Estimating Preferences Over Data to Inform Statistical Disclosure Methods Decisions

Elan Segarra

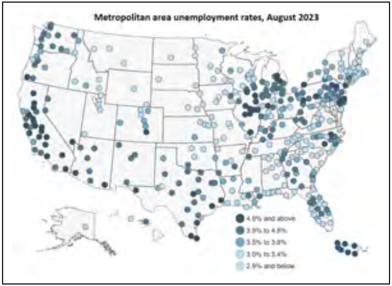
U.S. Bureau of Labor Statistics

FCSM

October 25th, 2023

Disclaimer: The views expressed herein are those of the author(s) and do not necessarily reflect those of the Federal Government, Department of Labor, or the Bureau of Labor Statistics. All results have been reviewed to ensure that no confidential information is disclosed.

BLS Publishes Many Many Statistics



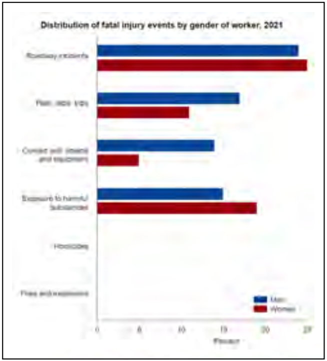


TABLE A-7. Fatal occupational injuries by worker characteristics and event or exposure, all United States, 2021

Worker Characteristics	Total fatal injuries (number)	Event or exposure(1)					
		Transportation incidents(2)	Violence and other injuries by persons or animals(3)	Contact with objects and equipment	Falls, slips, trips	Exposure to harmful substances or environments	Fires and explosions
Total	5,190	1,982	761	705	850	798	76
Employee status							
Wage and salary(4)	4,284	1,686	613	535	690	685	59
Self-employed(5)	906	296	148	170	160	113	17
Gender							
Women	448	175	111	23	49	86	-
Men	4,741	1,807	650	682	801	711	-

Many Dimensions of Publication Choice

How we implement statistical disclosure control (SDC) methods is one important dimension:

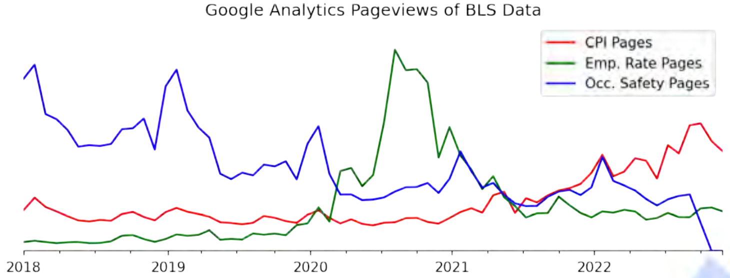
1. When using cell suppression approaches, which complementary cells should we suppress?
2. When using formally private methods, how should we divvy up the privacy budget among published statistics?

Many Dimensions of Publication Choice

How we implement statistical disclosure control (SDC) methods is one important dimension:

1. When using cell suppression approaches, which complementary cells should we suppress?
2. When using formally private methods, how should we divvy up the privacy budget among published statistics?

To help, we need a way to measure relative value of publications...



Project Overview

Goal: Quantify public data users' preferences over the statistics that BLS publishes.

Data:

- Census of Fatal Occupational Injuries (CFOI): Data being consumed
- Google Analytics: Data used to estimate value/preferences

Method: Estimate a nested logit model of consumer preferences.

Results:

- CFOI stats broken down by employment status have the highest value
- CFOI preference estimates *are not useful* for informing cell suppression disclosure control methods
- However they *are useful* for allotting privacy budgets in formally private disclosure control methods (e.g. differentially private noise infusion)

Subject of Interest: CFOI

The Census of Fatal Occupational Injuries (CFOI) is the subject data set for this project:

- Annual census of fatal work injuries collected since 1992
- Compiled by a Federal-State cooperative program which collects data info multiple sources (police reports, news, OSHA investigations, etc.)
- Compiled data includes narratives, injury codes (OIICS), geography, timing, and demographic information.

