A Semiparametric Test for Heterogeneous Risk

Author

Stephen M. Stohs
National Marine Fisheries Service
8604 La Jolla Shores Drive
La Jolla, CA 92037-1508
Stephen.Stohs@noaa.gov

Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Long Beach, California, July 23-26, 2006

A Semiparametric Test for Heterogeneous Risk

Stephen M. Stohs

May 31, 2006

Introduction

Various empirical settings feature repeated exposure to a risk that leads to a binary classification at the observation level. The standard modeling approach treats the outcomes as a series of Bernoulli trials, assigning a random variable equal to 1 for the outcome which is classified as "success" and 0 for the outcome which is classified as "failure." If the observations can be further classified by some natural grouping into clusters of small size, a question arises whether the probability of success is constant across clusters, or if some clusters face a significantly higher risk of success than others, warranting a less restrictive model which accommodates heterogeneous risk.

I consider two cases which fit the above description. The first example was suggested by a mentor in graduate school, who conjectured that the large number of sons among his offspring suggested that nonrandom factors may play a role in gender determination. To investigate this conjecture,

we can begin by thinking of gender determination as a random experiment, analogous to a Bernoulli coin toss, which begins with fertilization of the egg by a gamete (from the father) possessing either an X or a Y chromosome and ends with the birth of either a daughter or a son. Under the homogeneous risk assumption, the chance for a child of either gender is like an independent coin toss with constant and homogeneous probability of female birth across couples. Under the alternative hypothesis of heterogeneous risk, the gender determination process is heterogeneous, with some couples at greater risk of bearing sons, and other couples at greater risk of bearing daughters.

A second example, which potentially has general implications for managing economic production activities that pose risk to endangered species, arises in the context of incidental take of protected species by commercial fishermen. For an example, I consider California's large-mesh drift gillnet commercial fishery for swordfish and thresher shark (California DGN fishery). A drift gillnet fishing trip consists of a number of sets on the range from 1 to 20. Each set involves lowering a net into the water for about twelve hours then hauling it back up to retrieve the catch. The sets are roughly uniform with respect to duration and gear type and may each be regarded as one day's worth of fishing effort. If an endangered leatherback turtle is entangled more than one hour before the end of the set, it is virtually certain to die of suffocation.

Empirical evidence suggests that after controlling for geography and gear type, a reasonable first approximation to the set-level risk of leatherback

turtle take is provided by a Poisson distribution with homogeneous risk. However, the fact that a few trips experience multiple leatherback takes while many others experience none raises the question of whether some trips face significantly higher leatherback take risk than others. The evidence is confounded by variation in the number of sets per trip, raising the possibility that trips with multiple leatherback takes may simply reflect greater risk exposure.

A standard approach to testing the homogeneous risk hypothesis is to use a Chi square test based on the observed and expected number of successes within each cluster. However, the rule of thumb for using a Chi square test suggests the number of expected observations under the null hypothesis within at least most of the observation units should be five or more (Lindgren 1976), which generally will be far from the case if the data features a large number of clusters of small size. The Chi square test is known to produce better results when the sample size is at least four or five times the number of cells, and if the number of clusters is large relative to the cluster size, this is unlikely to be the case.

To test for heterogeneous risk when the data consists of many clusters of small size, I have developed an asymptotically valid semiparametric test of the null hypothesis that risk is homogeneous across clusters. Under the null hypothesis, I estimate the probability of success as the sample proportion of successes in the data. In order to properly account for the influence of cluster size variation on the number of successes, I assume that the number

of successes within a cluster is binomially distributed conditional on cluster size, with the same binomial parameter applying to all clusters. Then I use the empirical distribution of cluster size to compute an expected number of successes for each cluster size present in the data. Pearson's chi square goodness-of-fit test is applied to compare the expected and observed number of successes for each cluster size. A significant chi square statistic leads to a rejection of the homogeneous risk hypothesis in favor of the heterogeneous risk alternative.

My statistical test will be demonstrated on two data sets which are representative of the examples described above. The first is a sample of the number of male and female siblings within a collection of families for students in a large econometrics course at the University of California - San Diego. The homogeneous risk hypothesis is supported if the numbers of children of each gender reflect a collection of binomial distributions mixed by the distribution of family size. Too many families with excessively large proportions of sons or daughters will lead to a rejection in favor of the hypothesis that gender determination risk is heterogeneous across families.

The second data set consists of a historical collection of trip-level data for a geographically limited portion of the California drift gillnet fishery over the period from 1990-2005, which includes information on fishing effort (the number of sets[1]) and the number of incidental leatherback takes on each

[1]A drift gillnet set consists of lowering the net into the water and soaking it over a period of about 10 hours, then hauling the catch on board the fishing vessel.

trip. The risk of leatherback take on any individual set is very small, and is closely approximated by a Poisson random variable with homogeneous take risk at the set level. However, some trips resulted in more than one set with leatherback takes. Application of my test addresses the question of whether individual DGN fishing trips are subject to significantly heterogeneous take risk. The answer to this question has implications for controlling leatherback bycatch risk.

