Ecological Inference

Zach Siegel
UCLA Anderson School of Management
zachary.edmund.siegel@gmail.com

I am **willing** to share my project with future Stats203 students.

Contents

1 Introduction

In general, ecological inference refers to the process of gleaning information regarding the distriubtion of individual-level features from aggregate-level statistics (historically known as *ecological* statistics). Of course, often an aggregate-level statistic is in fact *sufficient* regarding inference of the distribution of individual-level features; for example, for many distributions, a sample mean is sufficient for estimating a population mean. The term *ecological inference* is therefore typically reserved for the special case of inferring the *joint distribution between individual-level binary features* from aggregate measurements of those features (i.e. sample proportions). When inappropriate inferences are drawn from aggregate measurements regarding individual features, one is succumbing to the *ecological fallacy*

For binary features with known marginal distributions, knowledge of the joint distribution is equivalent to knowledge of any of

- the covariance,
- the correlation, or
- the conditional distribution

of these features. Let X and Y have known marginal distributions $Bernoulli(p_X)$ and $Bernoulli(p_Y)$, respectively, then

$$\begin{aligned} Cor(X,Y) &= \frac{1}{\sigma_X \sigma_Y} Cov(X,Y) \\ &= \frac{1}{\sigma_X \sigma_Y} \left(E(XY) - EXEY\right) \\ &= \frac{1}{\sqrt{p_X(1-p_X)p_Y(1-p_Y)}} (\underbrace{P(XY=1)}_{\substack{=P(X=1|Y=1)p_Y \\ =P(Y=1|X=1)p_X}} - p_X p_Y). \end{aligned}$$

Throughout the literature, this problem is most frequently analyzed via the unknown conditional probability $P(X=1|Y=1)$, which we will refer to as $p_{X|Y}$. Several motivating examples will be given in Section 2, but the most common example of such a conditional probability is "probability of voting while Black". The term *ecological correlation* is used almost interchangeably with this conditional probability due to the above equivalence of information for binary variables, and in this section we will mostly analyze correlation rather than conditional probability.

The fundamental challenge of ecological inference, *aggregation bias*, is defined in [5] as:

> This is the effect of the information loss that occurs when individual-level data are aggregated into the observed marginals. The problem is that in some aggregate data collections, the type of information loss may be selective, so that inferences that do not take this into account will be biased.

At its core, the study of ecological correlation articulates the relationships between the following individual- and aggregate-level quantities, taking into account aggregation bias:

	Individual	Aggregate
Population	Individual-Level, Population Parameters	Aggregate-Level, Population Parameters
Sample	Individual-Level, Sample Statistics	Aggregate-Level, Sample Statistics

We will focus on the relationship between *individual-level* and *aggregate-level* (marginal) and *population parameters* and *sample statistics*. The population parameters and sample statistics at these levels have a relationship that is most easily described via a *two-way table* as follows.

- **Population Parameters**

		X		
		$= 1$	$= 0$	Marginal:
Y	$= 1$	p_{XY}	$p_{\neg XY} = p_Y - p_{XY}$	p_Y
	$= 0$	$p_{X\neg Y} = p_X - p_{XY}$	$p_{\neg X\neg Y} = 1 - p_X - p_Y + p_{XY}$	$1 - p_Y$
	Marginal:	p_X	$1 - p_X$	1

  Here, table entries denote joint probabilities and marginal probabilities fill the role of "aggregate" data. In the work of Gary King (notably [5]), table entries often instead contain conditional probabilities, which of course carry the exact same information ($p_{X|Y} = \frac{p_{XY}}{p_Y}$, for example).

- **Sample Statistics**

		X		
	Aggregation A	$= 1$	$= 0$	Marginal:
Y	$= 1$	?	?	$\hat{p}_{Y,A} = \bar{Y}_A$
	$= 0$	?	?	$1 - \hat{p}_{Y,A}$
	Marginal:	$\hat{p}_{X,A} = \bar{X}_A$	$1 - \hat{p}_{X,A}$	

  Here, individual-level entries are written as "?" (a question mark) to emphasize that they are not observed, whereas marginal statistics are observed. The label "Aggregation A" and the subscripts with A (which will not typically be practical to include) are essentially indices emphasizing that:

  - Multiple aggregate-level samples are available. That is, data may be available as $\{(\bar{X}_A, \bar{Y}_A) : A = 1, \ldots, m\}$,
  - There is aggregation bias, and so (X, Y) may not be independent of their grouping A. The aggregation A is used as an index instead of the typical counting indices i or j to emphasize that each observation of $(\overline{X}, \overline{Y})$ is not randomly sampled from the same population. This will be articulated in Section 1.2.

The quantities of interest are typically the joint (or conditional) population parameters (such as p_{XY}), though sometimes the joint sample statistics themselves (denoted "?" above) are desired. The quantities at hand are typically the aggregated sample statistics (such as $\bar{X}_A$).

The following sections explain when aggregation bias does and does not occur, and will introduce some notation common in the literature.

