Effect of Using True Variability for Baseline Cancer Rates

1.0 Introduction

Statistics is the cornerstone of many aspects of life but is mistrusted by many people. It is easy to see how statistics gets a bad rap ... because it is complicated, so few people really understand it, and it is easy to unintentionally or intentionally misinterpret data. One example of misinterpretation is investigated below. Most of us are familiar with the concept of calculating the average (mean or expectation value) and the standard deviation of a set of data. But what is done if there is only one data point. One approximation is to use the Poisson distribution which approximates the normal distribution for large numbers. The mean of a single data point, x, is x. And the standard deviation of the single data point, x, is the square root of x, √x. This approximation should only be used if there is only one estimate or measurement (**theoretical method**). If there is a set of measured values, then conventional parametric statistics should be used to calculate a mean and standard deviation of the distribution (**empirical method**). Unfortunately, this requirement is often overlooked in community health studies where census tract data is compared to county data. The theoretical method is usually used to calculate county (baseline) parametric statistics. The empirical method should be used, since the larger variability of all individual county census tracts is known.

Before investigating this problem, we will discuss polling statistics to illustrate concepts of confidence intervals and variability.

2.0 Polling Statistics

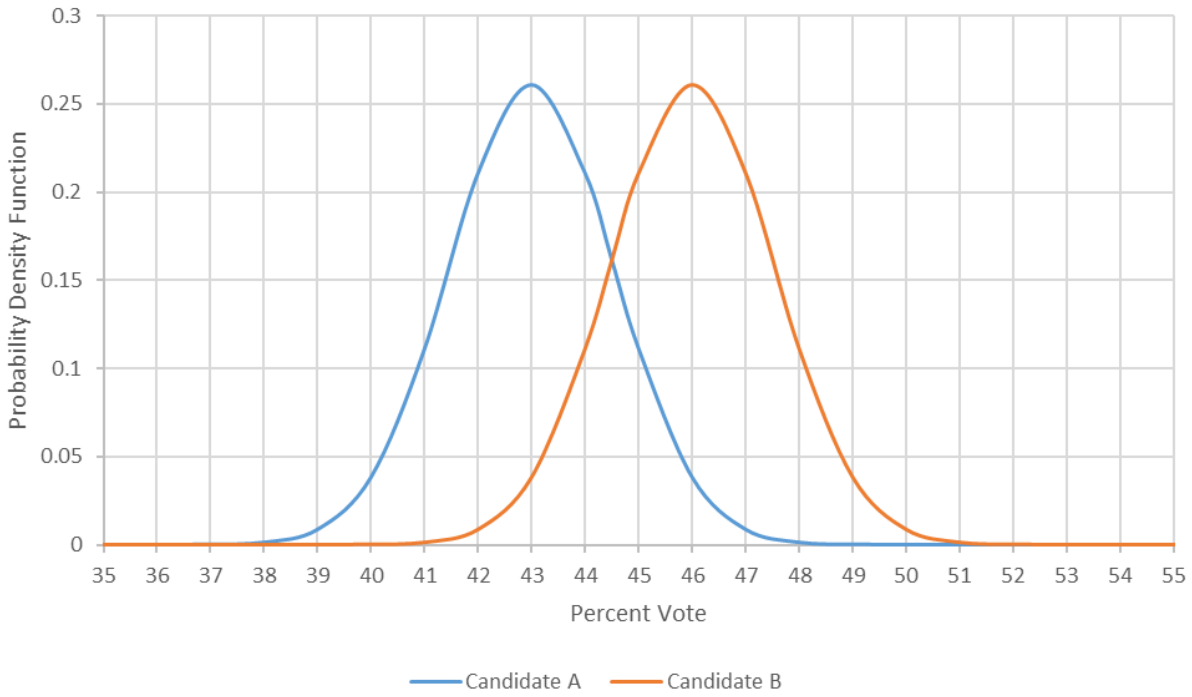
The U.S. has just survived another 18-month election, and was daily inundated with polling results. The TV pundits who show these polls will (sometimes) point out that most polls have a margin of error of +/- 3%. What does +/- 3% mean?