Description of the Probability Model and Semiparametric Test

The objective is to apply a version of Pearson's Chi square test to the question of whether the cluster-level risk is homogeneous. The challenges are that the average cluster size is small, resulting in a typical expected number of successes per cluster smaller than five, and that cluster size varies randomly according to an unknown distribution. The approach I chose was to develop a test statistic which considers the random variation in cluster size without requiring knowledge of its unknown distribution. The resulting test is semiparametric in that the conditional distribution of successes within each cluster is parametric, while the distribution of the number of Bernoulli trials across clusters is not.

Let N denote the random variable for cluster size, which assumes positive integer values over k observation observation units[2]. The distribution

[2]In my applications, the observation units are either families or DGN fishing trips.

of N is described by a cumulative distribution function $F(\cdot)$ with unknown properties. Assume that $\{N_i, \mathrm{i} = 1,\, 2,\, \ldots, \mathrm{k}\}$ denotes a sample of k observations on N. Further let X_{it} denote a Bernoulli (0/1) random variable with probability of success given by θ_i, for $i = 1, 2, \ldots, k$ and $t = 1, 2, \ldots, N_i$. We are interested in testing the null hypothesis that $\theta_i = \theta$ for all values of $i = 1, 2, \ldots, k$, where θ is a constant. For the i^{th} cluster, define the random variable

$$Y_i = \sum_{t=1}^{N_i} X_{it}, \tag{1}$$

which is the sum of N_i Bernoulli trials in the cluster, i.e., the number of successes for that cluster. Assuming the Bernoulli random variables which comprise the i^{th} cluster are exchangeable[3], we can model the conditional distribution of $Y_i \,|\, N_i$ as $Binom(N_i, \theta_i)$ in the unrestricted model and as $Binom(N_i, \theta)$ in the restricted case.

Suppose the econometrician has a sample consisting of observations on the number of successes y_i over n_i Bernoulli trials in each of k clusters, $\{(y_i, n_i),\, i = 1, 2, \ldots k\}$. Under the null hypothesis of homogeneous risk, the exact conditional distribution for the number of successes Y_i in the i^{th} cluster is a binomial distribution conditional on the number n_i of Bernoulli trials in the i^{th} cluster:

$$p(y \,|\, n_i) = \binom{n_i}{y} \theta^y (1-\theta)^{n_i - y}, \qquad i = 1, 2, \ldots, k. \tag{2}$$

[3]The Bernoulli random variables are considered to be exchangeable if their joint distribution within a cluster is not altered by permuting the labels $i = 1, 2, \ldots, N_i$.

Define y_{max} as the largest value of y which corresponds to one of the cells in the Chi square classification[4]. An expected number of observations for each value of $y = 1, 2, 3, \ldots, (y_{max} - 1)$ may be computed using the observed number of Bernoulli trials in each cluster[5] in conjunction with the conditional distribution of $y|n_i$:

$$\begin{aligned}
\hat{e}_y &= E(\text{number of clusters with } y \text{ successes}) \\
&= k\hat{p}_k(y) \\
&= k\sum_{n=1}^{\infty} \hat{p}(y \mid n)(\hat{F}_k(n) - \hat{F}_k(n-1)) \\
&= \sum_{i=1}^{k} \hat{p}_k(y \mid n_i), \qquad (3)
\end{aligned}$$

where the empirical conditional probability mass function (p.m.f.) for Y given n_i,

$$\hat{p}(y \mid n_i) = \binom{n_i}{y} \hat{\theta}^y (1 - \hat{\theta})^{n_i - y}. \qquad (4)$$

This is the estimate of the conditional distribution of Y_i given n_i based on the minimum Chi square estimate[6], $\hat{\theta}$, of θ.

The probability for the cell corresponding to y_{max} is computed as the sum

[4]The value of y_{max} should be chosen if possible to ensure a sufficiently high expected number of observations in each cell of the classification scheme.

[5]The equivalence of the following series of equations is demonstrated in Appendix B.

[6]The procedure for computing the minimum Chi square estimate is explained in Appendix A. An alternative approach is to use the maximum likelihood estimate (MLE) of the Bernoulli parameter, $\tilde{\theta} = \frac{\sum_{i=1}^{k} y_i}{\sum_{i=1}^{k} n_i}$. However, as noted in Hogg and Craig (Hogg & Craig 1978) and other sources, the Chi square statistic based on MLE parameter estimates results in an upward bias to the computed p-value.

of probabilities over the upper tail of the distribution beginning with y_{max}, which is most easily calculated using the complementary probability:

$$Pr\{Y \geq y_{max}\} = 1 - \sum_{y=0}^{y_{max}-1} \hat{p}_k(y). \tag{5}$$

The expected cell frequencies are calculated as

$$\hat{e}_y = \sum_{i=1}^{k} \hat{p}_k(y \,|\, n_i) \tag{6}$$

for $y = 0, 1, 2, \ldots, (y_{max} - 1)$ and

$$\hat{e}_{y_{max}} = kPr\{Y \geq y_{max}\} = k - \sum_{y=0}^{y_{max}-1} \hat{e}_y. \tag{7}$$