1.1 Ecological Correlation in *Random Samples*

Ecological inference is not a problem in general when aggregations are *random samples*. Suppose $\bar{X}$ and $\bar{Y}$ are random samples of X and Y in n individuals, which for the moment need not be binary, with means and variances μ_X, μ_Y, σ_X^2, and σ_Y^2. Then

$$\begin{aligned}
Cor(\bar{X},\bar{Y}) &= \frac{1}{\frac{1}{\sqrt{n}}\sigma_X \frac{1}{\sqrt{n}}\sigma_Y}\left[E(\bar{X}\bar{Y}) - E(\bar{X})E(\bar{Y})\right] \\
&= \frac{1}{\frac{1}{\sqrt{n}}\sigma_X \frac{1}{\sqrt{n}}\sigma_Y}\left[E(\frac{1}{n^2}\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i) - \mu_X\mu_Y\right] \\
&= \frac{1}{\frac{1}{\sqrt{n}}\sigma_X \frac{1}{\sqrt{n}}\sigma_Y}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E(X_iY_j) - \mu_X\mu_Y\right] \\
&= \frac{1}{\frac{1}{\sqrt{n}}\sigma_X \frac{1}{\sqrt{n}}\sigma_Y}\left[\frac{1}{n^2}(2\binom{n}{2}\mu_X\mu_Y + nE(X_1Y_1)) - \mu_X\mu_Y\right] \\
&= \frac{1}{\frac{1}{\sqrt{n}}\sigma_X \frac{1}{\sqrt{n}}\sigma_Y}\left[(1-\frac{1}{n})\mu_X\mu_Y + \frac{1}{n}E(X_1Y_1) - \mu_X\mu_Y\right] \\
&= \frac{1}{\sigma_X\sigma_Y}\left[E(X_1Y_1) - \mu_X\mu_Y\right] \\
&= Cor(X,Y).
\end{aligned}$$

This ensures that, given aggregate-level measurements $\bar{X}$ and $\bar{Y}$ *from random samples*, the correlation between the individual-level features X and Y can be estimated by estimating the correlation between $\bar{X}$ and $\bar{Y}$. For example, a Pearson sample correlation calculated for sample proportions is an estimator for individual correlation.

The rest of this report aims to explain the extent to which aggregate-level correlation *does not* estimate individual-level correlation.

1.2 Treating Aggregation as a Feature

A natural but uncommon (and eventually cumbersome) articulation of this problem is that individual i in the population has features (X_i, Y_i, A_i), where A_i denotes the aggregation into which they fall, and $A \not\perp (X,Y)$ in general.

Let $\bar{X}_A$ and $\bar{Y}_A$ be sample means (i.e. sample proportions) of the individual-level features X and Y in aggregation A, i.e. $\bar{X}_A$ estimates $E(X_i|A_i = A)$ by the formula

$$\bar{X}_A = \frac{1}{\sum_{i=1}^{n}\mathbf{1}(A_i = A)}\sum_{i=1}^{n} X_i\mathbf{1}(A_i = A) = \frac{1}{n_A}\sum_{i=1}^{n} X_i\mathbf{1}(A_i = A).$$

where the (random) variable n_A denotes how many samples are in aggregation A.

The aggregation index A is used rather than a common counting index i or j to emphasize that the units of aggregation need not be independent of X and Y, and so $\bar{X}_A$ and $\bar{Y}_A$ are not means of random samples in general. If aggregation A does constitute a random sample of individuals from

the entire population, meaning $A \perp (X, Y)$, then ecological inference is not a problem, as explained in the previous section.

1.3 Extreme Example : Ecological Correlation in a Highly Non-Random Sample

On the other end of the spectrum when X and Y are binary, consider the case where $A_i = X_i$, meaning A_i contains *all the information about* X_i. That is, a sample (X_i, Y_i) is in aggregation 1 (i.e. $A_i = 1$) if and only if $X_i = 1$, otherwise it is in aggregation 0 (i.e. $A_i = 0$). In this case, $(\bar{X}_0, \bar{Y}_0)$ and $(\bar{X}_1, \bar{Y}_1)$ are far from means of random samples, and calculating $Cor(\bar{X}, \bar{Y})$ takes more care, and $Cor(\bar{X}, \bar{Y}) \neq Cor(X, Y)$ (asymptotically or otherwise). Note that $\bar{X}_0 = 0$ and $\bar{X}_1 = 1$ exactly in this case.

Reader note: I initially hoped this made-up example would be short, but it was long and I left it in anyway. It may not be easy to follow in places.

The ecological correlation is

$$Cor(\bar{X}, \bar{Y}) = \frac{E(\bar{X}\bar{Y}) - E(\bar{X})E(\bar{Y})}{\sigma_{\bar{X}}\sigma_{\bar{Y}}}.$$

We can begin to calculate the quantities above using iterated expectation conditioned on aggregation A; none of the following are different from what they would be in a random sample case:

$$\begin{aligned}
E(\bar{X}) &= E\left(E(\bar{X}_A|A)\right) \\
&= P(A=0)E(\bar{X}_0) + P(A=1)E(\bar{X}_1) \\
&= (1-p_X)\cdot 0 + p_X \cdot 1 \quad \text{(because } X = 1 \text{ whenever } A = 1 \text{ in this extreme example)} \\
&= p_X. \\
E(\bar{Y}) &= E\left(E(\bar{Y}_A|A)\right) \\
&= P(A=0)E(\bar{Y}_0) + P(A=1)E(\bar{Y}_1) \\
&= (1-p_X)P(Y=1|X=0) + p_X P(Y=1|X=1) \\
&= (1-p_X)P(Y=1|X=0) + p_X P(Y=1|X=1) \\
&= (1-p_X)\frac{p_Y - p_{XY}}{1-p_X} + p_X \frac{p_{XY}}{p_X} \\
&= p_Y. \\
E(\bar{X}\bar{Y}) &- E(E(\bar{X}_A\bar{Y}_A|A)) \\
&= P(A=0)E(\underbrace{\bar{X}_0}_{=0}\bar{Y}_0) + P(A=1)E(\bar{X}_1\bar{Y}_1) \\
&= p_X E(\bar{Y}_1) \\
&= p_X p_{Y|X} \\
&= p_{XY}.
\end{aligned}$$

