed in the datastep';
run;
```

Each of the summations in equation (2) is provided separately here, in many cases by a call to PROC SUMMARY. As with the direct-adjustment program, multiple merges are needed because some of the quantities are local-area-specific and others are not. Only the quantities comprising the variable EXPECTED are stratum-specific, however.

The indirect-adjustment program could also be written to create the variable EXPECTED as a weighted sum from PROC SUMMARY, although I think that the code I use here is more convenient.

It may be somewhat less likely that the indirect adjustment algorithm will cause the erroneous-merge problem discussed above; however, this is still possible if there are no patients in the standard-case dataset in some stratum, if the rates are to be disaggregated by disease class or some other non-population-specific variable. Also, the user should ensure that each county (and, if there is more than one, each disease, analysis year or other cohort) contributes at least one patient of any age to the disease dataset, and that there is at least one person in each stratum of the standard population (which is almost certain). If there is any reason to doubt that a sparse aggregation is unnecessary, then despite the above illustration's using PROC SUMMARY instead of PROC FREQ, the latter should be used.

Both the direct- and indirect-adjustment programs could have been written using PROC SQL, and in many cases that might be a better approach. PROC SQL can replace all the merges with one nested SELECT statement. Depending on the application, it's even possible that the entire program, including all data manipulation, could be written in one call to PROC SQL. Such an approach might be particularly desirable if the merges that your real-life data require are more complicated than what this paper illustrates—for example, if you want one program to compute, display and compare adjusted rates for several disease cohorts or time periods. SQL allows possibly appealingly spartan code for these purposes, although it's possible that execution may be slower than the datastep-merge approach and that the code may be more difficult to understand. Properly constructed SQL also provides another alternative to the erroneous-merge problem solved in this paper with PROC FREQ's SPARSE keyword.

Here's the output from the indirect-adjustment program.

Example of Indirectly Adjusted Rates Cases in standard population

OBS	AGE	STDCASES
1	0	1,000
2	40	9,000
3	60	5,000
		=====
		15,000

Indirect rate computed in the datastep

OBS	COUNTY	INDIRECT
1	1	376
2	2	182

As discussed previously, unlike the direct rate, County #1's indirect rate is not wildly different from the general population's disease experience. For County #2, which lacks the population anomaly for the over-60 group, both the direct and the indirect adjustments are agreeably on the same order of magnitude. The crude rate for County #2 is $79 \div 37,500 = 211$ per 100,000.

Finally, note that if the underlying data are provided as proportions or as stratum-specific crude rates, the appropriate program will look quite a bit different—for one thing, probably needing fewer and less complicated merges. (Most likely, though, somebody somewhere needed to do some merges to produce the proportions from original case data.) However, the overall logic flow would still largely be the same.

4 Displaying the Results Graphically and Cartographically

Attached to this article are three graphics: two maps of Washington State and an output from PROC GPLOT. The rates portrayed represent hospitalization experience of county residents for some disease, in some year, under study at a research university in Seattle. Each county has its rate indirectly adjusted to the statewide experience for the same year. The standard error of the indirect rate is also computed using the techniques described above.

The first map shades each Washington county according to whether there is a significant difference at $\alpha = .05$ from the state crude rate, and, if so, whether the county rate is higher or lower than the state's. The second map does not address the question of significance, but it divides all 39 Washington counties into four quartiles (using a LEVELS=4 option in PROC GMAP's CHORO statement).

Notice that the highest- and lowest-rate counties do not necessarily have rates significantly different from the state rate, while some significant rates fall in the medium-quartile counties. This is because significance of an indirect rate is in part a function of number of disease cases, and there is much more incidence of disease in the larger counties, which nevertheless might not have exceptionally large or small rates.⁵ On the other hand, the counties with small populations are more likely to have unusually high or low rates, even though it is less likely that their rates will be significant.

The third graphic is a plot of each county's rate and confidence interval, arranged in order of rate. This combines the information from both maps, allowing the viewer to see which counties are particularly high or low, and also which are significant (because the confidence intervals do not overlap the state crude rate value). The maps have the advantage of showing which regions and which adjacent counties are tending towards high or low rates.

