Big Matters with Small Numbers: Narrow Populations

Richard Charnigo

University of Kentucky

www.richardcharnigo.net/RE/index.html

Motivating Example

The table below summarizes parent report of child asthma status among one state's participants in the National Survey of Children's Health from 2007.

Race/Ethnicity	Asthma	No Asthma
Black Non-Hisp	1	11
Multi Non-Hisp	7	46
Other Non-Hisp	2	28
White Non-Hisp	109	1314
Hispanic	11	184

Motivating Example

Consider estimating the rate of asthma among Black Non-Hisp children.

A reasonable estimate might be 1/12, or 8.3%.

However, we might also wish to perform one or more of the following three tasks:

- Provide a 95% confidence interval for the rate.
- Test a null hypothesis that the rate equals, say, 5%.
- Test a null hypothesis that the rate among Black Non-Hisp children is no different from the rate among White Non-Hisp children.

Motivating Example

Regarding the first task, we might think to apply the familiar formula from an introductory statistical methods course:

$$\widehat{p} \pm 1.96\sqrt{\widehat{p}(1-\widehat{p})/n},$$

where $\widehat{p}$ is the estimated rate and n is the sample size.

Doing so for the present example, we obtain

$$(1/12)\pm 1.96\sqrt{(1/12)(11/12)/12} = 0.083\pm 0.156.$$

Motivating Example

This result is unreasonable. A negative rate is not possible, so a 95% confidence interval for the rate should not include negative numbers.

Generally speaking, the familiar formula provides reasonable results when

$$n\widehat{p}(1-\widehat{p}) \geq 10,$$

although some authors are more generous and suggest 5 rather than 10 as a cutoff.

Motivating Example

However, in this example we have

$$n\widehat{p}(1 - \widehat{p}) = 0.92.$$

Likewise, performing the other two tasks — testing a null hypothesis that the rate equals 5% and testing a null hypothesis that the rate among Black Non-Hisp children is no different from the rate among White Non-Hisp children — using familiar testing procedures from an introductory statistical methods course requires

$$n\widehat{p}(1 - \widehat{p}) \geq 10,$$

which is not the case here.

Motivating Example

Note that the requirement

$$n\widehat{p}(1-\widehat{p}) \geq 10$$

effectively imposes two conditions:

First, $\widehat{p}$ cannot be too small, which is to say that the event cannot be too rare.

Second, n cannot be too small, which is to say that the population cannot be too narrow.

Motivating Example

Handling the first issue (rare events) was the subject of a presentation that I gave two years ago. Materials from that presentation are at {www.richardcharnigo.net/RE/index.html}.

Handling the second issue (narrow populations) is the subject of today's presentation. Before we proceed further, some comments on semantics may be helpful.

Motivating Example

A narrow population may be one for which there are, quite literally, very few individuals. However, more loosely speaking, we may regard a population as narrow if its size in relation to a larger population is small enough that even a good-sized sample for the larger population will translate into a small n for the narrow population. With this understanding, a small n becomes a working definition of a narrow population.

Another way of framing the distinction between the first issue (rare events) and the second issue (narrow populations) is that the former is about small numerators while the latter is about small denominators.

Binomial Approach for Inference on a Rate

Let us first address how one can test a null hypothesis that the rate equals, say, 5%. Knowing how to perform this task will provide a crucial insight on how to create a 95% confidence interval for the rate.

If the null hypothesis were true, then the number of events in a random sample of size 12 would follow a binomial distribution with parameters 12 (number of trials) and 0.05 ("success" probability on each trial). So, our hypothesis test will evaluate whether the observed number of events — in our example, one — represents an extreme occurrence for such a binomial distribution.

Binomial Approach for Inference on a Rate

Please refer to $\{$MCHNarrowPopulations.xls$\}$. Row 51 of Sheet "BinomialRate" indicates that the probability of observing 1 or fewer events in a random sample of size 12 would be 0.882 if the null hypothesis were true, while the probability of observing 1 or more events would be 0.460.