If candidate A receives 43% in a poll, then the true result is 43% +/- 3% or between 40% and 46%. The more correct statistical explanation is that … there is a 95% probability that the true result will be somewhere between 40% and 46%. This range is also known as the 95% confidence interval (CI)[1]. So, what does "true result" mean? It means that … if 100 similar polls were taken, then 95 out of 100 results would lie between 40% and 46%. That seems like lot, but it is the same as saying that 1 in 20 poll results would lie outside that range. The implication is that, all things being equal (and that is an important caveat), the ultimate result of a national vote will also give candidate A between 40% and 46% of the vote.

[1] For a normal distribution, the 95% confidence interval is defined as +/- $1.96\sigma_r$, where σ_r is the standard deviation

If the same poll results in candidate B receiving 46%, this likewise means, 46% +/- 3% or between 43% and 49%. At first glance, candidate B would appear to be leading by 3%, however the two confidence intervals also overlap by 3%. See Figure 1 for the two normal distributions for each candidate.

Figure 1. Candidate Poll



The two probability density functions (pdf) of each candidate overlap at 44.5%. The probability that candidate A would win in a sample poll is half the area under both curves or 32.2%/2 = 16.1%. Extrapolating to a national vote, and presuming that the sample poll was correctly representative of a national vote, there is an 83.9% chance that candidate B would win and a 16.1% chance that candidate A would win. These poll results and predicted probability of winning the election (rounded up) closely mirrored those of Donald Trump (candidate A) and Hillary Clinton (candidate B) before the polls closed on November 8th. The ultimate election result clearly demonstrated that the polls were not representative of the national vote and that the electoral college system complicates election predictions.

In order that a candidate "win" a poll, he/she should lead by more than twice the margin of error, or 2 x 3 = 6%. This would ensure a greater than 95% probability of winning the ultimate vote. The reason for the uncertainty or variability is that the poll is usually a very small sample (usually 1,000 voters) of the much larger population (over 100 million). The larger the sample size, the smaller the margin of error or confidence interval becomes.

Similar problems with statistical variability is seen in the interpretation of cancer registry data.