The observed number of clusters with $Y_i = y$ is counted using

$$o_y = \sum_{i=1}^{k} 1\{y_i = y\}, \tag{8}$$

for $y = 0, 1, 2, \ldots, (y_{max} - 1)$, where $1\{\cdots\}$ is the indicator function equal to 1 if the given condition is true and 0 otherwise, and

$$o_{y_{max}} = k - \sum_{y=0}^{y_{max}-1} o_y. \tag{9}$$

These two calculations give rise to vectors of observed and expected number of observations on each possible number of successes $0 \leq y \leq y_{max}$ over

all clusters in the data, $\hat{e} = [\hat{e}_0\, \hat{e}_1\, \ldots\, \hat{e}_{y_{max}}]'$ and $o = [o_1\, o_2\, \ldots\, o_{y_{max}}]'$.

The null hypothesis of homogeneous risk may be tested using a Chi square test based on comparing the observation vector o to the vector of the expected number of observations e:

$$\hat{\chi}^2_k = \sum_{y=0}^{y_{max}} \frac{(o_y - \hat{e}_y)^2}{\hat{e}_y}, \tag{10}$$

which has an asymptotic χ^2 distribution with $y_{max} - 1$ degrees of freedom, as one degree of freedom is sacrificed for the constraint that the total number of expected observations must sum to k, and a second degree of freedom is lost in estimating θ by the minimum chi square estimate subject to the homogeneity restriction ($\theta_i = \theta$ for $i = 1, 2, \ldots, k$).

Human Gender Determination

An introductory discussion of the role of probability in genetics and gender determination is provided in Feller (Feller 1968). Human gender is determined by a pair of chromosomes. In all individuals except for a negligible share of exceptions, females have an XX pair, while males have an XY pair. At first glance, the question of human gender determination seems straightforward: In reproduction, the female contributes an X-chromosome, while the male contributes either an X- or a Y-chromosome through a selection process whose outcome is commonly likened to that of a fair coin toss. Fer-

tilization of the egg in the former case ultimately results in the birth of a female child and in the latter case results in the birth of a male child.

Whether this simple symmetric model of human gender determination is accurate is an empirical question. There are a number of reasons this model might fall short of actual experience:

1. Nature may favor one gender over another, increasing the relative chance that, say, a sperm with an X chromosome is the one which successfully fertilizes the egg, or that female fetuses will survive gestation;

2. Genetic variation in the parent population could result in significant differences in the relative probability of male or female birth across couples;

3. Variations in mating behavior could account for differences in the probability of male births across couples;

4. Some societies have a gender preference which could induce statistical dependence between family size and gender distribution of offspring. For instance, in some Asian and Latin American cultures, male children are preferred to females. If a couple keeps trying until at least one male child is born, then if we assume a fixed chance of female birth on each attempt, we would expect to see an increasing percentage of females with respect to family size[7]

[7]This form of discretionary censoring could result in selectivity bias in the observed

No. of Children	Count	Percentage
1	29	12.9%
2	121	54.0%
3	51	22.8%
4	13	5.8%
5	6	2.7%
6	2	0.9%
7	2	0.9%

Table 1: Family size distribution

Hence it is entirely possible to find significant variation across couples in the odds of giving birth to daughters rather than sons.

Statistical Analysis

The test is applied below to a combined sample of students from two sections of Econometrics C, a required course in the undergraduate economics curriculum at the University of California, San Diego. The table shows the observed numbers of successes (female births) over the distribution of family size, assuming accurate reporting by the students who were polled[8].

The distribution of family sizes is displayed in the following table:

The table below shows the observed and expected numbers of female children based on the estimated marginal distribution of the number of female children given the empirical distribution of family size, with the expected

gender distribution, even if the fertilization process itself was equally likely to produce a male or a female embryo.

[8]Identical twins were treated as a single roll of the genetic dice, while fraternal twins were treated as separate rolls.

		Expected	
No. of Girls	Observed	MLE	Min χ^2
0	59	56.74	53.58
1	105	99.14	98.81
2	41	53.50	55.85
3	12	11.48	12.29
4	4	2.43	2.66
5 or more	3	0.5890	0.6659

Table 2: Observed and expected numbers of girls

	$(o-e)^2/e$	
No. of Girls	MLE	Min χ^2
0	0.0903	0.4117
1	0.3464	0.3767
2	2.9193	3.7086
3	0.0236	0.0010
4	1.0062	0.7419
5 or more	7.3758	6.2158
χ^2	11.7615	11.4558
p-value	0.0192	0.0219

Table 3: Chi square statistics and p-values

numbers of girls computed using both the MLE and the minimum Chi square estimates of the Bernoulli parameter, θ:

The intermediate steps in calculating the test statistic are displayed in the table below.