The variance of $\bar{X}$ and $\bar{Y}$, on the other hand, are quite different from the random sample case:

$$\begin{aligned}
\sigma^2_{\bar{X}} &= E(\bar{X}^2) - E(\bar{X})^2 \\
&= E(E(\bar{X}^2_A|A)) - p_X^2 \\
&= P(A=0)E(\bar{X}_0^2) + p(A=1)E(\bar{X}_1^2) - p_X^2 \\
&= (1-p_X)\cdot 0 + p_X \cdot 1 - p_X^2 \\
&= p_X - p_X^2. \\
\sigma^2_{\bar{Y}} &= E(\bar{Y}^2) - E(\bar{Y})^2 \\
&= E(E(\bar{Y}^2_A|A)) - p_Y^2 \\
&= P(A=0)E(\bar{Y}_0^2) + p(A=1)E(\bar{Y}_1^2) - p_Y^2 \\
&= (1-p_X)E(\bar{Y}_0^2) + p_X E(\bar{Y}_1^2) - p_Y^2,
\end{aligned}$$

note that $\bar{Y}_0 \xrightarrow{\mathcal{L}} p_{Y|\neg X} = \frac{p_Y - p_{XY}}{1-p_X}$ by LLN, and so by the Helly-Bray Theorem $E(\bar{Y}_0^2) \to \left(\frac{p_Y - p_{XY}}{1-p_X}\right)^2$, and similarly $E(\bar{Y}_1^2) \to \left(\frac{p_{XY}}{p_X}\right)^2$:

$$\begin{aligned}
\sigma^2_{\bar{Y}} &= (1-p_X)\left(\frac{p_Y - p_{XY}}{1-p_X}\right)^2 + p_X\left(\frac{p_{XY}}{p_X}\right)^2 - p_Y^2 \\
&= \frac{(p_Y p_X - p_{XY})^2}{p_X(1-p_X)}.
\end{aligned}$$

In this case,

$$\begin{aligned}
Cor(\bar{X},\bar{Y}) &= \frac{E(\bar{X}\bar{Y}) - E(\bar{X})E(\bar{Y})}{\sigma_{\bar{X}}\sigma_{\bar{Y}}} \\
&= \frac{p_{XY} - p_X p_Y}{p_X p_Y - p_{XY}} \\
&= -1,
\end{aligned}$$

whereas $Cor(X,Y) = \frac{p_{XY} - p_X p_Y}{\sqrt{p_X p_Y (1-p_X)(1-p_Y)}}$.

Note that in the iterated expectations conditioned on aggregation A, we essentially weight the within-aggregation expectations by the relative frequency of aggregation membership: $E(\overline{X}_1)$ and $E(\overline{Y}_1)$ are weighted by $P(A=1)(= P(X=1) = p_X)$, which is the probability that an individual falls into this unit of aggregation. This concept of weighting by membership will be elaborated on in *within-group correlation* and *ecological correlation*in Section 1.4 (the next section), which reveals the relationship between three very natural sample statistics describing correlation.

1.4 Relationship Between Sample Correlations at the *Individual*, *Within-Group*, and *Ecological* Levels

There are in fact several natural ways to conceive of sample correlation when either individual- or aggregate-level data are available, and three of them satisfy a useful relationship between each other that conveys the limitations of ecological correlation, first noted in [7]. We use the notation of Section 1.2 with n total individuals, where the i'th individual has features (X_i, Y_i, A_i), with A_i denoting the unit of aggregation into which this individual falls.

We define the following sample statistics that depend on aggregation of the data:

- Define n_A and $\bar{X}_A$ as in Section 1.2:

$$\bar{X}_A = \frac{1}{\sum_{i=1}^{n} \mathbf{1}(A_i = A)} \sum_{i=1}^{n} X_i \mathbf{1}(A_i = A) = \frac{1}{n_A} \sum_{i=1}^{n} X_i \mathbf{1}(A_i = A). (\star)$$

- Define (biased) sample variance S_X and S_Y in the usual way (over all n individuals), and let S_X^A and X_Y^A denote the (biased) sample standard deviations *within* aggregation A:

$$S_X^A = \sqrt{\frac{1}{n_A} \sum_{i=1}^{n} X_i^2 \mathbf{1}(A_i = A) - \overline{X}_A^2}, S_Y^A = \text{ analogous.}$$

- Let S_{XY}^A denote the (biased) sample covariance between X and Y within aggregation A:

$$S_{XY}^A = \frac{\frac{1}{n_A} \sum_{i=1}^{n} X_i Y_i \mathbf{1}(A_i = A) - \overline{X}_A \overline{Y}_A}{S_X^A S_Y^A}.$$

- Let η_{XA} and η_{YA} measure the extent to which X and Y vary over aggregations; that is, they are the sample standard deviations of $\overline{X}_A$ and $\overline{Y}_A$ with respect to A, weighted by n_A:

$$\eta_{XA} = \sqrt{\sum_{a=1}^{m} \frac{n_A}{n} \overline{X}_A^2 - \overline{X}^2}, \eta_{YA} = \text{ analogous.} (\star)$$

Finally, we define the three forms of correlation that will be related:

- Let r denote the *individual sample correlation*, i.e. the usual (biased) Pearson's correlation coefficient between X and Y

$$r = \frac{\frac{1}{n} \sum_{i=1}^{n} X_i Y_i - \overline{XY}}{S_X S_Y}.$$

- Let r_w denote a weighted average of the individual sample correlations within each aggregation:

$$r_W = \sum_{A=1}^{m} \frac{n_A}{n} \frac{S_{XY}^A}{S_X^A S_Y^A}.$$

- Define the *ecological correlation* r_e as the correlation of the $\overline{X}_A$ and $\overline{Y}_A$ aggregate statistics, weighted by population:
$$r_e = \frac{\sum_{A=1}^{m} \frac{n_A}{n} \overline{X}_A \overline{Y}_A - \overline{XY}}{\eta_{XA}\eta YA}.(\star)$$
Note that we use the fact that $\overline{(\overline{X}_A)} = \overline{X}$ and $\overline{(\overline{Y}_A)} = \overline{Y}$, which is not hard to show.