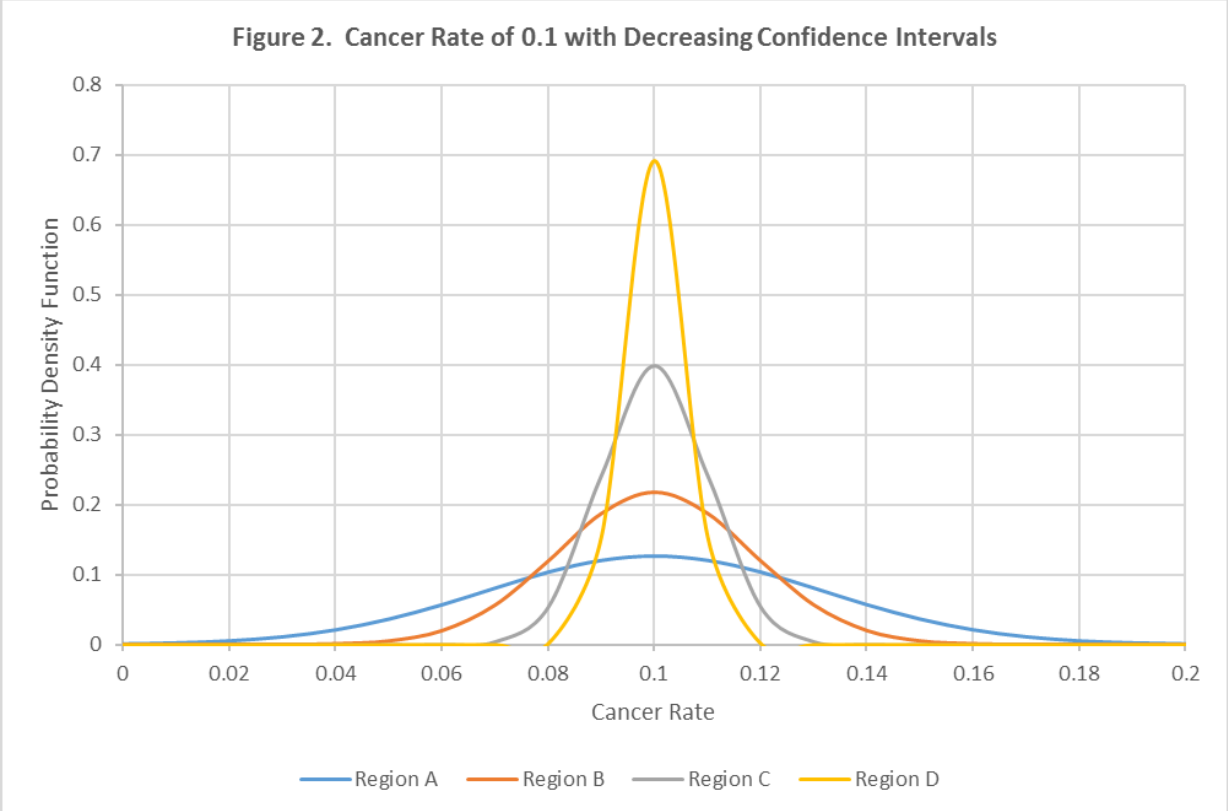
3.0 Cancer Statistics

The relative standard deviation (also called coefficient of variation) of a region's cancer rate (σ_r/r) is equal to the inverse square root of the number of cancer cases. Table 1 illustrates a variety of regions with increasing populations from 100 to 3,000. Assume the cancer rate remains the same at 0.1, and therefore the number of cancers in each region increases from 10 to 300.

Table 1. Relative Standard Deviation as Function of Inverse Square Root of No. of Cancers

Region	No. of Cancers	Population	Cancer Rate	Standard Deviation of No. of Cancers	Standard Deviation of Cancer Rate	Relative Standard Deviation (st. dev. / rate)	Coefficient of Variation (%)
	c	p	$r = c / p$	$\sigma_c = c^{1/2}$	$\sigma_r = \sigma_c / p$	$\sigma_r / r = c^{-1/2}$	$100 * \sigma_r / r$
A	10	100	0.1	3.2	0.032	0.32	32%
B	30	300	0.1	5.5	0.018	0.18	18%
C	100	1,000	0.1	10.0	0.010	0.10	10%
D	300	3,000	0.1	17.3	0.006	0.06	5.8%

The effect of reducing standard deviations and relative standard deviations as the number of cancers increases. Figure 2 visually shows the shrinking standard deviation for the four regions as the number of cancers increases.



The U.S. National Cancer Institute (NCI) publishes cancer statistics for the U.S., states and counties. Tables 2a and 2b show, as an example, bladder cancer incidence data extracted from NCI's web site for the U.S., all U.S. states, and all California counties, for the 5-year period, 2009 to 2013. Data columns show …

- State or California county
- Central estimate (CE) and $1\sigma_r$ confidence interval (CI)[2] of the age-adjusted bladder cancer incidence rate per year per 100,000 population. The confidence interval is provided in parentheses as a binary pair representing the lower and upper confidence limits (CL) of the central estimate
- Average number of incidence cases per year
- Population

The cancer rate data is given to three significant figures. The rate for the United States is 20.7 per year per 100,000. The confidence interval is (20.6, 20.7). This is an extremely narrow CI and implies that the rate for the U.S. is known with extreme accuracy. This is very misleading and untrue. The small CI is purely a function of the large population of the U.S., and the corresponding relatively large number of cancer incidence cases.

The effect of population on the confidence interval can be seen be inspecting the various state data in Table 2a. The most populous state, California, with a population of over 37 million, has a relatively small CI of 18.2 (18.0, 18.4). In contrast, Wyoming with the lowest population of 563,858, has a larger CI of 22.6 (20.9, 24,4).

Turning to Table 2b, the low populous California county of Mariposa has a population of 18,216 and an even larger CI of 23.9 (16.8, 34.0). Inspection of the remaining Cis for states and counties show a similar inverse functional dependence of CI to population.

Another interesting comparison is the U.S., Indiana and Santa Barbara County.