Whether the MLE or the minimum Chi square estimate of the Bernoulli parameter is used to develop the expected number of female children, the re-

sulting p-value for the calculated value of the test statistic is between 1% and 5%, so a classical hypothesis test of the null hypothesis that the probability of having daughters is a like homogeneous Bernoulli coin toss would reject the null hypothesis in favor of the alternative hypothesis that the probability of having daughters is heterogeneous across families (and in particular with respect to family size) at the 5% but not the 1% significance level.

A quick look at the data for the observed and expected number of daughters indicates that the observed number of daughters was less than expected for families with only two children, and greater than expected for families with five or more children. This is consistent with a behavioral model where at least a significant share of couples have a target level for the number (or percentage) of sons in their brood; if the threshold number of sons is reached after having only two children, such couples stop. If such a couple has a disproportionate number of daughters by the time they have four children, they try for one more son.

Should Fewer Cells be Used?

The results presented above show that the uppermost cell in the table of expected frequencies does not satisfy the standard rule-of-thumb that the smallest cell frequency should not be less than five. However, attempting to remedy this situation through reducing the uppermost value of y in the partition is problematic. The following table shows that the p-value increases monotonically as the cutoff value for the uppermost cell is decreased:

		MLE		Min χ^2	
Cell Limit	Degrees of Freedom	χ^2	p-value	χ^2	p-value
2	1	1.4052	0.23586	0.57784	0.44716
3	2	4.6642	0.09709	4.5089	0.10493
4	3	8.1032	0.043927	8.1012	0.043966
5	4	11.761	0.019216	11.456	0.021892

Table 4: Cell limits and p-values

The difficulty is that the test statistic only possesses an asymptotic Chi square distribution under the null hypothesis; the distribution under any particular form of the alternative is generally unknown, but is only potentially detectible in the deviations of observed counts from their expected counts at the individual cell level. Aggregating cells at the top end of the table implicitly reduces the set of alternatives which can be detected by the chi square test by summing residuals across any cells which are pooled in order to increase the expected cell count.

For example, suppose that under the homogeneity assumption, the observed count of families with five or more daughters is significantly above its expected count, while the observed count of families with only two daughters is far below the expected count. This *prema facie* evidence of a departure from the homogeneity hypothesis would be obscured if the top categories were combined into a single category for "two or more daughters."

Testing Homogeneity of Leatherback Bycatch Risk over Drift Gillnet Fishing Trips

In this section, I consider statistical questions which arise in connection with managing leatherback turtle bycatch in the California Drift Gillnet (DGN) Fishery. The fishery is described in detail in the Pacific Fishery Management Council's Fishery Management Plan for Highly Migratory Species (Council 2003); a summary is provided below.

Description of the California DGN Fishery

The California DGN fishery traces its origin to the late 1970s when incidental catches of pelagic sharks in a Southern California coastal set net fishery motivated a group of 15 fishing vessel owners to experiment with large-mesh nets targeting thresher shark. Subsequently, California's swordfish industry transformed from primarily a harpoon fishery to a DGN fishery in the late 1970s, and landings soared to a historical high of 286 metric tons (mt) by 1984. After 1981, swordfish became the primary target species for the fleet, because it commands a higher price-per-pound than thresher shark, resulting in a decline in reported thresher shark landings to lows of the late 1980s and early 1990s. The number of DGN vessels landing swordfish declined from 228 in 1985 to 43 in 2004.

Historically, the California DGN fleet has operated within US Exclusive Economic Zone (EEZ) waters adjacent to the state to about 150 mi offshore,

ranging from the U.S.-Mexico border in the south and as far north as the Columbia River during El Niño years. The majority of the current DGN fishing effort is concentrated in the southern California bight due in part to a leatherback turtle time/area closure north of Pt. Conception. Fishing activity is highly dependent on seasonal oceanographic conditions that create temperature fronts that concentrate feed for swordfish. The DGN fishery typically begins in late May and continues through the end of January, although 90 percent of the fishing effort typically occurs from mid-August to the end of December.

Drift gillnet fishing requires specialized inputs, including a crew of 2-3 (including the captain) and appropriate gear which includes a gillnet and a boat (30-85 feet long, with 60 percent of the vessels less than 50 ft in total length) outfitted to transport the fishermen to access the fishing grounds, to permit setting and retrieval of the gillnet, and to facilitate storage of the fish until landing them.

A typical drift gillnet fishing trip consists of between 5 and 15 "sets" of the net, with about 6 sets on average. Nets are typically set in the evening, allowed to soak overnight, and then retrieved in the morning. The average soak time is around 10 hours. The vessel remains attached to one end of the net during the soak period, drifting with the net. During retrieval, the net is pulled over the stern by a hydraulic net reel. As the net is pulled, anything caught in the net can usually be seen coming to the surface, at which point the reel is slowed and stopped if the catch is too large. The catch is either

pulled aboard in the net, or if too large, tied with a line, so as not to be lost, and winched aboard. Once onboard, entangled fish are removed from the net using routine procedures.