The p-value for the hypothesis test — called an exact test by statisticians because it does not rely on large-sample normal approximations — is defined to be double the minimum of three numbers: the two aforementioned probabilities and 0.5. The doubling handles a two-sided alternative hypothesis, while the 0.5 ensures that the p-value is not greater than 1.

Binomial Approach for Inference on a Rate

In the present example, the p-value is 0.919. We do not reject the null hypothesis.

Intuitively, the probability of observing 1 or fewer events in a random sample of size 12 was not so small to persuade us that the event rate was larger than 5%, while the probability of observing 1 or more events in a random sample of size 12 was not so small to persuade us that the event rate was smaller than 5%.

Binomial Approach for Inference on a Rate

In {MCHNarrowPopulations.xls}, the Sheet "BinomialRate" can be revised to accommodate a different data set. First change the "1" and "12" in the BINOMDIST function used to define the entries of Column C to the observed number of events and the sample size in the different data set.

Then change the "0" and "12" in the BINOMDIST function used to define the entries of Column D to the observed number of events less one and the sample size.

Binomial Approach for Inference on a Rate

Here is a second example, with which we will illustrate revision of Sheet “BinomialRate” in {MCHNarrowPopulations.xls} to accommodate a different data set.

In one state, the National Survey of Children’s Health from 2007 reported that 5 out of 16 children aged 5 or younger with emotional, behavioral, or developmental issues had injuries requiring medical attention, compared to 6 out of 56 children without such issues.

Let us test the null hypothesis that the rate of injuries requiring medical attention among children aged 5 or younger with emotional, behavioral, or developmental issues is 5%.

Binomial Approach for Inference on a Rate

Change the “1” and “12” in the BINOMDIST function used to define the entries of Column C to “5” and ”16”. Then change the “0” and “12” in the BINOMDIST function used to define the entries of Column D to “4” and “16”.

We obtain a p-value of 0.002, and so we reject the null hypothesis. This is because 5 out of 16 is too large to accord with a 5% event rate. As indicated in Column D, there is less than a 0.001 probability of observing 5 or more events in a random sample of size 16 if the event rate is only 5%.

Binomial Approach for Inference on a Rate

To construct a 95% confidence interval, we can use the principle that p_0 should be included in the confidence interval if and only if we would accept the null hypothesis that the rate is p_0. Statisticians refer to this principle as inversion.

See Column F on Sheet “BinomialRate” in {MCHNarrowPopulations.xls} for illustration of the inversion principle. A 95% confidence interval for the rate of asthma among Black Non-Hisp children is 0.003 to 0.384. The confidence interval is wide, reflecting great uncertainty based on the small sample size of 12, but no negative numbers are included.

Hypergeometric Approach to Comparing Rates

Let us now address the task of how to test a null hypothesis that the asthma rate among Black Non-Hisp children is no different from the rate among White Non-Hisp children or, equivalently, the ratio of the rates equals one.

Estimates of these rates are $1/12 = 8.3\%$ and $109/1423 = 7.7\%$ respectively. However, as noted earlier,

$$n\widehat{p}(1-\widehat{p}) = 0.92 < 10$$

for the Black Non-Hisp children. So, the familiar testing procedures from an introductory statistical methods course are not applicable.

Hypergeometric Approach to Comparing Rates

Therefore, we will now describe a testing procedure — called Fisher's exact test — that may be used in this situation.

If the null hypothesis were true, then given random samples of 12 Black Non-Hisp and 1423 White Non-Hisp children with 110 asthma events in total, the number of asthma events among the Black Non-Hisp children should follow a hypergeometric distribution with parameters 12 (number of Black Non-Hisp children), 110 (total number of asthma events), and 1435 (total number of children).

Hypergeometric Approach to Comparing Rates

As shown on Sheet “HypergeometricRates” of {MCHNarrowPopulations.xls}, there would be a 0.383 probability of no asthma events among the Black Non-Hisp children, a 0.384 probability of one event, a 0.175 probability of two events, and so forth.

