Sources of Population Data on Sexual Orientation and Gender Identity (SOGI)

Sexual Orientation and Gender Identity Subcommittee, CDPH/LHJ Population Data Task Force[1]

Contact: David Crow (david.crow@cdph.ca.gov)

August 2024

Executive Summary

Objective: The Sexual Orientation and Gender Identity (SOGI) subcommittee of the California Department of Public Health/Local Health Jurisdictions' (CDPH/LHJ) Population Data Task force was charged with **identifying sources of population data on SOGI subpopulations** (e.g., transgender individuals, lesbians, gay men, etc.) to use as denominators in rate estimates.

Main Findings:

- No data source contains a population-level enumeration of members of either different sexual orientation groups or different gender identity groups (referred to collectively as SOGI data, groups, or subpopulations).
- Several read-to-use estimates exist of some SOGI subpopulations at the state and county levels, but these estimates have drawbacks and limitations:
    - AskCHIS, the online search tool of the California Health Interview Survey (CHIS)
    - The Williams Institute
- There are three probability sample surveys that can be used to estimate the size of SOGI subpopulations in California:
    - The California Health Interview Survey (CHIS) (UCLA)
    - The Household Pulse Survey (HPS) (Census Bureau)
    - The Behavioral Risk Factor Surveillance System (BRFSS) (CDC)
- Though CHIS is designed to be representative at the county level, estimates of SOGI subpopulations for some counties are not very precise—and, for the smallest counties, are not available at all.
- Estimates obtained through HPS and BRFSS are only stable (i.e., precise, having narrow confidence intervals) at the statewide level.
- It is possible to obtain more granular estimates, at the county level or lower, by combining the behavioral survey data or using model-based techniques (including "bottom-up" spatial modelling and small area estimation, SAE), or both.

[1] The authors of this report are, in alphabetical order, Angalar Chi, David Crow, Scott Fujimoto, Abera Galleta, Robbie Snyder. Michael Samuel gave invaluable editing suggestions.

- In accordance with the emerging consensus about best practices for measuring SOGI status, two of the three surveys discussed here, CHIS and HPS, ask about gender identity using the "two-step" method—that is, a question on sex at birth and another on current sex or gender, with a follow-up confirmation when these differ.
- However, CHIS and HPS omit response options that allow for transgender individuals to identify their current sex/gender (e.g., male-to-female, female-to-male, etc.).[2] This complicates breaking down transgender individuals by current gender; doing so relies on constructing identity from two or more variables in the survey (for example, comparing current sex to sex at birth) and making assumptions (e.g., about the distribution of transgender men and women in the population).

Main Data Sources for SOGI Population Estimates

After a thorough examination of possible data sources, the subcommittee found no data source with a population-level enumeration of members of different SOGI subgroups. There are, as far as we know, at least two sources for existing, ready-to-use estimates of SOGI populations at the state or county level that could serve as denominators for rate estimates. However, these estimates have limitations of which users should be aware. An alternative is to estimate the size of SOGI populations using modelling and data from the behavioral health surveys; as we describe in the section titled "Modelling CHIS Data," statistical modelling can be used for more precise, granular estimates. We describe the ready-to-use SOGI estimates and then turn to the behavioral surveys.

Ready-to-Use Estimates of Some SOGI Subpopulations

At least two sources exist for estimates of the size of SOGI subpopulations that are ready to use "off the shelf"—that is, they require no analysis or modelling. These estimates could be used as denominators in rate calculations. However, each has drawbacks that could limit their usefulness. The two sources are:

1) AskCHIS
2) The Williams Institute

AskCHIS

The UCLA Health Policy Center, which carries out the California Health Interview Survey (CHIS), provides an easy-to-use, online query tool, AskCHIS, for getting descriptive statistics (percentages and confidence intervals) for questions on the survey—including those on sexual orientation and gender identity. These estimates are available at the statewide and county geographical levels.