Subject of Interest: CFOI

The Census of Fatal Occupational Injuries (CFOI) is the subject data set for this project:

- Annual census of fatal work injuries collected since 1992
- Compiled by a Federal-State cooperative program which collects data info multiple sources (police reports, news, OSHA investigations, etc.)
- Compiled data includes narratives, injury codes (OIICS), geography, timing, and demographic information.

Disclosure control is particularly difficult for CFOI:

- It is a census, the counts are small, and some data is public
- BLS publishes many tables/figures/statistics using CFOI (e.g. industry and occupational breakdowns)
- Currently use cell suppression to protect confidentiality

Model of Consumer Choice Over Statistics

There are three levels to the model of data consumer preferences:

1. A **statistic** is an individual number
   - Ex: The count of work fatalities in the construction sector (NAICS 23)
2. A **publication** is a collection of statistics
   - Ex: A table/figure of work fatality counts by industry (2 digit NAICS)
3. A **market** consists of a set of publications at a specific time from which a data consumer can choose
   - Ex: On BLS.gov a data consumer can choose to view fatality counts by employee status, industry, occupation, or age group

Key insight: Observing which publications are chosen (i.e. clicked on) reveals preferences over the underlying statistics

Model of Consumer Choice: Nested Logit

Consumer i has indirect utility from publication p in market t given by

$$U_{ipt} = \underbrace{\frac{1}{|\mathcal{S}_p|} \sum_{s \in \mathcal{S}_p} X_{st}\beta + W_{pt}\theta + \xi_{pt}}_{\equiv \delta_{pt}} + \varsigma_{ig} + (1-\sigma)\epsilon_{ipt}$$

$\mathcal{S}_p$: Set of statistics included in publication p
X_{st} : Observable characteristics of statistic s (eg ind. breakdown)
W_{pt} : Observable characteristics of publication p (eg bar chart)
ξ_{jt} : Unobservable stat. characteristics (eg ugly presentation)
ς_{ig} : Unobservable correlated nest shock (eg broken site link)
ϵ_{ipt} : Unobservable characteristics (eg researcher vs layperson)

Objects of interest: β and θ quantify preferences across characteristics

Nest Structure

BLS

Model of Consumer Choice: Utility Maximization

If users choose the data product (e.g. table) that maximizes their indirect utility, then the observed "market share" of publication p in time period t is given by

$$s_{pt} = \frac{\exp\left(\frac{\delta_{pt}}{1-\sigma}\right)}{D_g^{\sigma} \sum_h D_h^{1-\sigma}} \qquad \text{where} \qquad D_g = \sum_{k \in g} \exp\left(\frac{\delta_{kt}}{1-\sigma}\right)$$

Note: "Market share" $= s_{pt} = \frac{q_{pt}}{M_t} = \frac{\text{pageviews}}{\text{site visitors}}$

Model of Consumer Choice: Utility Maximization

If users choose the data product (e.g. table) that maximizes their indirect utility, then the observed "market share" of publication p in time period t is given by

$$s_{pt} = \frac{\exp\left(\frac{\delta_{pt}}{1-\sigma}\right)}{D_g^{\sigma} \sum_h D_h^{1-\sigma}} \qquad \text{where} \qquad D_g = \sum_{k \in g} \exp\left(\frac{\delta_{kt}}{1-\sigma}\right)$$

Note: "Market share" $= s_{pt} = \frac{q_{pt}}{M_t} = \frac{\text{pageviews}}{\text{site visitors}}$

Estimation:

1. Inversion step (the "magic" of logit):

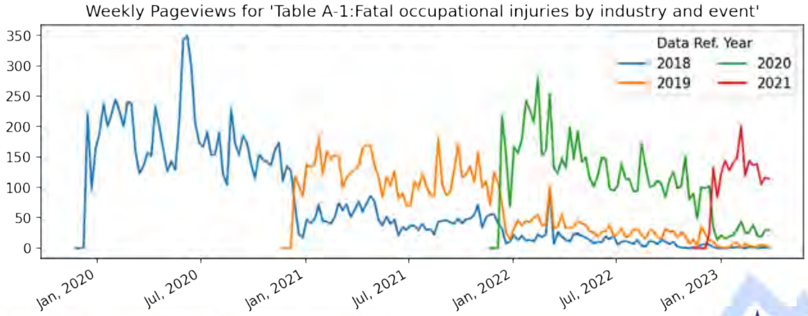
$$\ln s_{pt} - \ln s_{0t} = \tilde{X}_{pt}\beta + W_{pt}\theta + \sigma \ln s_{p|g} + \xi_{pt}$$

2. Estimate using traditional regression techniques (e.g. Two Stage Least Squares)

Data: Google Analytics

Google Analytics

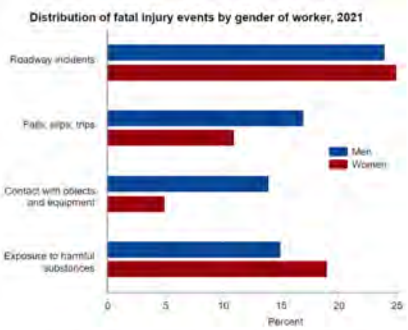
- Granular data on pageviews, duration, and even demographics
- There are 28 different tables/figures each published over multiple reference years
- Relative pageviews function as our measure of choice among consumers



Data: Extracting Characteristics

Characteristics are manually coded for each site, such as

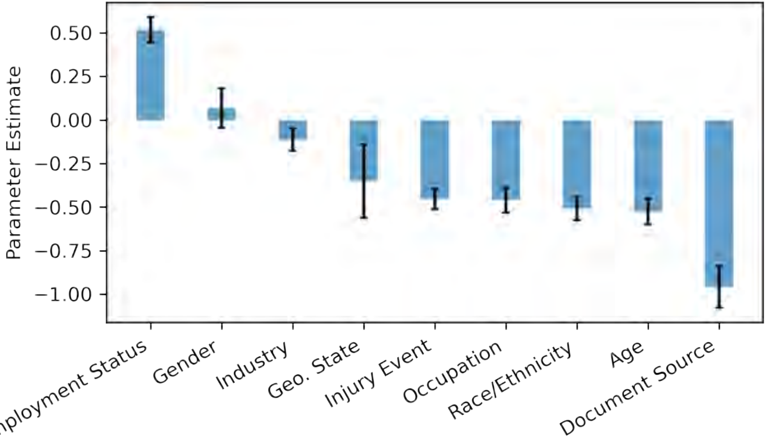
- *ind* = includes breakdown by industry sectors
- *format* = bar chart, table, time series etc.
- *multiyear* = includes more than one year of data
- *curr_year* = includes most recent RY (at view time)
- *exposition* = includes exposition along with data



$$\Rightarrow \quad x_{pt} = \begin{bmatrix} curr\_year \\ ind \\ gender \\ event \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ \vdots \end{bmatrix}$$

$$w_{pt} = \begin{bmatrix} multiyear \\ format\_bar \\ exposition \\ \vdots \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}$$

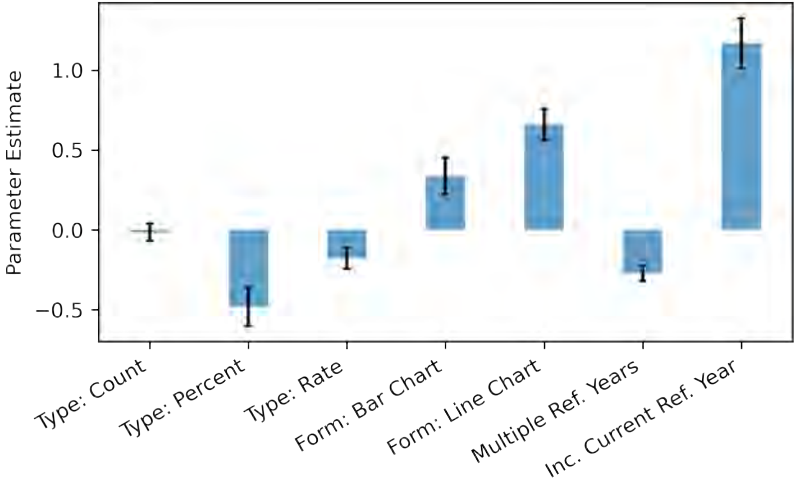
Estimation Results: Statistic Characteristics