Cohort	**Population**	**Central Estimate (CI)**
US	309,266,000	20.7 (20.6, 20.7)
Indiana	6,488,511	20.7 (20.2, 21.2)
Santa Barbara County	423,620	20.7 (18.9, 22.7)

All three areas have the same best estimate of bladder cancer incidence rate of 20.7. However, as the population decreases, the CI increases.

[2] For a normal distribution, the 68% confidence interval is defined as +/- $1\sigma_r$, where σ_r is the standard deviation

4.0 Theoretical Calculation of Standard Deviation and Confidence Intervals

A simple theoretical method of calculating the standard deviation of a number of counts, in this case number of bladder cancer incidence cases, is the Poisson approximation …

$$\sigma_c = c^{1/2}$$

$$\sigma_r = \sigma_c / p$$

where σ_c = standard deviation of number of cancers
σ_r = standard deviation of cancer rate
c = number of cancers
p = population

Confidence intervals can be calculated for $1\sigma_r$ as is done here or any other multiple of σ_r. For instance, the +/- 95% CI is simply +/- $1.96\sigma_r$. The Poison is a good approximation for large numbers, otherwise other methods can be employed, for instance the gamma distribution approximation.

Calculation of lower and upper CL is illustrated below for several states and counties, to demonstrate agreement with NCI data. Surprisingly, the calculations above for a $1\sigma_r$ CI match exactly, to 3 significant figures, to the CIs that the NCI claims to represent the 95% CI based on the $1.96\sigma_r$. This has been communicated to NCI.

New York

raw rate = cases * 100,000 / population = 5,176 * 100,000 / 19,402,641 = 26.677

age factor = central estimate / raw rate = 23.5 / 26.677 = 0.88091

standard deviation (σ_r) = (age factor * 100,000 * $(cases)^{1/2}$) / population
= (0.88091 * 100,000 * $5{,}176^{1/2}$) / 19,402,641 = 0.32664

lower $1\sigma_r$ CL = central estimate - σ_r = 23.5 - 0.32664 = 23.2

upper $1\sigma_r$ CL = central estimate + σ_r = 23.5 + 0.32664 = 23.8

Indiana

raw rate = cases * 100,000 / population = 1,473 * 100,000 / 6,488,511 = 22.702

age factor = central estimate / raw rate = 20.7 / 22.702 = 0.91181

standard deviation (σ_r) = (age factor * 100,000 * $(cases)^{1/2}$) / population
= (0.91181 * 100,000 * $1{,}473^{1/2}$) / 6,488,511 = 0.53934

lower $1\sigma_r$ CL = central estimate - σ_r = 20.7 - 0.53934 = 20.2

upper $1\sigma_r$ CL = central estimate + σ_r = 20.7 + 0.53934 = 21.2

California

raw rate = cases * 100,000 / population = 6,684 * 100,000 / 37,326,524 = 17.907

age factor = central estimate / raw rate = 18.2 / 17.907 = 1.0164

standard deviation (σ_r) = (age factor * 100,000 * (cases)$^{1/2}$) / population
= (1.0164 * 100,000 * 6,684$^{1/2}$) / 37,326,524 = 0.22262

lower $1\sigma_r$ CL = central estimate - σ_r = 18.2 - 0.22262 = 18.0

upper $1\sigma_r$ CL = central estimate + σ_r = 18.2 + 0.22262 = 18.4

Los Angeles County

raw rate = cases * 100,000 / population = 1,521 * 100,000 / 9,834,062 = 15.467

age factor = central estimate / raw rate = 16.4 / 15.467 = 1.0603

standard deviation (σ_r) = (age factor * 100,000 * (cases)$^{1/2}$) / population
= (1.0603 * 100,000 * 1,512$^{1/2}$) / 9,834,062 = 0.41925

lower $1\sigma_r$ CL = central estimate - σ_r = 16.4 - 0.41925 = 16.0

upper $1\sigma_r$ CL = central estimate + σ_r = 16.4 + 0.41925 = 16.8

5.0 Empirical Calculation of Standard Deviation and Confidence Intervals

Use of **theoretical** approximations, like the Poisson approximation, to estimate standard deviation and confidence intervals should only be done when empirical data is unavailable. **Empirical** CIs can be calculated using conventional parametric statistics of individual US states and state counties, as shown in the lower portions of Tables 1a and 1b.