Net length ranges from 4,500 ft to 6,000 ft and averages 5,760 ft while net depth ranges from 145 ft to 165 ft and averages 150 ft. The top of the net is attached to a float line and the bottom to a weighted lead line. Although termed "gillnets," the nets actually catch fish by entanglement, rather than literally trapping them by the gills. Nets are also size selective; large fish such as swordfish become entangled while smaller fish pass through the mesh. Unfortunately, the mesh is not sufficiently large to permit the passage of large charismatic megafauna such as leatherback turtles, which occasionally become entangled. As the nets entrap animals at a submersion depth of 36 feet or greater and are only hauled up after soaking for a period of 12 hours or so, there is a high probability that oxygen-breathing animals which become entangled (such as sea turtles and marine mammals) will drown before the net is hauled up.

The Endangered Species Act and Marine Mammal Protection Act are federal laws which impose strict limits on the allowable level of protected species bycatch. Leatherback turtles are granted protection under the Endangered Species Act, and hence a key concern in commercial fisheries management is to limit the risk of accidently capturing or killing them. Since 1990, observers have been sent out on drift gillnet fishing boats to closely monitor the number of leatherback turtles and other species which are caught as by-

catch. Over the period from 1990 through 2005, a total of 23 leatherback turtles were captured as bycatch over the course of approximately 7000 observed[9] drift gillnet fishing trips. As 21 of these trips occurred in the portion of the drift gillnet fishing grounds to the north of Pt. Conception, the risk of leatherback bycatch was deemed excessive and this portion of the fishing grounds was closed to fishing from 2001-2005 over the peak fishing period (August 15-November 15). This "turtle conservation area" will be reopened to a limited amount of fishing effort for the 2005-2006 fishing season, but fishing trips will be subject to 100% observer coverage, and fishing effort will be immediately halted at any point in the season if two leatherback turtles are caught as bycatch.

Questions of interest in connection with leatherback bycatch include the following:

1. Is the risk of leatherback bycatch heterogeneous across fishing seasons?

2. Is the risk of leatherback bycatch geographically heterogeneous – that is, are some areas of higher bycatch risk than others?

3. Is the risk of leatherback bycatch heterogeneous across trips?

Given the availability of close to 7,000 observed drift gillnet sets from 1990-2005, including geographic markers for the approximate location where fishing occurred, the first two of these questions are amenable to a standard

[9]Observed trips involve the literal inclusion of an observer as a passenger on shipboard during a fishing trip; the observer keeps a running tally of the catch and bycatch of the different species which are caught in each drift gillnet set.

Chi square testing approach. However, in the case of the third question, the fact that the average drift gillnet trip only consists of about six sets of fishing, and that a total of only 23 leatherback takes have occurred over the period from 1990-2005, imply that a standard Chi-square testing approach will suffer the shortcoming of an excessive number of cells with an observed count of 0 "successes[10]."

Statistical Analysis

A significant difference in DGN leatherback bycatch risk has been demonstrated to exist between the areas north and south of Pt. Conception[11] (Carretta, Price, Petersen & Read 2004). For purposes of this paper, the focus is limited to the area north of Pt. Conception where bycatch risk is relatively higher. The data were extracted from the California Drift Gillnet Observer Database, and are a representative sample of approximately 20% of the fishing effort which took place for the portion of California DGN fishery North of Pt. Conception over the period from 1990-2004. The distribution of the number of sets per trip is given in Table 5 below.

Based on the empirical conditional distribution of y given n, the observed and expected numbers of trips with 0, 1, or 2 or more leatherback incidental

[10]Environmentalists might prefer to classify leatherback bycatch as "failures" on semantic grounds.

[11]Pt. Conception lies at $37°27'$ North Latitude, and represents a dividing line between the geographically and ecologically distinct southern and northern ranges of the DGN fishery.

takes are shown in the subsequent table[12]. The expected number of takes in the bottom row of the table remains somewhat unsatisfactory in light of the rule-of-thumb stipulation that this value should exceed five, but this problem is somewhat intrinsic to a context with a very small "success" rate[13]. The Chi square statistics for the two cases, and the associated p-values (based on a Chi-square distribution with 1 degree of freedom[14]), are displayed in Table 7.

Test results from using the MLE and from using the minimum Chi-square estimate of θ are comparable, with a slightly larger p-value in the latter case reflecting that the Chi-square statistic was minimized over θ. In both cases, the results for testing the null hypothesis of homogeneous leatherback take risk across DGN fishing trips is inconclusive: A classical hypothesis testing approach would suggest the null hypothesis could be accepted at the 1% significance level, but would be rejected at the 5% significance level. A glance at the computed values of $(o-e)^2/e$ for each different level of number of takes per trip suggests that the two trips with two leatherback takes were significantly higher than expected under the homogeneity hypothesis, suggesting that risk is sometimes unusually high. On the other hand, only two trips out of 490 with two leatherback takes suggests that even if risk is heterogeneous, the occasions when risk is inordinately high are rare.