Though the exact values of $\widehat{\beta}$ are difficult to interpret, their relative ordering reflects which breakdowns are more valued by data consumers.

Full Regression Results

Estimation Results: Publication Characteristics



Similarly, the exact values of $\widehat{\theta}$ are difficult to interpret, but they suggest which publication chars. are valued by data consumers.

Full Regression Results

Statistical Disclosure Control (SDC)

Consider these 2 tables of synthetic CFOI statistics:

Table 1: Fatalities by Age and Year

	Year			
Age	2019	2020	2021	Total
< 20	2	3	8	13
20-34	29	27	34	90
35-54	51	46	55	152
≥ 55	49	43	57	149
Total	131	119	154	404

Table 2: Fatalities by Age and Industry

	Industry (for 2021)			
Age	Cons.	Mfg.	Trade	Total
< 20	4	3	1	8
20-34	15	9	10	34
35-54	29	15	11	55
≥ 55	31	12	14	57
Total	79	39	36	154

How can we use our estimated preferences over data to inform our SDC methods?

We consider two application cases:

1. Cell Suppression Problem
2. Formally Private Noise Injection

BLS

Using the Estimation Results

Theoretically, we can use the estimated preference parameters to construct average valuations of any relevant statistics, e.g.:

- Table of fatalities by age and year: $\overline{U}_1 = -2.75$
- Table of fatalities by age and industry: $\overline{U}_2 = -2.18$

Since $\overline{U}_1 < \overline{U}_2$ we can conclude the second table is *generally* more valuable than the first on average.

Using the Estimation Results

Theoretically, we can use the estimated preference parameters to construct average valuations of any relevant statistics, e.g.:

- Table of fatalities by age and year: $\overline{U}_1 = -2.75$
- Table of fatalities by age and industry: $\overline{U}_2 = -2.18$

Since $\overline{U}_1 < \overline{U}_2$ we can conclude the second table is *generally* more valuable than the first on average.

There are two issues:

1. Negative utility/valuations may be inappropriate (e.g. for dividing a privacy budget)
2. Difficult to ascribe relative valuation since utility is equivalent up to positive affine transformation.

Using the Estimation Results

Approach:

- Identify a set of statistics and push them through the nested logit demand model
- Yields choice probabilities that can be used like valuations
- They are all positive and between 0 and 1 so can function nicely in disclosure control methods

Example

Given those 2 table options then the model predicts:

- $P(\text{Choose Table 1}) = s_1 = 0.361$
- $P(\text{Choose Table 2}) = s_2 = 0.639$

This estimates that the publication of fatalities by age and industry is approximately 77% more valuable (on average) as a publication that breaks down by age and industry.

Disclosure Application: Cell Suppression Problem

SDC Method: Tabular Cell Suppression

- Sensitive cells are suppressed to protect confidentiality.
- Additional cells (i.e. complementary/secondary) often need to be suppressed.

	Industry (for 2021)			
Age	Cons.	Mfg.	Trade	Total
< 20	4	3	1	8
20-34	15	9	10	34
35-54	29	15	11	55
≥ 55	31	12	14	57
Total	79	39	36	154

Disclosure Application: Cell Suppression Problem

SDC Method: Tabular Cell Suppression

- Sensitive cells are suppressed to protect confidentiality.
- Additional cells (i.e. complementary/secondary) often need to be suppressed.