The definitions marked with a $(\star)$ can be calculated only using the observations $\{(\bar{X}_A, \bar{Y}_A) : A = 1, \ldots, m\}$ and can be considered "given" in the typical scenario; quantities defined without this mark *cannot* be calculated from aggregate-level observations, and can only be estimated subject to the limitations described in the following sections.

Theorem 1. *The following relationship holds:*

$$r_e = k_1 r - k_2 r_w, \text{ where } k_1 = \frac{1}{\eta_{XA}\eta_{YA}} \text{ and } k_2 = \frac{\sqrt{1-\eta_{XA}^2}\sqrt{1-\eta_{YA}^2}}{\eta_{XA}\eta_{YA}}. \tag{1}$$

and so $r_e = r$ if and only if

$$r_w = \underbrace{\left(\frac{1-\eta_{XA}\eta_{YA}}{\sqrt{1-\eta_{XA}^2}\sqrt{1-\eta_{YA}^2}}\right)}_{c} r.$$

It is easy to show $c \geq 1$; furthermore, in practice, r_w is smaller than r except in contrived examples. This means $r_e > r$ except in contrived examples.

Proof. This fact was first noted in [7] and it is noted in [8] that the proof consists only of algebra, though as far as I can tell it is not a short explanation and it is not included here. This fact is also noted (with different notation) in [2]. □

1.5 Asymptotic Distribution of Individual-Level and Aggregate-Level (Ecological) Sample Correlations

Here we address the case in which aggregate (population) correlation is equal to individual-level (population) correlation, namely when aggregations constitute random samples. In this case, it is possible to estimate the individual population correlation between two features from the sample correlation of aggregate statistics, but they will have different rates of convergence.

As in lecture notes 12.1 from Statistics 203, letting ρ denote the population correlation between X and Y and defining the sample correlation as $r = \frac{S_{XY}}{S_X S_Y}$, we have

$$\sqrt{n}(r - \rho) \xrightarrow{\mathcal{L}} N(0, \gamma^2),$$

where

$$\gamma^2 = \frac{1}{4}\rho^2 \left[\frac{C(XX,XX)}{\sigma_X^4} + 2\frac{C(XX,YY)}{\sigma_X^2\sigma_Y^2} + \frac{C(YY,YY)}{\sigma_Y^4}\right] - \rho\left[\frac{C(XX,XY)}{\sigma_X^3\sigma_Y} + \frac{C(XY,YY)}{\sigma_X\sigma_Y^3}\right] + \frac{C(XY,XY)}{\sigma_X^2\sigma_Y^2},$$

where

$$C(XX, XX) = Cov((X - \mu_X)^2, (Y - \mu_Y))$$
$$C(XX, XY) = Cov((X - \mu_X)^2, (X - \mu_X)(Y - \mu_Y)),$$

etc. for the other terms.

What about the asymptotic distribution of the ecological correlation r_e from Section 1.4? We have two options for attacking this problem:

- Replicate the analysis of the asymptotic distribution of r, using the CLT and the delta method (Cramér's Theorem).
- Utilize the relationship between r, r_w and r_e from Equation 1 and apply the Delta Method to the transformation of r to derive an asymptotic distribution of r_e.

These are left to a future project! This does not appear in any of the literature I have reviewed.

1.6 Convergence Rate of Ecological Correlation Versus Individual Correlation in Random Samples

This also does not appear in any literature I reviewed. I am interested in studying this topic further. Surely this has either been addressed, or is easily derivable in a manner similar to analysis of Pearson's correlation in any statistics textbook.

2 Examples and Literature Review

2.1 Motivating Examples

The following examples show that the considerations related to ecological inference are applicable across nearly the entirety of empirical statistical research.

2.1.1 Voting Demographics

Throughout Gary King's work, the most common running example is that of (unknown) political preferences among demographic groups.

Throughout US electoral politics, aggregate voting information is typically available at the national, state, district, and precinct level, but not the individual level. Furthermore, the US Census measures regional demographic makeup at the individual level but presents its data only at the "tract", "block group", and "block" levels (depending on the nature of the data, it may be available at the block level or only at the coarser tract level).

King presents almost all ecological inference concepts via the example of inferring "probability of voting while Black" β_i^B from "fraction who voted" and "fraction who are Black" in a given aggregation. This is a conditional probability ($\beta_i^B = P(\text{Voted}|\text{Black})$); as explained in Section 1, in

this context (of known marginals) a conditional probability is informationally equivalent to a joint probability or a correlation between two binary variables. Figure 1 demonstrates this relationship.

		Voted		
		Yes	No	Tot:
Race	Black	β_i^B	$1-\beta_i^B$	X_i
	Non-Black	β_i^w	$1-\beta_i^w$	$1-X_i$
	Tot:	T_i	$1-T_i$	1

Figure 1: *A two-way (or "four-fold") table describing the relationship between known aggregate-level statistics and unknown individual-level statistics/parameters. X_i represents the (known) fraction of voting age people in district i who are Black, T_i represents the (known) fraction of people in district i who vote, and β_i^B and β_i^w represent the (unkown) fraction of Black and white people who vote, respectively.*

2.1.2 Gene Co-Expression

The submitted paper [2] questions the assumption that gene or protein co-expression within cells can be inferred from correlation between genetic material or proteins in tissues. In particular, this implies that pairs of genes or proteins that are correlated in tissues need not be part of any molecular cellular mechanism.