[12]The expected number of trips was computed two ways: first using the MLE of the Bernoulli parameter ($\tilde{\theta}$), then using the minimum Chi square estimate ($\hat{\theta}$)

[13]The difficulty could potentially be remedied with a sufficiently large data set.

[14]There are three cells in the contingency table, but 1 degree of freedom is lost due to the summing-up condition, and a second is lost due to estimation of the Bernoulli parameter.

Sets per Trip	Number of Trips	Percentage
1	39	8.0%
2	37	7.6%
3	30	6.1%
4	32	6.5%
5	112	22.9%
6	59	12.0%
7	50	10.2%
8	40	8.2%
9	33	6.7%
10	19	3.9%
11	17	3.5%
12	8	1.6%
13	6	1.2%
14	3	0.6%
15	3	0.6%
16	1	0.2%
17	0	0.0%
18	0	0.0%
19	1	0.2%

Table 5: Distribution of number of sets per trip

		Expected	
Takes	Observed	MLE	Min χ^2
0	471	469.49	466.06
1	17	20.03	23.29
2 or more	2	0.48	0.65

Table 6: Observed and expected numbers of trips with 0-2 leatherback takes

	$(o-e)^2/e$	
Bycatch	MLE	Min χ^2
0	0.004876	0.05242
1	0.45955	1.6982
2 or more	4.8331	2.7725
χ^2	5.2976	4.5231
p-value	0.021	0.033

Table 7: Chi square statistics and p-values

Conclusion

This paper has developed and illustrated the use of an asymptotically valid Chi square test for heterogeneous risk which is particularly suited in situations where the risk of success is small on any individual Bernoulli trial, and where the data are naturally aggregated into a large number of clusters of small individual size. The test is semiparametric in that it makes no parametric assumptions about the distribution of cluster size, but assumes the number of "successful" outcomes within each cluster follows a (parametric) binomial distribution conditional on cluster size. Under the null hypothesis of homogeneous risk, a Chi-square test is developed based on the empirical marginal distribution of the number of successes over clusters, and it is shown in Appendix 2 that the resulting test statistic retains the asymptotic properties of the standard Chi-square goodness-of-fit test statistic.

The test was demonstrated using two datasets. The first provides evidence for the heterogeneity of gender determination across different families. The second provides evidence on the heterogeneity of leatherback turtle by-

catch risk over drift gillnet commercial fishing trips over a geographically constrained region. The empirical results in both cases present weak evidence that the risk is heterogeneous across clusters, with a rejection of the homogeneity hypothesis at the 5% level but not the 1% level.

One question for future research is that of what to do if the homogeneity hypothesis is rejected in favor of the alternative hypothesis that risk is heterogeneous across clusters. Possible approaches include using a negative binomial model to capture the overdispersion in the data, or to use a hierarchical approach which models θ_i as a random parameter across clusters, thereby explicitly modeling the heterogeneous risk.

A second question is that of how best to use the disaggregate data at the sample unit level to conduct a homogeneity test of high statistical power. The example for the case of aggregating the top cell of the chi square statistic for the gender test illustrates the loss of statistical power which can result from aggregation. By aggregating the raw data up to the empirical marginal distribution of y, the method presented here helps to solve the problem of insufficiently small expected cell counts at the potential cost of masking heterogeneity across sample units which is not reflected in the marginal distribution of y. A method which directly quantifies the variance in cluster-level residuals (such as a conditional moment test) might potentially increase statistical power by avoiding the cancelation effects which naturally occur with aggregation.

References

Andrews, D. W. K. (1988), 'Chi-square diagnostic tests for econometric models', *Journal of Econometrics* **37**(1), 135–156.

Cameron, A. C. & Trivedi, P. K. (1998), *Regression analysis of count data*, first edn, Cambridge University Press.

Carretta, J. V., Price, T., Petersen, D. & Read, R. (2004), 'Estimates of marine mammal, sea turtle, and seabird mortality in the california drift gillnet fishery for swordfish and thresher shark, 1996-2002', *Marine Fisheries Review* **66**(2), 21–30.

Casillas-Olvera, G. & Bessler, D. A. (2006), 'Probability forecasting and central bank accountability', *Journal of Policy Modeling* **28**, 223–234.

Cochran, W. G. (1952), 'The chi-square test of goodness of fit', *The Annals of Mathematical Statistics* **23**(3), 315–345.

Council, P. F. M. (2003), Fishery management plan and environmental impact statement for u.s. west coast fisheries for highly migratory species, Technical report.

Cramér, H. (1946), *Mathematical Models of Statistics*, Princeton University Press.