[2] CHIS does record a variable, "GENDIDEN", that breaks down transgender identity by current sex/gender (see the data dictionaries for "Constructed Variables"). However, CHIS does not make this variable available in the funder data files provided to CDPH; to conduct analyses using GENDIDEN, researchers must use CHIS's paid consulting services.

Pros:

- Easy to use and understand.
- Available for most California counties (six northern counties are grouped together).
- Query tool allows for pooling over several years and comparing between years and counties.

Cons:

- Estimates rounded to nearest thousand and may be too imprecise for some purposes.
- Gender identity doesn't distinguish between transgender men, transgender women, and non-binary individuals.

Williams Institute

The Williams Institute, in UCLA's Law School, is a leading research center "dedicated to conducting rigorous, independent research on sexual orientation and gender identity law and public policy" (from its website). The institute's 2022 report "How Many Adults and Youth Identify as Transgender in the United States?" contains state-by-state estimates of transgender individuals, broken down by race/ethnicity and age.[3]

Pros:

- Contains national and state-level estimates of transgender population.
- Estimates youth (13-17) as well as adult populations.
- National estimates distinguish between transgender men, transgender women, and "noncomforming" individuals.
- Provides helpful breakdowns by race/ethnicity and age.
- Detailed explanation of methodology used in calculations.

Cons:

- State-level estimates don't distinguish between transgender men, transgender women, and non-binary populations.
- No estimates more granular than state-level.
- Reports are updated irregularly; so far, only in 2015 and 2022.
- Rounded to nearest thousand.

Behavioral Health Surveys

In addition to these ready-to-use estimates, three behavioral health surveys exist that furnish data to allow for *estimates of* the size of different SOGI subpopulations:

[3] The institute also published state-level estimates of LGBT adults in a 2023 report, "Adult LGBT Population in the United States." However, the report does not distinguish among LGBT subpopulations, lumping them together to produce a single number. Thus, the report will not be useful for those researching a particular SOGI subpopulation.

1) the California Health Interview Survey (CHIS)
2) the Household Pulse Survey (HPS); and
3) the Behavioral Risk Factor Surveillance System.

These surveys interview large numbers of respondents; are recurring carried out regularly, which allows data to be pooled over several survey periods; have areal variables (i.e., for administrative subdivisions of the state) that, in theory, allow for estimates of SOGI populations at the substate level; adhere to high methodological standards and document their methods thoroughly; and conform, mostly, to the emerging consensus on best practices for survey questions on sexual orientation and gender identity.

California Health Interview Survey (CHIS)

CHIS is a large-scale behavioral health survey (~22,000 respondents per year) carried out in California annually since 2001 by the University of California, Los Angeles's (UCLA) Center for Health Policy Research. The questionnaire covers chronic health conditions, access to health care and insurance, mental and behavioral health topics, and risk factors. The survey is designed to be representative at the county level, though low counts of SOGI populations may make county-level estimates infeasible for smaller counties. CHIS has asked about sexual orientation since its inception in 2001 and about gender identity since 2014.

CDPH has a restricted access data set that has much more granularity than AskCHIS. CDPH researchers can request the restricted data from victoria.daher@cdph.ca.gov. (These data are not available to researchers working elsewhere.)

Pros

- **Large survey** (23-24k per year)
- **Representative at county level**
- **Highly granular geographic** variables, including regions (according to several different regionalization schemes, such as the Covered California Pricing Regions), urban/rural categorizations (again, according to several different schemes), county, ZIP Code, and, in the restricted datasets (to which CDPH has access), longitude, latitude, and census tract.[4]
- **Gender identity and sexual orientation largely conform to emerging consensus on best practices**, including the two-step question on gender identity with verification.
- **Thoroughly documented methodology**.