2.1.3 College Admissions: Simpson's Paradox

In the 1970s, The University of California, Berkeley was sued on the grounds of gender bias for admitting a (statistically significantly) higher fraction of men than women to its graduate school overall. Each of the departments was instructed to audit its admissions practices, and nearly every department actually reported admitting a *higher* fraction of women than men.

The school analyzed the results and discovered that women tended to apply more frequently to departments with lower overall acceptance rates. For simplicity, suppose there are two departments: Literature, which receives less funding but has few firm prerequisites, resulting in many applications with low acceptance rates; Science, which is well-funded and therefore can admit many students, and furthermore has firm prerequisites, meaning most of those who apply are qualified and are in fact admitted.

Men (possibly due to discrimination elsewhere within US education) tend to apply in greater proportion to Science programs, with high admission rates; women tend to apply to Literature programs with low admission rates. Therefore, even though the Science and Literature departments may both admit women at as high or higher rates than men, the overall admission rate for women may be much lower.

This phenomenon is known as *Simpson's Paradox*, referring specifically to the case when proportions of a feature within subpopulations are higher for one group but, due to unequal distribution among the subpopulations, the proportion of a feature among the entire population is lower. A general problem in ecological inference is that these proportions may not be equal in subpopulations and the general population, but it is only an example of Simpson's Paradox if the dominant subpopulation is different in the subpopulation versus the entire population.

Figure 2 illustrates that when the proportion of women applying to Science is equal to that of men, the proportion of women who are admitted is greater than that of men; when the rate of women applying to Science falls sufficiently below that of men, their overall acceptance rate falls below that of men. Throughout, the acceptance rate of women to both departments is higher than that of men.

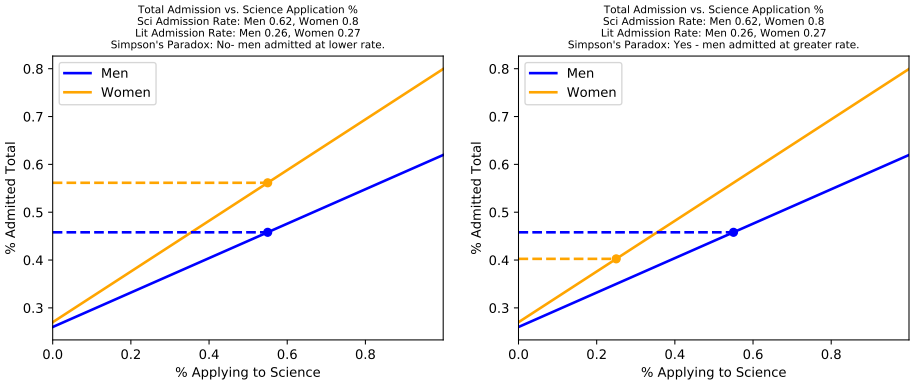

Figure 2: *The overall acceptance rates (and acceptance rates relative to each other) of men and women depend not only on their acceptance rates within each department, but on their likelihood of applying to the different departments.*

This can be articulated in the language of Section 1.2 using r, r_w, and r_e as follows: Individual i has three features: (X_i, Y_i, A_i), where

$$X_i = \begin{cases} 1 : \text{male} \\ 0 : \text{female}, \end{cases}$$

$$Y_i = \begin{cases} 1 : \text{admitted} \\ 0 : \text{not admitted}, \end{cases}$$

$$A_i = \begin{cases} 1 : \text{applied to Science} \\ 0 : \text{applied to Literature}. \end{cases}$$

Then $\bar{X}_A$ and $\bar{Y}_A$ denote the fraction of men who apply to a department, and the fraction admitted to a department in department A.

- The individual (Pearson) correlation between maleness and admission is positive: $r = Cor(X, Y) > 0$, meaning on average, male applicants tend to be admitted at higher rates.
- The ecological correlation r_e (calculated using iterated expectations conditioned on department) is also positive: $Cor(\overline{X}, \overline{Y}) > 0$, meaning departments with more male applicants tend to have higher admission rates.
- The individual correlation *within* department A is negative: $Cor_{|A}(X, Y) < 0$ for all A, leading to a negative r_w (the average of these correlations weighted by department size),

meaning that, on average, within departments, maleness correlates with *lower* admission.

The relationship between r, r_e, and r_w in Equation 1 clarifies how these three quantities can have different signs.

2.1.4 Literacy

In the founding paper of the study of ecological inference [7], Robinson notes that aggregate literacy statistics may lead to wrong conclusions. This paper (with racial insensitivity reflective of 1950s scholarship) gives two examples using literacy rates in the 1930 Census's nine "geographic divisions" of the United States.

First, he compares "percent illiterate" to "percent [Black]":

> The data of this section show that the individual correlation between color and illiteracy is .203, while the ecological correlation is .946. In this instance, the two correlations a t leasst have the same sign, and that sign is consistent with our knowledge that educational standards in the United States are lower for [Black Americans] than for whites.

In another example (with more offensive undertones):

> consider another correlation where we also know what the sign ought to be, viz., that between nativity and illiteracy. We know that educational standards are lower for the foreign born than for the native born, and therefore that there ought to be a positive correlation between foreign birth and illiteracy. This surmise is corroborated by the individual correlation between foreign birth and illiteracy...The individual correlation...is .118. However, the ecological correlation between foreign birth and illiteracy...is -0.619! When the ecological correlation is computed on a state rather than a divisional basis, its value is -0.526.