Fisher’s exact test works by adding up all probabilities less than or equal to the probability associated with the actually observed number of asthma events among Black Non-Hisp children.

Hypergeometric Approach to Comparing Rates

The sum of these probabilities is the p-value for testing the null hypothesis of equal asthma rates.

Since the actually observed number of asthma events among Black Non-Hisp children was 1, and since the associated probability of 0.384 is larger than every other probability, the p-value in this example turns out to be 1. There is no basis for rejection of the null hypothesis. Intuitively, this is because 8.3% and 7.7% are quite close.

Hypergeometric Approach to Comparing Rates

To revise Sheet “HypergeometricRates” in {MCHNarrowPopulations.xls} to accommodate a different data set, change the “12” and “110” and “1435” in the HYPGEOMDIST function used to define the entries of Column C to the smaller sample size, the total number of events in the two samples, and the sum of the two sample sizes in the different data set.

Then change the “.38431” used to define the entries of Column D to that entry of Column C corresponding to the actual number of observed events. Make sure to round up.

Hypergeometric Approach to Comparing Rates

Considering our second example, in which 5 out of 16 children aged 5 or younger with emotional, behavioral, or developmental issues had injuries requiring medical attention, compared to 6 out of 56 children without such issues, let us test the null hypothesis that the rates of injuries requiring medical attention are the same for children with emotional, behavioral, or developmental issues as for children without such issues.

Referring to Sheet “HypergeometricRates” in {MCHNarrowPopulations.xls}, change the “12” and “110” and “1435” to “16” and “11” and “72”. Then change the “.38431” to “.04693”.

Hypergeometric Approach to Comparing Rates

The p-value from Fisher's exact test is 0.059, so we do not reject the null hypothesis.

We come close, however, because the 5 out of 16 is 31.2% compared to 10.7% for the 6 out of 56.

Bayesian Methods

All of the inferential procedures described thus far are frequentist. Roughly speaking, this means we use only the sample at hand to make inferences about the population from which the sample is drawn.

Moreover, parameters describing that population are viewed as fixed, even though they are unknown to the researcher. This is why we make statements like “We are 95% confident that the rate is between 10% and 30%” rather than “There is a 95% probability that the rate is between 10% and 30%”.

Bayesian Methods

In contrast, Bayesian inference uses both the sample at hand and a set of prior beliefs to make inferences about the population from which the sample is drawn.

Moreover, parameters describing that population are viewed as themselves random, and so we can actually make statements like “There is a 95% probability that the rate is between 10% and 30%”.

Bayesian Methods

To provide an intuitive motivation for Bayesian inference, suppose that I flip a coin and observe a heads.

Frequentist inference would then estimate the probability of obtaining a heads with that coin as 100% (1 success in 1 trial).

However, Bayesian inference might estimate the probability of obtaining a heads with that coin as close to 50%, drawing upon prior beliefs in the form of having in the past observed flips come up heads and tails in roughly equal numbers with other coins.

Bayesian Methods

More explicitly, we can quantify prior beliefs by imagining an auxiliary thought experiment in which, say, a coin was flipped 200 times to yield 100 heads.

Combining the prior beliefs (100 successes in 200 trials) with the results from the sample at hand (1 success in 1 trial), we may adopt $101/201 = 50.2\%$ as a single number estimate of the probability of obtaining a heads.

Or, if we are really fervent about a Bayesian interpretation and regard the probability of obtaining a heads as itself random, then 50.2% is a measure of central tendency for the distribution of the probability of obtaining a heads.

Bayesian Methods

Now let us apply Bayesian inference to the task of providing a 95% confidence interval for the asthma rate among Black Non-Hisp children.

The first question we face is how to specify prior beliefs. This can be done in one of several ways. For instance:

- We can use current year's data from other minority race/ethnicity groups in the state.
- We can use previous years' data from Black Non-Hisp children in the state.
- We can use current year's data from Black Non-Hisp children in neighboring states.