Cons

- **Small-*N* problem** (see section below on "Modelling SOGI Populations"). Though CHIS is a relatively large survey, LGBTQ+ identification is relatively infrequent; so, disaggregating the data by different geographical units, and cross-tabulating the data with covariates, quickly leads to small (or zero) cell sizes.
- CHIS **gender identity item does not distinguish between transgender men and transgender women**. **Estimating separate populations of transgender men and women**

[4] Analysis at the census tract requires

would, then, **rely on assumptions** that a respondent's self-reported sex corresponds to their post-transition (medical, social, etc.) identity; that the current gender identity is the opposite of the one assigned at birth; or about the distribution of transgender men and women in the general population (e.g., that this distribution mirrors that of men and women in the general population). All three of these assumptions are problematic—among other things, none appears to allow for the possibility of nonbinary identification—but may be preferable to not attempting to different between transgender men and women *at all*.[5] (See description below of the Sexually Transmitted Disease Control Branch's calculations for the statewide transgender population for an example of how such assumptions might be used.)

- **Estimation at the ZIP Code- or census tract-level may be challenging**, even with the modelling techniques described below.[6]

See this 2018 document "California Health Interview Survey Sexual Orientation and Gender Identity Working Group: Summary and Final Recommendations" and the report "CHIS Sexual Orientation and Gender Identity 2022 Cognitive Pretest Report: Findings from Waves I and II" for discussions of best practices on SOGI questions—including a cross-tabulation of gender identity with sex. CHIS recommends pooling several years of data for estimating gender identity.

Household Pulse Survey

In April 2020, the Census Bureau began collecting information on the sexual orientation and gender identity of respondents to its Household Pulse Survey (HPS). The HPS was originally intended to "measure household experiences during the coronavirus pandemic," but quickly expanded to include other social and economic topics. The U.S. Census Bureau uses a two-weeks on, two-weeks off collection and dissemination approach. The survey is administered online, and the data files are released twice per month. HPS began to ask about SOGI topics in Phase 3.2 (August 2021); as of April 2024, HPS is in Phase 4.2. Though the survey is national in scope, each edition has around 4,500 California respondents. The public use data files are available on the U.S. Census Bureau website; the Census Bureau does make restricted use files available, but the procedure for obtaining them is cumbersome.

Pros:

- **Continuously administered survey** on a two-weeks-on, two weeks off schedule, allowing for **possible trend analysis**
- Public use data files are **released twice per month**
- Approximately **4,500 – 5,000 respondents per survey for California**

[5] The same reasoning applies to the sexual orientation variable, which lumps together gay men and lesbian women. Here, too, we would take the additional step of cross-tabbing sexual orientation with self-reported sex or gender to create estimates for gay men and lesbian women.

[6] In addition, CHIS does not provide geo-identified census tract data in the funder files. Census tract-level analyses require a written request and IRB approval (CHIS 2022 Source Adult Data Dictionary, Section 5, "Restricted Variables").

- **Large N overall** (~126k; 28 editions with SOGI questions × 4,500 Californians per edition) due to being continuously administered
- Restricted use datasets allow for **highly granular analysis**; the "MAFID" (Master Address File ID) variable has respondent's addresses, which can be cross-referenced with very granular geographic levels, including county, ZIP Codes, city, census tract, and other geographic information. However, the restricted use data is difficult to obtain.
- HPS uses the **two-step with confirmation method for gender identity**.
- **Thorough descriptions of methodology**, sampling, and data collection.

Cons:

- **No substate analysis** possible with publicly available data.
- **Restricted use data are hard to get**: researchers must submit an application to the nearest Federal Statistical Research Data Center (FSDRC) with a project proposal and obtain a federal security clearance.
- Response options to the **gender identity question do not distinguish between transgender men and transgender women**. Thus, estimating populations of transgender men and women would rely on relatively strong assumptions (i.e., assumptions that are plausible but hold true only when specific conditions are met).