2.1.5 Relationship to Other Problems

Ecological inference shares qualities with other problems involving missing data. The following problems are distinct from ecological inference, but share similarities worth mentioning. Below is some brainstorming on this idea.

- The field of *differential privacy* studies the wide range of data manipulation techniques that can intentionally obscure information about *individual* observations in a dataset while preserving all *aggregate* features of the data to the extent that is possible. This is sometimes done by systematically adjusting observations with noise or by re-sampling from an emprical distribution to generate fictional observations. While the resulting (manipulated) data may be presented on an "individual-level", what is actually preserved from the original data is aggregate-level features.

  Differentially private data management is becoming a widely-adopted norm in commercial and government information technology, and there are efforts to legally codify such privacy guarantees.

  Generally, features of the original data that are not explicitly preserved in the mutated dataset are lost. There may still be opportunities to study how to preserve individual-level correlations in a differentially private dataset. Interpreting correlations in differentially private data may

resemble the ecological inference problem. Sensitivity analysis in the field of differential privacy may yield contributions to our understanding of ecological inference.

- Sensitivity analysis for hypothesis testing when mean imputation is used for missing data also explores the limitation of using a sample average in place of individual observations. This is a topic related to clinical trials as in [6].

2.2 Literature Review

The first articulation of the ecological inference problem is by Robinson in 1950 [7]. Apparently (as mentioned in several other texts), for the following twenty or so years, debate raged within the field of sociology regarding whether or not ecological inference was being incorrectly applied. For example, studies on the district-level correlation between a demographic feature and "having voted" were criticized for equating ecological correlation with individual correlation, but were also often defended with the explanation that the ecological correlations themselves were of interest, as the districts constituted a meaningful "unit of analysis" in electoral politics.

Shively [8] from 1969 provides a review of ecological inference. They describe both Goodman's regression and the method of bounds, along with the assumptions necessary to apply both. They give examples where ecological correlation has the opposite sign of individual correlation.

Hannan [3] from 1972 again provides a review of ecological inference. They give a detailed mathematical framework that describes Goodman's regression more in the style of a textbook than a journal article.

Clark and Avery [1] in 1972 provide yet another review of ecological inference and focuses on the specific nature of bias due to *proximity aggregation*.

In [2], it is noted that the correlation between molecular markers at the tissue level is not equivalent to correlation within cells. In particular, this precludes the (tempting) possibility of inferring gene or protein co-expression within cells (and thereby cellular mechanisms) via their coexistence within tissues. They articulate the relationship between sample correlations (Pearson's, ecological, and within-aggregation correlation) described in Equation 1 slightly differently:

$$\frac{\rho_{uv}}{\rho_{xy}} = \sqrt[2]{\left(1+\frac{\sigma_{ix}^2}{\sigma_u^2}\right)\cdot\left(1+\frac{\sigma_{iy}^2}{\sigma_v^2}\right)} - \frac{\sigma_{ix}}{\sigma_u}\cdot\frac{\sigma_{iy}}{\sigma_v}\cdot\frac{\rho_{ixiy}}{\rho_{xy}}.$$

In 1997, Gary King made the first lasting contributions on this topic since the 1970s in the his 1997 book [4], and wrote the most comprehensive reference book on the subject in 2004 in [5]. Throughout Dr. King's work, the most common running example is that of (unknown) political preferences among demographic groups.

3 Methods of Ecological Inference

For this section, we exclusively adopt the example and notation of Gary King in [4], which this section closely follows, as introduced previously in Table 1, which is repeated below. Using the

notation in Section 1.4, we use the $(\star)$ to denote quantities that are *observed*, namely aggregate statistics; quantities without this symbol are those that we aim to calculate, estimate, or bound.

Consider the voters in *precinct* i, of which there are p within a single larger *district*. Let superscripts ^{B} and ^{w} denote Black and white demographics (capitalization of "Black" and non-capitalization of "white" is intentional per anti-racist best-practices). Let

- T_i = proportion of individuals who vote in district i (the proportion that Turns out to vote) $(\star)$
- N_i^B, N_i^w = number of voting-age Black, white individuals in precinct i. $(\star)$
- N_i^{BT}, N_i^{wT} = number of Black, white individuals who vote in precinct i. Without the subscript, these denote the number of those individuals in the entire district.
- X_i = proportion of individuals who are Black (an eXplanatory variable). $(\star)$
- $\beta_i^B = \frac{N_i^{BT}}{N_i^B}$ = proportion of voting-age Black individuals who vote in district i. This is a realization/estimate of the probability of an individual voting given their demographic and precinct: $\beta_i^B = P(\text{votes}|\text{Black}, \text{precinct } i)$.
- $\beta_i^w = \frac{N_i^B}{N_i^{wT}}$ = proportion of voting-age white individuals who vote in district i. This is a realization/estimate of the probability of an individual voting given their demographic and precinct: $\beta_i^w = P(\text{votes}|\text{white}, \text{precinct } i)$.
- $B^B = \frac{\sum_{i=1}^p N_i^{BT}}{N^B}$ the proportion of voting-age Black individuals who vote over the entire district.

		Voted		
		Yes	No	Tot:
Race	Black	β_i^B	$1-\beta_i^B$	X_i
	Non-Black	β_i^w	$1-\beta_i^w$	$1-X_i$
	Tot:	T_i	$1-T_i$	1

The problem is to estimate the β and B parameters (and the N parameters they comprise) - every quantity without a $(\star)$ - from observations of the X and T variables.