Bayesian Methods

All three of the above options have drawbacks:

- The first assumes a perhaps unrealistic homogeneity across minority race/ethnicity groups.

- The second disregards the redundancy between the previous years' data and the current year's data.

- The third assumes a perhaps unrealistic geographic homogeneity. However, we will go with the third option for illustrative purposes.

Bayesian Methods

The current year's data from Black Non-Hisp children in neighboring states provides estimated rates of 0, 0, 0.191, 0.163, 0, and 0.222, as indicated on Sheet "Bayesian" of {MCHNarrowPopulations.xls}.

The corresponding estimated rates for White Non-Hisp children are 0.054, 0.057, 0.056, 0.037, 0.068, and 0.058.

Bayesian Methods

Using what statisticians call the method of moments, which I have automated on Sheet "Bayesian" of {MCHNarrowPopulations.xls}, the neighboring states' estimated rates suggest the following numbers of prior successes and failures: 0.63 and 5.96 for Black Non-Hisp, 29.24 and 502.71 for White Non-Hisp.

Note that statisticians use "success" to refer to the occurrence of some usually adverse event, in this case asthma, which contrasts with how we use "success" in everyday speech.

Bayesian Methods

Note, first, that prior successes and failures do not actually have to be integers and, second, that prior successes and failures are essentially downweighted versions of the actual numbers of successes and failures from the neighboring states.

This downweighting distinguishes the Bayesian approach from a simple aggregation of the actual numbers from the neighboring states, so that the actual numbers from the neighboring states have some impact on our inferences but not as much impact as the actual numbers from the state of interest.

Bayesian Methods

Combining the prior successes and failures with the observed successes and failures, our single number estimates for the asthma rates among Black Non-Hisp and White Non-Hisp children respectively are 8.8% (slightly higher than the frequentist estimate 8.3%) and 7.1% (slightly lower than the frequentist estimate 7.7%).

Given the prior successes and failures along with the observed successes and failures, we postulate that the asthma rate has a beta distribution with first parameter equal to the sum of the prior and observed successes plus one and second parameter equal to the sum of the prior and observed failures plus one. This beta distribution is called a posterior distribution.

Bayesian Methods

Then, as illustrated on Sheet "Bayesian" of {MCHNarrowPopulations.xls}, we randomly draw 1000 realizations from the posterior distribution of the asthma rate for Black Non-Hisp and 1000 realizations from the posterior distribution of the asthma rate for White Non-Hisp. We also form ratios from these random draws.

Note that, since these draws are random, Excel will recalculate them whenever the spreadsheet is manipulated. If one wished to avoid that, one could copy and paste the draws and select "Values Only" in the "Paste Options" menu.

Bayesian Methods

We obtain a 95% confidence interval — actually, Bayesians would call it a 95% credible interval — for the asthma rate among Black Non-Hisp children by taking the 2.5 and 97.5 percentiles of the random draws from the posterior distribution of the asthma rate for Black Non-Hisp. In other words, we take the middle 95% of the random draws to define the 95% credible interval.

The 95% credible interval will be different every time the spreadsheet is manipulated. As I write this script, I see 0.026 and 0.288 as the lower and upper limits. While still quite wide, this 95% credible interval is narrower than the 0.003 to 0.384 obtained earlier using a frequentist approach. Thus, adding information about prior beliefs reduced our uncertainty about the asthma rate for Black Non-Hisp.

Bayesian Methods

We also obtain a 95% credible interval for the asthma rate among White Non-Hisp children (0.060 to 0.082) and for the ratio of asthma rates (0.383 to 4.192).

As for hypothesis testing, from a strict Bayesian perspective the null hypothesis will be false with probability one. This relates to the idea from an introductory statistics course that the probability of a continuous random variable landing exactly on a specific number is zero. Here the specific number is the value proposed for the parameter under the null hypothesis.