Feller, W. (1968), *An Introduction to Probability Theory and its Applications*, Vol. 1, third edn, Wiley.

Hogg, R. V. & Craig, A. T. (1978), *An Introduction to Mathematical Statistics*, fourth edn, Macmillan.

Lindgren, B. W. (1976), *Statistical Theory*, third edn, Macmillan.

Tate, M. W. & Hyer, L. A. (1973), 'Inaccuracy of the chi-square test of goodness of fit when expected frequencies are small', *Journal of the American Statistical Association* **68**(344), 836–841.

Yates, J. F. (1982), 'External correspondence: Decompositions of the mean probability score', *Organizational Behavior and Human Performance* **30**(1), 132–156.

Minimum Chi Square Estimation of the Homogeneous Bernoulli Parameter

An initial estimate for the Bernoulli parameter θ may be obtained from the method of maximum likelihood.

The likelihood for a sample of observations $(y_i\,,\,n_i)$ is given by

$$\begin{aligned} L &= \prod_{i=1}^{k} f(y_i, n_i; \theta_i) \\ &= \prod_{i=1}^{k} p(y_i \,|\, n_i; \theta_i) f(n_i) \\ &= \prod_{i=1}^{k} \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{n_i - y_i} f(n_i), \end{aligned} \tag{11}$$

where $f(y_i, n_i; \theta_i)$ is the joint p.m.f. of the observation (y_i, n_i), $f(n_i)$ is the (unknown) marginal p.m.f. for N and $p(y_i \mid n_i; \theta_i) = \binom{n_i}{y_i}\theta_i^{y_i}(1-\theta_i)^{n_i-y_i}$ is the conditional likelihood of the observation y_i given n_i.

Noting that the likelihood is unique up to a factor which does not depend upon the parameter(s) of interest, it may be more simply written as

$$L = \prod_{i=1}^{k} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i} \tag{12}$$

which has a corresponding log likelihood

$$\log L = \sum_{i=1}^{k} y_i \log(\theta_i) + (n_i - y_i)\log(1-\theta_i). \tag{13}$$

In the unrestricted case, first-order conditions for maximizing the likelihood are given by

$$\frac{\partial \log L}{\partial \theta_i} = \frac{y_i}{\theta_i} - \frac{n_i - y_i}{1-\theta_i} = 0, \tag{14}$$

which leads to the unrestricted maximum likelihood estimates

$$\tilde{\theta}_i = \frac{y_i}{n_i} \tag{15}$$

for $i = 1, 2, \ldots, k$.

Under the homogeneity restriction ($\theta_i = \theta$ for $i = 1, 2, \ldots, k$), the log

likelihood reduces to

$$\begin{aligned} \log L &= \sum_{i=1}^{k} y_i \log(\theta) + (n_i - y_i)\log(1-\theta) \\ &= (\sum_{i=1}^{k} y_i)\log(\theta) + (\sum_{i=1}^{k} n_i - \sum_{i=1}^{k} y_i)\log(1-\theta), \end{aligned} \tag{16}$$

and the first order condition for maximizing the likelihood is given by

$$\frac{d\log L}{d\theta} = \frac{\sum_{i=1}^{k} y_i}{\theta} - \frac{\sum_{i=1}^{k} n_i - \sum_{i=1}^{k} y_i}{1-\theta} = 0, \tag{17}$$

whose solution provides the maximum likelihood estimate under the homogeneity restriction of

$$\tilde{\theta} = \frac{\sum_{i=1}^{k} y_i}{\sum_{i=1}^{k} n_i}. \tag{18}$$

The test statistic presented in this paper was computed using both the MLE and the minimum Chi square estimates of the Bernoulli parameter θ under the homogeneity restriction. The minimum Chi square estimate of θ was obtained by taking the MLE as an initial value, then using matlab's fminsearch function to iterate to the value $\hat{\theta}$ which minimized the Chi square statistic based on the observed data sample.

Verification of the Formula for Expected Number of Observations

The main body of the paper asserted that the expected number of observations for each value of $y = 1, 2, 3, \ldots, \infty$ could be computed using $e_y = \sum_{i=1}^{k} \hat{p}_k(y \mid n_i)$, and this is proved as follows:

$$
\begin{aligned}
e_y &= k\hat{p}_k(y) \\
&= k \sum_{n=1}^{\infty} \hat{p}(y \mid n)[\hat{F}_k(n) - \hat{F}_k(n-1)] \\
&= \sum_{n=1}^{\infty} \hat{p}(y \mid n)[k(\hat{F}_k(n) - \hat{F}_k(n-1))] \\
&= \sum_{n=1}^{\infty} \hat{p}(y \mid n)(\sum_{i=1}^{k} 1\{n_i \leq n\} - \sum_{i=1}^{k} 1\{n_i \leq n-1\}) \\
&= \sum_{n=1}^{\infty} \hat{p}(y \mid n) \sum_{i=1}^{k} (1\{n_i \leq n\} - 1\{n_i \leq n-1\}) \\
&= \sum_{n=1}^{\infty} \hat{p}(y \mid n) \sum_{i=1}^{k} 1\{n_i = n\} \\
&= \sum_{i=1}^{k} \hat{p}(y \mid n) \sum_{n=1}^{\infty} 1\{n_i = n\} \\
&= \sum_{i=1}^{k} \hat{p}(y \mid n), \qquad (19)
\end{aligned}
$$

since $\sum_{n=1}^{\infty} 1\{n_i = n\} = 1$ for each $i = 1, 2, \ldots, k$.