3.1 Goodman's Regression

Consider the accounting identity:

$$T_i = \beta_i^B \cdot X_i + \beta_i^w \cdot (1 - X_i), \tag{2}$$

$$= (\beta_i^B - \beta_i^w)X_i + \beta_i^w, \tag{3}$$

where (2) can be clarified as:

$$\overbrace{T_i}^{P(\text{vote})} = \overbrace{\underbrace{\beta_i^B}_{P(\text{vote}|\text{Black})} \cdot \underbrace{X_i}_{P(\text{Black})}}^{P(\text{vote}\cap\text{Black})} + \overbrace{\underbrace{\beta_i^w}_{P(\text{vote}|\text{white})} \cdot \underbrace{(1-X_i)}_{P(\text{white})}}^{P(\text{vote}\cap\text{white})} \text{ in district } i.$$

Note from the form of (3) that if $T_i = aX_i + b$, then $b = \beta_i^w$ and $(a + b) = \beta_i^B$. Using this basic idea, we can estimate β_i^B and β_i^w by assuming they have the same values B^B and B^w for all precincts (all i) in the district and regressing $\{T_i : i = 1, \ldots, p\}$ on $\{X_i : i = 1, \ldots p\}$.

King notes that this method has been the most common approach to statistical inference since its study began with [7]. He succinctly describes the central limiting assumption of this method:

> ...the basic accounting identity in Equation [(3)] has twice as many unkowns ($2p$) as observations (p), that is β_i^B and β_i^w for each of p precincts, computing unrestricted estimates of all unknowns seems hopeless. The Goodman model resolves the proliferation of parameters by making the extreme "constancy assumption": $\beta_i^B = B^B$ and $\beta_i^w = B^w$ for all i. if this assumption is appropriate, then Equation [(2)] becomes manageable, since it has only two parameters:
>
> $$T_i = B^B X_i + B^w (1 - X_i).$$
>
> The problem is that if the assumption is wrong,..., the answers this model produces will often be wrong.

The issue is that the population parameters governing the samples β_i^B and β_i^w (which are unobserved) are *not* identical in different precincts; in other words, there is a correlation between the joint individual variables ("voted", "race") and the aggregations (denoted A in Section 1.2). Gary King's articulation of this concept, illustrated in Figure 3, a touchstone of his work on this topic.

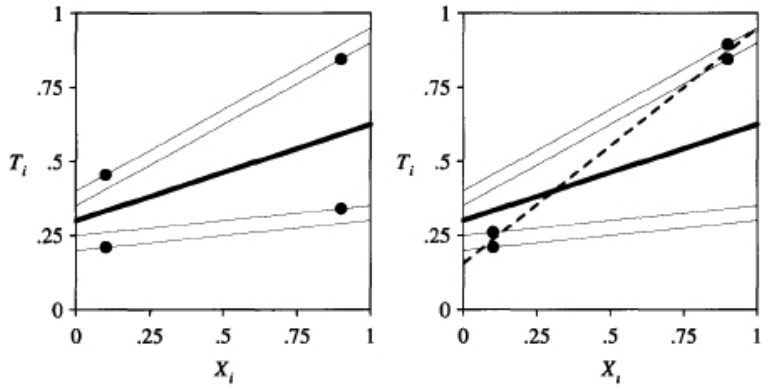

Figure 3: *From [4], "Each graph in this figure has four data points. Each point corresponds to a particular precinct with actual values X_i and T_i (where $T_i = \beta_i^B X_i + \beta_i^w (1 - X_i)$). The line through each point shows the theoretical relationship between T_i and X_i that would arise were the parameters (β_i^B, β_i^w) held constant but the value of X_i were allowed to vary from zero to one. The lines from which the points are selected are identical in both graphs. In the left graph, the slope of the lines is uncorrelated with the value of X_i at each observed point, but in the right graph points with larger values of X_i have lines with steeper slopes. The goal is to estimate a regression line from the points alone that is the average of the individual lines (represented as the dark solid line in both graphs). Since the lines for the observations are the same in the two graphs, the average line we wish to estimate is the same in both. This same dark line on the left is also exactly the regression line fit to its four points, which is an example of unbiased regression estimates that result when the slopes and X_i are uncorrelated. On the right graph, the dashed line is the regression line fit to its points, but because X_i and the slopes are correlation, it is a biased estimate of the (dark solid) average line."*

3.2 Method of Bounds and Tomography Lines

King notes that the number of black voters in precinct i, N_i^{BT} (which is unknown), cannot exceed the number of Black individuals in that district N_i^B nor the number of voters in that district N_i^T (which are both known), yielding the fact $N_i^{BT} \leq \min(N_i^T, N_i^B)$. Furthermore, $\beta_i^B = \frac{N_i^{BT}}{N_i^B}$, so we can also bound β_i^B.

This logic yields the following bounds:

$$\begin{array}{rll} \max\left(0, \frac{T_i - (1-X_i)}{X_i}\right) & \leq \beta_i^B & \leq \min\left(\frac{T_i}{X_i}, 1\right) \\ \max\left(0, \frac{T_i - X_i}{1-X_i}\right) & \leq \beta_i^w & \leq \min\left(\frac{T_i}{1-X_i}, 1\right). \end{array}$$

Furthermore, [4] notes, we know even more information about these unknown parameters than these bounds. By re-arranging Equation (3), we see that the two unknowns are related linearly by known constants:

$$\beta_i^w = \left(\frac{T_i}{1 - X_i}\right) - \left(\frac{X_i}{1 - X_i}\right) \beta_i^B.$$

This, finally, *is* all the information we know *a priori* about β_i^B and β_i^w. Thus, for each precinct i, our pair of observations (X_i, T_i) yields a *line segment* on which β_i^B and β_i^w may lie.