Bayesian Methods

That said, if the credible interval does not contain the value proposed for the parameter under the null hypothesis, we can say that the posterior distribution is incompatible with the null hypothesis in the sense that the parameter is unlikely to be close to the value proposed under the null hypothesis.

For example, the asthma rate among White Non-Hisp children is unlikely to be close to 0.05, but we cannot make a similar assertion for the asthma rate among Black Non-Hisp children.

Bayesian Methods

In {MCHNarrowPopulations.xls}, revising Sheet "Bayesian" to accommodate a new data set will entail changing the entries under "Estimates from Adjacent States" to reflect the new prior beliefs as well as the entries by "Observed Successes" and "Observed Failures" to reflect the new data set.

This will yield new entries by "Beta Parameter 1" and "Beta Parameter 2", which should then be input into the BETAINV functions of Columns J and K in lieu of 2.633897, 17.96344, 139.2405, and 1817.706.

Advantages and Disadvantages of Probability Modeling

Strengths of the probability modeling methods based on the Binomial and Hypergeometric distributions are as follows:

- One avoids having to specify a source of prior beliefs, which would entail some element of subjectivity.

- Many people are already familiar with the interpretations of frequentist confidence intervals and hypothesis tests.

- Besides my implementations of these methods in Excel, one may employ built-in implementations of these methods in statistical software packages such as SAS (via PROC FREQ).

Advantages and Disadvantages of Probability Modeling

Weaknesses of the probability modeling methods based on the Binomial and Hypergeometric distributions are as follows:

- By not using information about prior beliefs, one increases the uncertainty associated with one's inferences.

- Fisher's exact test is geared toward the very special null hypothesis that the rate ratio equals one.

- While the interpretations of frequentist confidence intervals are familiar, they are awkward since they implicitly rely on a notion of repeated sampling.

Advantages and Disadvantages of Bayesian Analysis

Strengths of the Bayesian methods are as follows:

- By using information about prior beliefs, one reduces the uncertainty associated with one's inferences.

- One's inference about a rate ratio is not confined to the question of whether the rate ratio equals one.

- The interpretations of Bayesian credible intervals are less awkward than the interpretations of frequentist confidence intervals.

Advantages and Disadvantages of Bayesian Analysis

Weaknesses of the Bayesian methods are as follows:

- One must specify a source of prior beliefs, which entails some element of subjectivity.

- Strictly speaking, a Bayesian perspective does not accommodate testing a null hypothesis against a two-sided alternative. So, one's inferences are based on the credible intervals.

- I am not aware of a statistical software package that implements these Bayesian methods as I have done in Excel without requiring some substantial programming by the user.

Conceptual Differences between Frequentist and Bayesian Inference

The conceptual differences between frequentist and Bayesian inference are roughly summarized as follows:

Frequentist inference views a population parameter as a fixed but unknown number that is estimated using solely the data from a sample drawn from that population.

Both hypothesis testing and confidence intervals are part and parcel of frequentist inference.

Conceptual Differences between Frequentist and Bayesian Inference

Bayesian inference views a population parameter as a random quantity whose distribution is estimated using not only the data from a sample drawn from that population but also prior beliefs, which in practice may involve data from samples drawn from other populations.

Moreover, because a continuous random variable does not exactly equal a specific value with positive probability, under typical circumstances a null hypothesis is false with probability one. As such, Bayesian inference emphasizes credible intervals.

Confidentiality Issues

Some data sets with narrow populations may, despite not containing traditional identifiers such as names or social security numbers, allow for some individuals to be identified by their neighbors or colleagues.

Although this is somewhat unlikely to occur with survey data (because who completed the survey will not generally be known), this may occur if we are speaking of data from a population rather than from a sample.

In other words, confidentiality may be a big concern when we quite literally have a narrow population.

Confidentiality Issues

Because confidentiality may be a big concern, data sets from narrow populations may be summarized rather than presented in their entireties.

For instance, records for individuals may not be made available, but numbers of events in each of various strata may be furnished along with their respective denominators.

In some cases, even reporting the numbers of events may be revelatory. Then the numbers of events may be suppressed as well.