Using the US state data in Table 1a, the **theoretical** US bladder cancer incidence rate is 20.7 (20.6, 20.7). The narrow confidence interval implies that most Americans can expect a rate between 20.6 and 20.7. However, inspection of the state rates shows that this is patently untrue. State rates range from 13.7 (12.9, 14.5) to 29.3 (28.1, 30.6). The **empirical** US rate is 20.3 (16.3, 24.3) using a +/- $1\sigma_r$ CI (68%), and 20.3 (12.4, 28.1) using a +/- $1.96\sigma_r$ CI (95%).

Using the California County data in Table 1b, the **theoretical** California state bladder cancer incidence rate is 18.2 (18.0, 18.4). The narrow confidence interval implies that most Californians can expect a rate between 18.0 and 18.4. However, inspection of the county rates shows that this is patently untrue. County rates range from 10.7 (8.6, 13.3) to 34.1 (21.3, 52.9). The **empirical** California rate is 20.3 (16.3, 24.3) using a +/- $1\sigma_r$ CI (68%), and 20.3 (12.4, 28.1) using a +/- $1.96\sigma_r$ CI (95%).

Clearly empirical confidence intervals are always much larger, and more representative of contributing states or counties, than theoretical confidence intervals.

Cancer registry data is also available at the census tract level. Thus, for instance, Los Angeles County is comprised of approximately 1,300 census tracts. Populations of these census tracts are obviously much smaller that their corresponding counties and therefore the census tract confidence intervals are correspondingly larger.

6.0 Community Health Studies

Community health studies are often performed in communities surrounding industrial facilities including refineries, hazardous waste site, landfills, nuclear plants, and other facilities where there is a concern for off-site health effects. Frequently, these studies will calculate cancer rates in a small number of census tracts surrounding the facility and compare those rates to the county rates. If the confidence interval of the census tract is greater than, and does not overlap, the county confidence interval, then the census tract is claimed to have a higher cancer rate than the county average.

An alternate comparison method is to divide the census tract rate (and CI) by the county average rate (and CI), to give a "rate ratio" (and associated CI). Depending on the investigator, census tract rate ratios exceeding 1.5 or 2.0 will be indicative of increased cancer rates. Some investigators may (or may not) acknowledge that if the confidence interval includes the null value, 1.0, then results are statistically inconclusive.

The county average and CI would be designated the "baseline cancer rate" with the implication that every census tract in the county should have a rate consistent with the county rate and CI, unless impacted adversely by environmental releases from the facility under investigation. If environmental releases and subsequent exposures from the facility are either non-existent or too small to be measured, the investigators may be obliged to add a waiver to their conclusions saying that *"there is no evidence that increased cancer rates are a result of operations at the facility."* But nevertheless, the implication remains.

The previous review of state and county cancer rates illustrates the wide disparity between theoretical and empirical (real) cancer rate confidence intervals. The size of cancer rate confidence intervals is extremely important in community health studies and resulting implied community health impacts. This is illustrated below in the form of a **hypothetical** health study.

7.0 Hypothetical Humboldt Bay Nuclear Power Plant Community Cancer Study

Humboldt Bay Nuclear Power Plant (HBNPP) was a 63 MWe nuclear reactor owned by Pacific Gas and Electric Company. It was the seventh commercial nuclear power plant to be licensed in the U.S. and operated from August 1963 to July 1976. Decommissioning is progressing.

A hypothetical community cancer study cancer study could be performed using the data in Table 2b. Four different outcomes are possible using combinations of theoretical and empirical California cancer rates and CIs, and +/- $1\sigma_r$ (+/- 68%) and +/- $1.96\sigma_r$ (+/- 95%) CIs. See Tables 3a, 3b, 4a and 4b.

Theoretical California +/- $1\sigma_r$ (+/- 68%) - See Table 3a

Method 1.

Humboldt County 28.5 (24.7, 32.8)
California 18.2 (18.0, 18.4)
Humboldt County rate exceeds California rate by more than 50%
Humboldt County CI does not overlap California CI

Method 2.