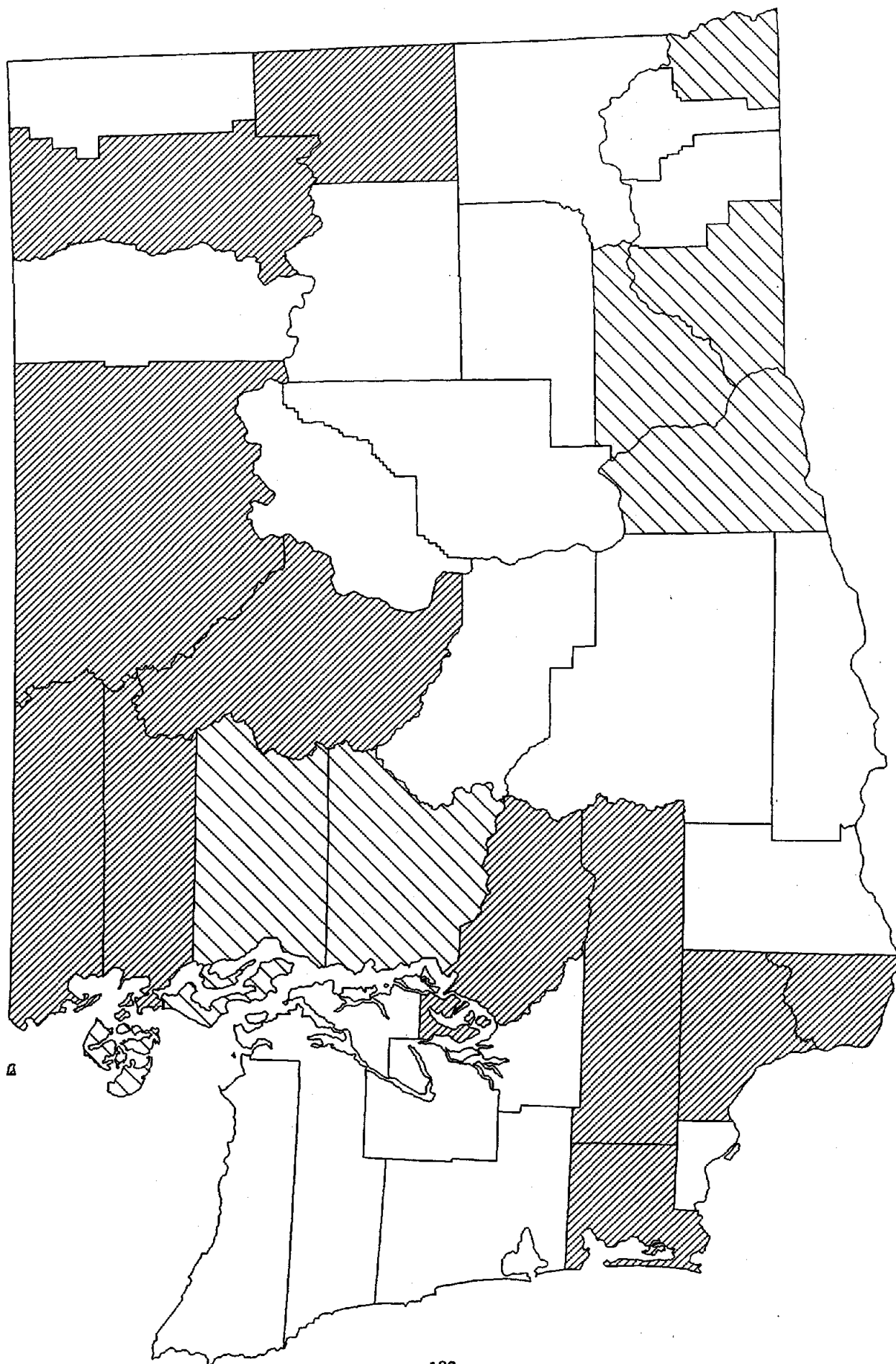
No correction for multiple comparisons was applied to the map or bar plot, even though one could make a case that some of the conclusions of county-rate significance are a Type I error, found merely by chance due to the large number of rates being examined.

⁵The large counties are somewhat unlikely to be very different numerically from the state, since that would influence the state rate itself. King County, Washington, for instance, which includes Seattle, has a fairly large portion of all hospital admissions in Washington.