	Industry (for 2021)			
Age	Cons.	Mfg.	Trade	Total
< 20	4	3	1	8
20-34	15	9	10	34
35-54	29	15	11	55
≥ 55	31	12	14	57
Total	79	39	36	154

Primary Suppression

Disclosure Application: Cell Suppression Problem

SDC Method: Tabular Cell Suppression

- Sensitive cells are suppressed to protect confidentiality.
- Additional cells (i.e. complementary/secondary) often need to be suppressed.

Potential Set of Complementary Suppressions

	Industry (for 2021)			
Age	Cons.	Mfg.	Trade	Total
< 20	4	3	1	8
20-34	15	9	10	34
35-54	29	15	11	55
≥ 55	31	12	14	57
Total	79	39	36	154

Primary Suppression

There are often multiple options for complementary suppressions. Estimated valuations over these cells can help guide decisions.

Disclosure Application: Cell Suppression Problem

Table 1: Fatalities by Age and Year

Age	Year 2019	2020	2021	Total
< 20	2	3	8	13
20-34	29	27	34	90
35-54	51	46	55	152
≥ 55	49	43	57	149
Total	131	119	154	404

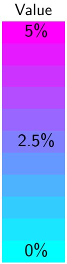


Table 2: Fatalities by Age and Industry

Age	Industry (for 2021) Cons.	Mfg.	Trade	Total
< 20	4	3	1	8
20-34	15	9	10	34
35-54	29	15	11	55
≥ 55	31	12	14	57
Total	79	39	36	154

We can then use these individual cell valuations as an input into any CSP solver to find optimal complementary suppressions.

Disclosure Application: Cell Suppression Problem

Table 1: Fatalities by Age and Year

Age	Year 2019	Year 2020	Year 2021	Total
< 20	2	3	8	13
20-34	29	27	34	90
35-54	51	46	55	152
≥ 55	49	43	57	149
Total	131	119	154	404

Value
5%
2.5%
0%

Table 2: Fatalities by Age and Industry

Age	Industry (for 2021) Cons.	Industry (for 2021) Mfg.	Industry (for 2021) Trade	Total
< 20	4	3	1	8
20-34	15	9	10	34
35-54	29	15	11	55
≥ 55	31	12	14	57
Total	79	39	36	154

We can then use these individual cell valuations as an input into any CSP solver to find optimal complementary suppressions.

Note: Doesn't perform markedly different from optimizing the number of suppressed cells because:

1. Estimated cell valuations do not have much variation
2. GA data don't allow for disentangling many cell valuations

Disclosure Application: Differentially Private Tables

Differential Privacy (DP):

- Publication property that provides a provable confidentiality guarantee
- Increasing adoption of DP across the FSS
- Mechanisms involve noise injection (e.g. perturbing cells)
- Typically involve a privacy budget, e.g. ε
    - Larger $\varepsilon \Rightarrow$ less noise and less security

Disclosure Application: Differentially Private Tables

Differential Privacy (DP):

- Publication property that provides a provable confidentiality guarantee
- Increasing adoption of DP across the FSS
- Mechanisms involve noise injection (e.g. perturbing cells)
- Typically involve a privacy budget, e.g. ε
  - Larger $\varepsilon \Rightarrow$ less noise and less security

How to use estimated valuations:

- Allocate the privacy budget, ε, among publications, e.g.
  - For Table 1 and 2 use 0.361ε and 0.639ε.
  - More generally, for P potential publications
    1. Estimate valuations: $\widehat{s_1}, \ldots, \widehat{s_P}$
    2. For publication p use $\widehat{s_p}\varepsilon$ in the DP mechanism
- Maintains publication accuracy where most valued by data consumers

Conclusion

Summary:

- Presented a proof of concept of how to use estimated preferences over statistics to inform SDC methods
- Developed a model of consumer choice over publications and estimated it for CFOI using Google Analytics data
- Found significant heterogeneity in preferences over publication characteristics
- Though these estimated preferences are not pivotal for solving cell suppression problem, they are useful for allotting privacy budgets in formally private SDC methods

Next Steps:

- Expand the model to the random coefficients case (i.e. BLP) to allow for the inclusion of consumer demographics
- Explore the framework with other BLS data products

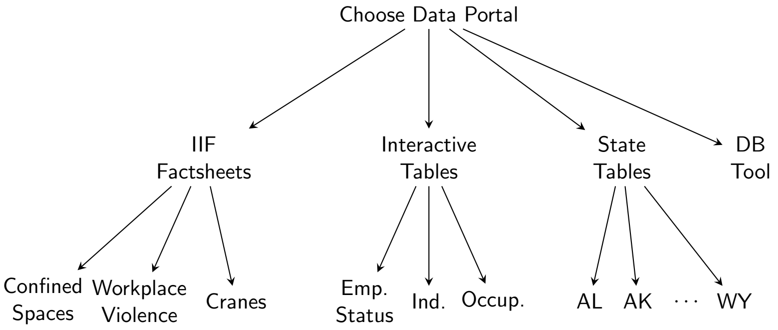
CONTACT INFORMATION

Thank You!

Elan Segarra
U.S. Bureau of Labor Statistics
Office of Compensation and Working Conditions
Segarra.Elan@BLS.gov

Nested Logit Model: Decision Tree

The nests (i.e. groups of related statistics/tables) line up with the likely route a user takes to access the statistic or table:



Can possibly generalize to other access mediums (e.g. twitter or API) or even to entire catalog of BLS data products.

Estimation

To put all of this in a more familiar form, if we define

$$y_{jt} = \ln q_{jt} - \ln q_{0t}$$

Then we have

$$y_{jt} = X_{jt}\beta + \sigma \ln s_{j|g} + \xi_{jt}$$

and since y_{jt}, X_{jt}, and $s_{j|g}$ are all observed we can use traditional regression methods to estimate β and σ.

Estimation

To put all of this in a more familiar form, if we define

$$y_{jt} = \ln q_{jt} - \ln q_{0t}$$

Then we have

$$y_{jt} = X_{jt}\beta + \sigma \ln s_{j|g} + \xi_{jt}$$

and since y_{jt}, X_{jt}, and $s_{j|g}$ are all observed we can use traditional regression methods to estimate β and σ.

Open question: Can we use OLS or is there endogeneity that would require something like IV?

- Reminder: ξ_{jt} are unobs. table characteristics
- Seems like obs. table characteristics (X_{jt}) are exogenously determined by BLS

Regression Results (Full)

	(1)	(2)
Employment Status	0.518**	0.518**
	(0.037)	(0.038)
Industry	-0.112**	-0.112**
	(0.030)	(0.033)
Occupation	-0.459**	-0.459**
	(0.034)	(0.036)
Gender	0.069	0.069
	(0.072)	(0.058)
Injury Event	-0.453**	-0.453**
	(0.025)	(0.029)
Age	-0.525**	-0.525**
	(0.042)	(0.038)
Geo. State	-0.350**	-0.350**
	(0.088)	(0.107)
Race/Ethnicity	-0.509**	-0.509**
	(0.034)	(0.034)
Document Source	-0.958**	-0.958**
	(0.054)	(0.061)
Constant	-3.308**	-3.308**
	(0.182)	(0.118)

** indicates significance at the 0.01 level.

	(1)	(2)
Type: Count	-0.016	-0.016
	(0.030)	(0.028)
Type: Percent	-0.482**	-0.482**
	(0.075)	(0.062)
Type: Rate	-0.179**	-0.179**
	(0.036)	(0.034)
Form: Bar Chart	0.337**	0.337**
	(0.051)	(0.060)
Form: Line Chart	0.660**	0.660**
	(0.058)	(0.051)
Multiple Ref. Years	-0.274**	-0.274**
	(0.021)	(0.025)
Inc. Current Ref. Year	1.170**	1.170**
	(0.059)	(0.080)
Nest Shares	0.358**	0.358**
	(0.029)	(0.037)
Mkt FE	Yes	Yes
Robust SE	No	Yes
N	5870	5870

** indicates significance at the 0.01 level.