Rate ratio of Humboldt / CA 1.57 (1.32, 1.81)
Rate ratio exceeds 1.5
Rate ratio CI does not include 1.0

Empirical California +/- $1\sigma_r$ (+/- 68%) - See Table 3b

Method 1.

Humboldt County 28.5 (24.7, 32.8)
California 20.3 (16.3, 24.3)
Humboldt County rate exceeds California rate by less than 50%
Humboldt County CI does not overlap California CI

Method 2.

Rate ratio of Humboldt / CA 1.40 (1.05, 1.76)
Rate ratio does not exceed 1.5
Rate ratio CI does not include 1.0

Theoretical California +/- $1.96\sigma_r$ (+/- 95%) - See Table 4a

Method 1.

Humboldt County 28.5 (19.9, 37.1)
California 18.2 (17.8, 18.6)
Humboldt County rate exceeds California rate by more than 50%
Humboldt County CI does not overlap California CI

Method 2.

Rate ratio of Humboldt / CA 1.57 (1.09, 2.04)
Rate ratio exceeds 1.5
Rate ratio CI does not include 1.0

Empirical California +/- 1.96σ_r (+/- 95%) - See Table 4b

Method 1.

Humboldt County 28.5 (19.9, 37.1)
California 20.3 (12.4, 28.2)
Humboldt County rate exceeds California rate by less than 50%
Humboldt County CI overlaps California CI

Method 2.

Rate ratio of Humboldt / CA 1.40 (0.71, 2.09)
Rate ratio does not exceed 1.5
Rate ratio CI includes 1.0

Clearly, use of the correct empirical California cancer rate and +/- 1.96σ_r (+/- 95%) CI produce a very different conclusion than using the theoretical California cancer rate and +/- 1σ_r (+/- 68%) CI. See Figures 3 and 4 illustrating comparison of the two distributions for California and Humboldt County, one using the theoretical confidence interval and one using the empirical confidence interval. The distributions on Figure 3 appear vastly different and it would be understandable to conclude that Humboldt County has elevated bladder cancer rates. However, using the more correct empirical confidence interval in Figure 4, we see a very different picture with distributions with a large overlap of distributions. It would be harder to credibly conclude that Humboldt County has statistically significant increased bladder cancer rates.

See also Figure 5, illustrating comparison of the theoretical and empirical rate ratios. The theoretical confidence interval significantly exceeds the null value of 1.0 with very little overlap. In contrast, the empirical confidence interval is not only closer to the null value with a mean that is less than 1.5, but the there is considerable overlap of the null value. Indeed, the lower 95% confidence limit of 0.71 is well below the null value.

8.0 Conclusions

Many, if not most, community health studies incorrectly use the theoretical CI for the baseline cancer rate which tends to exaggerate differences between local and baseline cancer rates, and therefore potentially reach incorrect conclusions.

Figure 3. Cancer Rates for Humboldt County and California Using the Theoretical California Confidence Interval