References

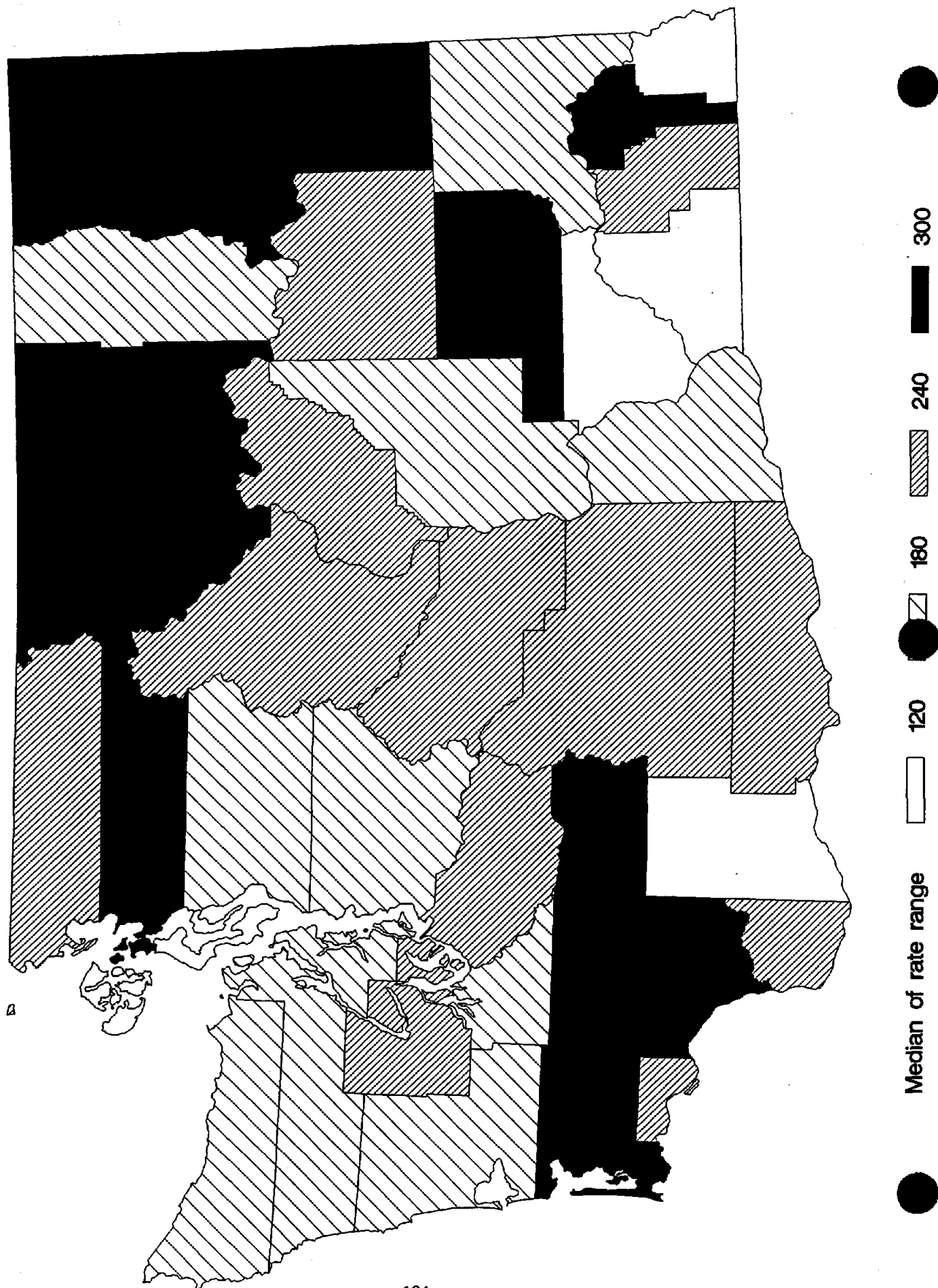
- [1] State Center for Health Statistics. *Statistical Primer*. South Carolina Division of Health Services, Dept. of Human Resources, PO Box 2091, Raleigh, NC 27602. Vol. 1, Number 1, undated.
- [2] Joseph L. Fleiss. *Statistical Methods for Rates and Proportions*. John Wiley and Sons, New York, second edition, 1981.
- [3] Astrid Brigitte Parvin. An examination of the relative performance of direct and indirect methods of standardization. Master's thesis, University of Washington, 1994.
- [4] Cain K and Diehr P. Testing the null hypothesis in small area analysis. *Health Services Research*, 27:267-294, 1992.
- [5] Diehr P, Cain K, Kreuter W, and Rosenkranz S. Can small-area analysis detect variation in surgery rates?: the power of small-area variation analysis. *Medical Care*, 30:484-502, 1992.
- [6] Bernard Rosner. *Fundamentals of Biostatistics*. Duxbury Press, Boston, second edition, 1986.

Statistical Difference from Statewide Rate



Difference from state mean of 200.0 less N/S greater

County Rate Quartiles



Rate Confidence Intervals, Ordered by Rate