Behavioral Risk Factor Surveillance System (BRFSS)

The Behavioral Risk Factor Surveillance System (BRFSS) is our country's premier behavioral and attitudinal survey on public health topics. The Centers for Disease Control and Prevention (CDC) has carried out BRFSS since 1984, when it began in 15 states; today, it covers all 50 states in addition to three U.S. territories. The questionnaire covers health-related risk behaviors, chronic health conditions, and use of preventive services, among other topics. BRFSS conducts around 400,000 interviews per year, making it the largest behavioral health survey in the world. These include around 9,000 Californians per survey (although the COVID pandemic resulted in fewer interviews during 2021-2022). BRFSS began to inquire about sexual orientation in 2006 and gender identity, in 2016. CDPH administers California's BRFSS survey and posts the Data User Agreement to obtain the Public Use File on its website.

Pros:

- BRFSS response set to **gender identity question *does* explicitly distinguish between transgender men and women**. No additional assumptions would be necessary to estimate populations of transgender men and women.
- Approximately **5,000 – 9,000 respondents per year** for California
- **Large N overall** (~150k for sexual orientation, 22 editions × 7,000 Californians per edition; and ~56,000, 8 editions × 7,000 Californians per edition).
- W**ell-documented methodology**.

Cons:

- BRFSS departs from the emerging consensus on practices in that it does *not* use the **two-step with confirmation method for gender identity**—resulting, in theory, in measurement error for this variable.

- Geocoded areal variables include only region, county, and ZIP Code
- Small *N* problem.

See the online document "Years Survey Included Sexual and Gender Minority (SGM)-related Questions" for further discussions of BRFSS SOGI questions and additional resources.

Modelling SOGI Populations

To get useful substate population estimates, statistical models, perhaps combining the CHIS, HPS, and BRFSS in some fashion, are probably needed. Two widely used approaches are **"bottom-up" spatial modelling** and **small area estimation** (SAE). Fortunately, these are theoretically and practically well-developed approaches, already implemented in several statistical software packages. Unfortunately, both are complicated and require a fairly high degree of technical sophistication to understand and use.

Bottom-Up Approaches and Spatial Modelling

The essence of "bottom-up" modelling approaches to estimating populations at granular levels lies in taking a fully enumerated set of geographic locations, arranged into a "polygon" or a "grid" ("lattice") structure, and using estimates from locations there *is* information on to "fill in" estimates for the locations with missing data. Bottom-up modelling accommodates data from, potentially, many different sources (outdated or partial census data, local survey estimates, registry records from decentralized health systems, "ancillary" geospatial covariates, like road network data or building "footprints," etc.). The procedure is to first get, from these data sources, numbers for as many locations in the grid as possible—so-called "spatially disaggregated population estimates." then "fill in" estimates for the *missing* locations, in between those locations where there *are* data, by relying on assumptions about how the geographical proximity of one community to another influences shared characteristics of both.

One such assumption is, following Tobler's law of geography ("everything is related to everything else, but near things are more related than distant things"), that **contiguous geographical units**—say, counties—**are more alike than are counties farther apart from one another**. A second assumption frequently used in bottom-up models is that **changes in quantities from one county to the next are gradual and continuous**, not abrupt.

Spatial Analysis in Epidemiology, *Spatio and Spatio-Temporal Bayesian Models with R-INLA*, and the online book *An Introduction to Spatial Econometrics* are accessible theoretical introductions, and *Applied Spatial Data Analysis in R* mixes theory with applications. The statistical computing program "R" has several packages for "bottom-up" spatial analysis, including `INLA` and `btb` (for spatial models), and `automap` (for kriging). Several free books and tutorials are available online, including Gómez-Rubio's *Bayesian inference with INLA*, Moraga's *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*, and the tutorial "Intro to Modelling Using INLA." SAS implements spatial smoothing in the procedures ("procs") KRIGE2D, VARIOGRAM, and

KDE. The tutorials "Using SAS for Spatial Analysis," "Everything in its Place," and "Spatial Smoothing Using SAS" (slides 21-39) provide practical guidance for spatial modelling in SAS.