What now?

You are invited to visit the website {www.richardcharnigo.net/RE/index.html}, which contains this presentation as well as the one I gave two years ago on rare events.

You are also invited to review whatever parts of the Ancillary Notes may interest you. They are contained in the same file as this presentation. The Ancillary Notes present some technical details for the binomial and hypergeometric distributions, some further comments about Bayesian inference, and some references for supplementary reading.

Ancillary Notes

Binomial distribution

Let n be a positive integer (number of trials) and p a positive number less than 1 (success probability). A random variable X has the binomial distribution with parameters n and p if

$$P(X = j) = \binom{n}{j} p^j (1-p)^{n-j}$$

for any integer j between 0 and n, where $\binom{n}{j}$ — read "n choose j" — is defined as $n!/[j!(n-j)!]$.

Above, $X = j$ is the event that there are j successes in the n trials.

Ancillary Notes

Hypergeometric distribution

Let n be a positive integer (number of people in smaller of two samples), a be a positive integer (number of events in two samples), and m be a positive integer (number of people in two samples) greater than n and a. A random variable X has the hypergeometric distribution with parameters n, a, and m if

$$P(X = j) = \binom{a}{j}\binom{m-a}{n-j}/\binom{m}{n}$$

for any integer j between max$\{0, a+n-m\}$ and min$\{a, n\}$.

Above, $X = j$ is the event that j out of the a events occur in the smaller of the two samples.

Ancillary Notes

Bayesian inference

Let p denote a success probability. Necessarily p is between 0 and 1. The probability of observing a successes and $n-a$ failures in n trials is

$$\binom{n}{a}p^a(1-p)^{n-a}.$$

This is proportional, in p, to

$$p^a(1-p)^{n-a}.$$

If our prior beliefs are tantamount to b successes and $m-b$ failures in m trials, then the prior beliefs can be expressed as

$$\binom{m}{b}p^b(1-p)^{m-b}.$$

This is proportional, in p, to

$$p^b(1-p)^{m-b}.$$

Note that the last expression is well-defined even if b and $m-b$ are not integers.

Ancillary Notes

To combine the observations with the prior beliefs, we multiply:

$p^a(1-p)^{n-a} \times p^b(1-p)^{m-b} = p^{a+b}(1-p)^{n+m-a-b}$.

Next, we note that

$$C(a,b,n,m)p^{a+b}(1-p)^{n+m-a-b}$$

defines the Beta probability distribution with parameters $a+b+1$ and $n+m-a-b+1$, where $C(a,b,n,m)$ — which depends on a,b,n,m but not p — ensures that the area under the curve from $p=0$ to $p=1$ is one.

Ancillary Notes

Hence, combining the observations with the prior beliefs leads to the postulate that the posterior distribution of p should be Beta with parameters $a + b + 1$ and $n + m - a - b + 1$.

Finally, to specify the prior successes and prior failures using the method of moments, we use the formulas

$$b = \bar{x}[\bar{x}(1 - \bar{x})/s^2 - 1]$$

and

$$m = [\bar{x}(1 - \bar{x})/s^2 - 1],$$

where $\bar{x}$ and s^2 denote the sample mean and variance of the estimated rates on which the prior beliefs rely.

Ancillary Notes

References

A good general reference for statistical methods is *Fundamentals of Biostatistics*, Sixth Edition, by Bernard Rosner (Duxbury, 2006).

Equations 7.44 (hypothesis test) and 6.20 (confidence interval) present the binomial approach for inference about a rate.

Equation 10.11 (Fisher's exact test) presents the hypergeometric approach to comparing rates.

Ancillary Notes

A good general reference for Bayesian inference is *Bayesian Data Analysis* by Gelman, Carlin, Stern, and Rubin (Chapman & Hall/CRC, 1995).

The approach presented herein – called empirical Bayes — is described in Section 5.1 and represents an approximation to the hierarchical Bayesian analysis described in Section 5.3.