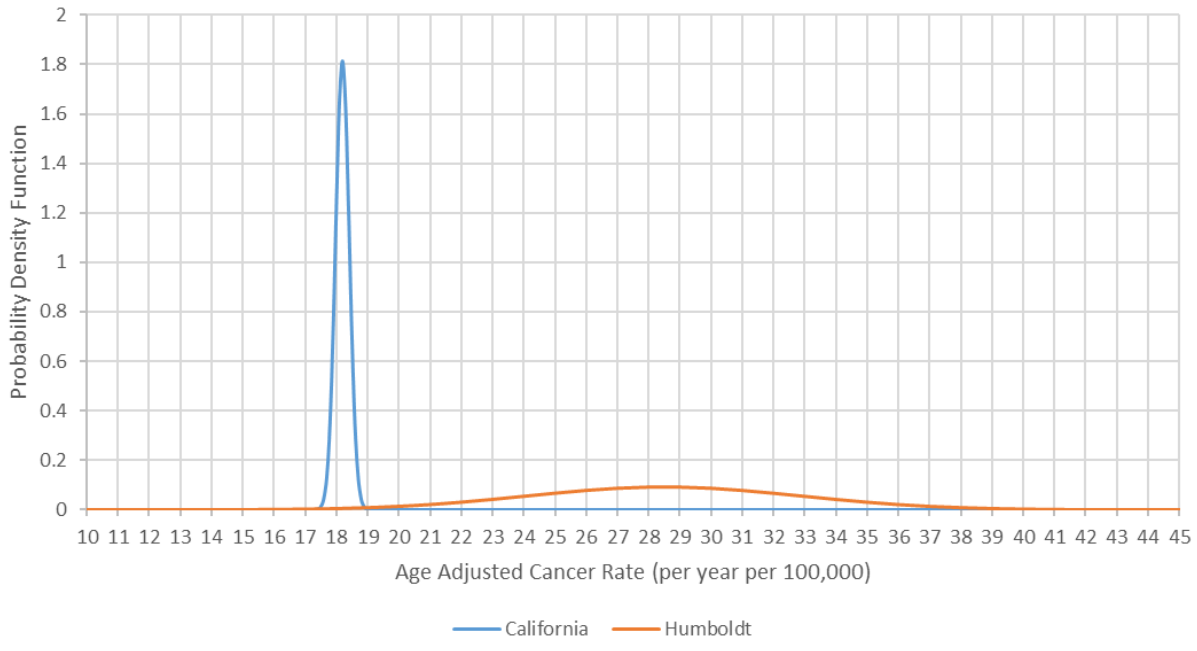


Figure 4. Cancer Rates for Humboldt County and California Using the Empirical California Confidence Interval

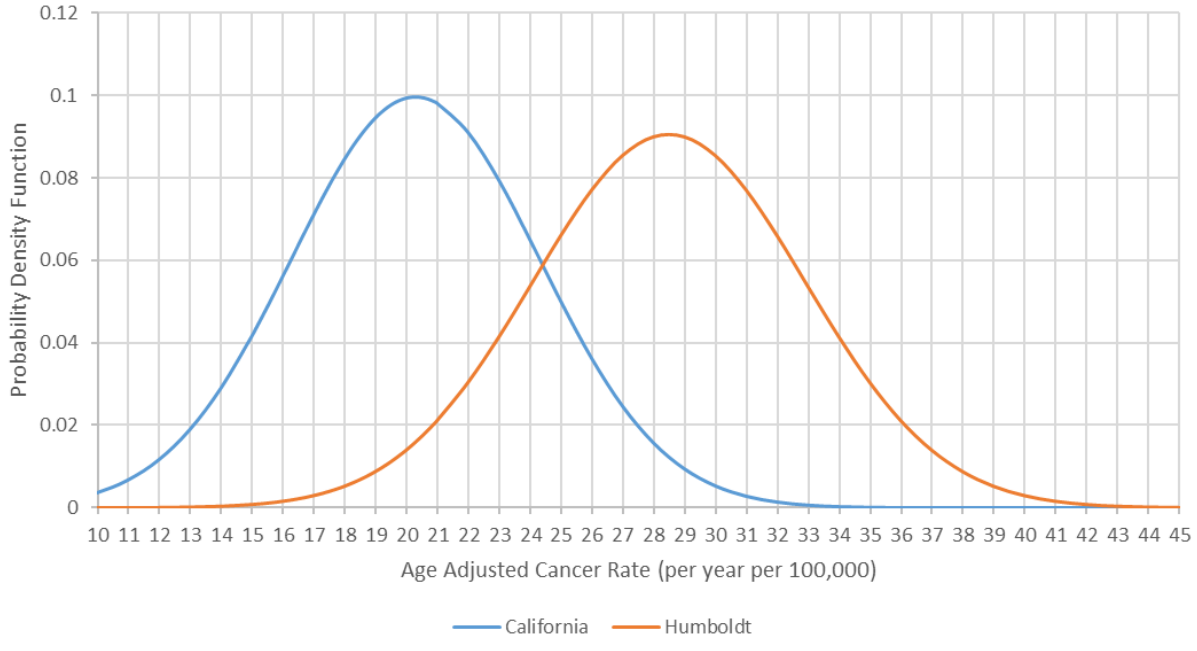


Figure 5. Bladder Cancer Rate Ratio for Humboldt County and California Using Theoretical and Empirical California Confidence Intervals