Small Area Estimation

Small area estimation (SAE) is a suite of **model-based statistical techniques to estimate quantities or populations in "sub-domains"** (such as geographical areas) **too small to allow survey data alone to yield sufficiently precise estimates**. SAE can even provide estimates for sub-domains that are missing altogether from the survey. The key intuition is "**exploiting similarity**": individuals with a given demographic, behavioral, and attitudinal profile living in a sub-domain—we use counties as an example, though SAE is often used for much smaller geographies (such as census tracts)—with few (or no) survey respondents are like individuals with similar demographic and attitudinal characteristics in *other* counties, and counties with a given socioeconomic profile are like *other* counties with similar profiles.

In order to estimate a quantity for a given county with SAE, models combine information on survey respondents from *that* county with information on similar individuals in *other* counties, information on similar counties, and sometimes "auxiliary" information obtained from sources outside the survey, such as census data. In contrast to spatial modelling, where the proximity of one county to another furnishes the key assumptions, in SAE the distance of counties from one another is less important. SAE combines information on individuals in different counties regardless of whether these counties are close or far from each other. In fact, researchers don't need to know anything about where the counties are located in space to carry out SAE—though if they *do* know where these counties are relative to one another, they can use that information to improve small area estimates. In practice, this means using a statistical model known as a multi-level model (because they include predictors from different geographical levels, such as individuals nested within counties).

General theoretical references for SAE include Rao and Molina's highly technical *Small Area Estimation* and Longford's more accessible *Missing Data and Small Area Estimation*. The statistical package R implements SAE in the packages `sae` and `emdi`, and multi-level models in `lme4`; the Internet book *Data Science with R: A Resource Compendium* (Chapter 23) provides additional references, including links to tutorials. The SAS procedures MIXED and GLIMMIX implement multi-level models, a by-product of which is area estimates. Mukhopadhyay and McDowell's paper *Small Area Estimation for Survey Data Analysis Using SAS* and Hindmarsh's Ph.D. Dissertation *Small area estimation for health surveys* give practical guidance, including SAS code, for implementing small area models.

Bottom-up spatial modelling and SAE can complement one another.

Proof of Concept: Modelling CHIS Data

To obtain substate estimates using survey data, the subcommittee ("we") carried out two simple, preliminary analyses using CHIS survey data. The first was a simple **weighted cross-tabulation of sexual orientation by county** (executed in SAS EGP 8.2 using PROC SURVEYFREQ), combining data

from 2014 to 2022 (the years for which data are available for both sexual orientation and gender identity). This is a "direct" estimate—that is, based only on the unmodelled survey data.

Table 1 shows the results for the three most common sexual orientation categories, "Straight/Heterosexual," "Gay/Lesbian," and "Bisexual." (Note that both males and females are included in all these categories.) The three columns under each SO category are that category's estimated proportion of the county population, the raw frequency (i.e., number of respondents in each category), and the survey's estimate of the total population category, after applying survey weights (technically known as "expansion factors"). This total could, in theory, be the denominator for SOGI rate calculations.

The cross-tabulation **doesn't appear to be very suitable for estimating SOGI populations for all California counties**. For one thing, the frequencies in many counties are so low that UCLA suppresses the data, in accordance with its data de-identification guidelines. Some of the estimated proportions seem unrealistically high, especially in the "bisexual" category: 9.8% of adults in Del Norte County are bisexual? And 8.5% in San Benito, plus 8.3% in Butte?

Small Area Estimation Using CHIS Data

The second preliminary analysis comprises three statistical models, two of which were **simple small area estimation (SAE) models**, that we compare to one another (**Table 2**). Here, we took estimating the proportion of gay males by county as a test case. This is a good test case because it involves slicing and dicing the data more finely than we did in Table 1: we are stratifying a main variable, sexual orientation, by a second variable, sex. As with the cross-tabulation in Table 1, we pooled CHIS data across the years 2014-2022.

The **first, baseline model was a direct estimate**—again, using just the individuals in the survey, without any modelling or spatial smoothing—**obtained by a logistic regression (PROC SURVEYLOGISTIC in SAS EGP 8.2) of the constructed variable "gay male" on each of the counties as predictors**. This is a convenient way, and equivalent to, making a table of the percentage of gay males in each county. However, since the model outputs are coefficients and proportions, and not counts, the model is not subject to UCLA's data suppression criteria for low counts.