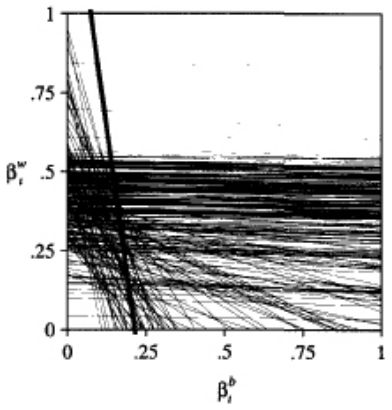

Figure 4: *For each observation of aggregate measurements (X_i, T_i), the maximum deterministic information possible regarding the quantities of interest (β_i^B, β_i^w) is that the pair lies on a line segment termed a* tomography line *in [4].*

Gary King, in both [4] and again in [5] mentions the use of "tomography lines" to represent the deterministic information contained in ecological correlations. It is an analogy to tomography lines in radiology, where they represent all the information one has regarding an object's location given a scan of the medium in which it is embedded.

3.3 Combination of Goodman's Regression and Method of Bounds

King's achievement in [4] (chapter 6) is to combine Goodman's Regression and the Method of Bounds via a statistical prior on (β_i^B, β_i^w). Namely, he assumes that they follow a truncated normal distribution:

$$(\beta_i^B, \beta_i^w) \sim TN(\beta_i^B, \beta_i^w | \Sigma),$$

where the truncation arises from the bounds in the previous section, and the normality is an generalization of the (fallacious) assumption of homogeneity between precincts mentioned in Section 3.1 (that $\beta_i^B = B^B$ for all i).

Via these mild statistical assumptions regarding (β_i^B, β_i^w), King improved Goodman's method for the first time in almost 50 years.

4 Conclusion

To date, the most savvy practitioners of ecological inference have been political campaigns. With or without strong theoretical frameworks, campaigns have been cross-referencing voter files (lists of registered voters that include the elections in which they did or did not vote [not, of course, for whom they voted]), census information, and precinct-level voting data for decades. In particular, voter files purchased in the dominant political consultancy marketplaces are almost always augmented

with priors on racial makeup of registered voters based on analysis of their last name, location, and age. By adding dimensions to aggregate data, the intersection between any two subpopulations conditioned on other features becomes smaller, and thus less inference is needed. That is, by adding race beliefs to voter files, we can somewhat confidently answer "what is the percent of Black Americans who voted?" and then come closer to answering "what percentage of Black Americans voted for a Democrat?" and other related questions.

Moving forward in an age of big data, tracking, mass surveillance, and the Internet of Things, it is likely that a proliferation of individual-level data will make previous Census-based aggregate population studies outdated. On the other hand, as our data collection becomes more sophisticated, so too does our world become more complex, and it's likely that the desire for more specific data will keep pace with its availability. The desire to estimate relationships at levels finer than those in the data will likely persist. I mentioned a possible relationship between ecological inference and sensitivity to missing data in general optimization problems, and this type of concern is likely to become more prominent as data-driven automation becomes more widespread.

On the other hand, the current "wild west" of data from ethically dubious tracking may give way not to a future of unbridled surveillance but to one of robust data privacy. I have mentioned the natural relationship that comes to mind between ecological inference and differential privacy. It is likely that as the data available to population researchers becomes richer, its limitations will be engineered in intentionally rather than artifacts of its collection mechanism. This may enable finer-grained data while still preserving individual privacy in census data and voter files. Making that possible is in some ways an exercise in ecological inference.

Ecological inference has always been presented as a problem that the ignorant would rather overlook, and still not much useful methodology has been produced to address it. Most of the literature has been about the limitations of attempting ecological inference in the first place, and there has been significant criticism of all the methods proposed to address it. In particular, Goodman's Regression was never widely adopted outside of a small group of sociological researchers, and even in that community it received about as much criticism as adoption. Perhaps the topic itself is too broad to be useful, and each research community will find their own domain-specific ways to acknowledge and address ecological inference.

Note that I left Sections 1.5 and 1.6 incomplete. This project is a survey, and these analyses of asymptotic behavior, while exceedingly relevant to this Large Sample Theory course, are not the focus of any of the resources I relied on.

References

[1] W. A. V. Clark and Karen L. Avery. The effects of data aggregation in statistical analysis. *Geographical Analysis*, 8(4):428–438, 1976.

[2] Yuchao Li Guoyu Wu. Distinct characteristics of correlation analysis at the single-cell and the population level. *Journal of the Royal Society Interface*, 63, 2020.

[3] Michael T Hannan. Approaches to the aggregation problem. *Stanford Sociology Technical Reports and Working Papers 1961-1993*, (46), 1972.

[4] Gary King. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press, 1997.

[5] Gary King, Ori Rosen, Martin Tanner, Gary King, Ori Rosen, and Martin A. Tanner. *Ecological Inference: New Methodological Strategies*. Cambridge University Press, New York, 2004.

[6] Devan V. Mehrotra, Fang Liu, and Thomas Permutt. Missing data in clinical trials: control-based mean imputation and sensitivity analysis. *Pharmaceutical Statistics*, 16(5):378–392, 2017.

[7] W. S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357, 1950.

[8] W. Phillips Shively. "ecological" inference: The use of aggregate data to study individuals. *American Political Science Review*, 63(4):1183–1196, 1969.