Asymptotic Convergence of the Test Statistic

The semiparametric test described in this paper assumes a nonparametric model for the distribution of cluster size, represented by the random variable N, and a parametric (binomial) distribution of the number of successes, Y, conditional on cluster size, with conditional p.m.f.

$$p(y \mid N) = \binom{N}{y} \theta^y (1-\theta)^{N-y}. \tag{20}$$

The test statistic is a Chi-square statistic based on the distribution of counts from the empirical marginal distribution of y, whose theoretical counterpart is given by

$$\begin{aligned} p(y) &= \sum_{n=1}^{\infty} p(y \mid n) f(n) \\ &= \sum_{n=1}^{\infty} p(y \mid n)(F(n) - F(n-1)), \end{aligned} \tag{21}$$

where $F(n) \equiv 0$ by definition[15].

The empirical distribution function for N may be expressed as

$$\hat{F}_k(n) = \frac{\sum_{i=1}^{k} 1\{n_i \leq n\}}{k}, \tag{22}$$

[15]Although there may be real-world reasons that $F(0) = Pr\{N = 0\} > 0$ in particular settings of interest, we implicitly rule this out by conditioning on $N > 0$, thereby ignoring clusters with no data to be observed.

where $1\{\cdot\}$ denotes the indicator function, equal to 1 if the parenthesized condition is true and 0 otherwise, and k is the number of clusters. The corresponding empirical probability mass function (or probability histogram) is given by[16]

$$\hat{f}_k(n) = \hat{F}_k(n) - \hat{F}_k(n-1). \tag{23}$$

The empirical marginal distribution of y is given by

$$\hat{p}_k(y) = \sum_{n=1}^{\infty} \hat{p}(y \mid n)[\hat{F}_k(n) - \hat{F}_k(n-1)]. \tag{24}$$

A well-known result in probability theory shows that

$$plim\,\{F_k(n)\} = F(n), \tag{25}$$

that is, the empirical distribution function asymptotically approaches the cumulative distribution function of the underlying population distribution. Further, under suitable regularity assumptions the minimum Chi square estimator $\hat{\theta}$ is a consistent estimator of the population Bernoulli parameter θ. Noting that the expression for $\hat{p}_k(y)$ is a continuous function of $\hat{F}_k(n)$, $\hat{F}_k(n-1)$, and the minimum Chi square estimator $\hat{\theta}$, it follows from the Continuous Mapping Theorem that

$$plim\,\{\hat{p}_k(y)\} = p(y). \tag{26}$$

[16]For the purpose of the following definition, we take $\hat{F}_k(0) \equiv 0$.

Next denote by

$$\chi_k^2 = \sum_{y=0}^{y_{max}} \frac{(o_y - e_y)^2}{e_y}, \tag{27}$$

the test statistic based on a sample consisting of k clusters, but based on the (unknown) marginal p.m.f. for Y, $p(y)$ instead of the estimated marginal p.m.f. $\hat{p}_k(y)$. We accept on the basis of earlier proof (Cramér 1946) that χ_k^2 converges in distribution to a Chi square random variable with $y_{max} - 1$ degrees of freedom, then note that χ_k^2 is a continuous function of e_y for each value of $y = 0, 1, 2, \ldots, y_{max}$, provided $e_y \neq 0$. Since

$$\hat{\chi}_k^2 = \sum_{y=0}^{y_{max}} \frac{(o_y - \hat{e}_y)^2}{\hat{e}_y}, \tag{28}$$

is formally equivalent to χ_k^2 except that $\hat{e}_y$ replaces e_y in each term, and $plim\,\{\hat{p}_k(y)\} = p(y)$ implies that $\frac{(o_y - \hat{e}_y)^2}{\hat{e}_y}$ converges in probability to $\frac{(o_y - e_y)^2}{e_y}$. This fact and some additional analysis[17] may be used to write

$$\hat{\chi}_k^2 = \chi_k^2 + R_k, \tag{29}$$

where $plim R_k = 0$. It follows from Slutsky's theorem that $\hat{\chi}_k^2$ has the same limiting distribution as that of χ_k^2.

[17] One can use a Taylor expansion of $f(e) = \frac{(o-e)^2}{e}$ about the points e_y, then sum the deviations from $\frac{(o_y - e_y)^2}{e_y}$ to obtain the residual difference R_k which converges to 0 in probability.