The four columns under each model are (1) the estimated proportion of gay males in each county; (2) the total number of gay males, to be used directly in the denominator (obtained by multiplying the adult population in each county by the proportion estimated in the first column); and the (3) lower and (4) upper bounds of the confidence interval (a plausible range in which the true population percentage lies).

The **second model** (again based on pooled 2014-2022 data) **is the simplest SAE model there is: a "random intercepts" model**, estimated with PROC GLIMMIX in SAS. Here, as in the first, logistic model, counties are the only predictors, but the model uses a weighted average of the within-county, direct estimate and the global, statewide mean. This "smooths" the county estimates

toward the statewide mean, especially for small counties. This second model is a significant improvement over the first.

Finally, the **third model is based on the second**, **but with year added as a linear predictor**. This model estimates a baseline (random intercept) for each county, but further posits that the proportion of gay males increases statewide at the same linear rate. The estimates are for the last year in the survey, 2022. The assumption of the same linear change for all counties oversimplifies reality somewhat (as all models do), but improves on the second model by accounting for change over time.

We could improve upon these models with more sophisticated ones that use more predictors, incorporate the insights of the "bottom-up" modelling approach, and add HPS and BRFSS. More sophisticated models may allow us to drill down to finer levels of granularity, such as ZIP Codes, and perhaps add more stratifiers (race/ethnicity, age category, etc.).

Conclusion

The State and local **governments need estimates of the numbers of LGBTQ+ individuals**—not only for denominators to calculate rate estimates, but also to understand the size of these populations and anticipate the scale of their unique health and social needs. Unfortunately, there is no source of data with a population-level enumeration of SOGI subgroups. There are a few existing estimates of these populations that are ready to use in denominators, but each set of estimates has drawbacks that may limit their usefulness. The Census Bureau may add SOGI questions to the 2030 decennial census, but **until a population-level enumeration of the LBGTQ+ population is available, extrapolation of population estimates from surveys is our best option.**

A paramount important factor in determining which survey(s) to use for a particular task is the response categories provided by the surveys themselves. One of the biggest differences between the surveys is that the BRFSS gender identity question breaks down transgender identity by sex (male-to-female vs. female-to-male) while CHIS and HPS do not.

Table 1. Sexual Orientation for Adults by County
(California Health Interview Survey, CHIS, Pooled 2014-2022 Data)

	Straight/Heterosexual					Gay/Lesbian					Bisexual				
County	%	LB	UB	Freq.	Total	%	LB	UB	Freq.	Total	%	LB	UB	Freq.	Total
Alameda	90.7%	89.5%	91.9%	5,074	1,099,178	3.4%	2.7%	4.0%	84	40,948	4.4%	3.5%	5.4%	260	53,524
Alpine	98.5%	95.6%	100.0%	46	992	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.
Amador	94.9%	91.7%	98.1%	322	26,910	2.7%	0.2%	5.2%	Supp.	757	Supp.	Supp.	Supp.	Supp.	Supp.
Butte	87.7%	85.0%	90.5%	1,986	142,247	2.0%	0.8%	3.1%	38	3,185	8.3%	5.9%	10.7%	94	13,458
Calaveras	95.4%	90.1%	100.0%	571	26,786	Supp.	Supp.	Supp.	Supp.	Supp.	2.4%	0.0%	7.6%	10	678
Colusa	94.3%	90.8%	97.9%	302	12,857	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.
Contra Costa	91.4%	89.2%	93.5%	3,662	780,168	2.7%	0.9%	4.4%	46	22,657	3.8%	2.7%	5.0%	137	32,805
Del Norte	85.6%	76.5%	94.8%	279	18,099	Supp.	Supp.	Supp.	Supp.	Supp.	9.8%	0.8%	18.7%	15	2,064
El Dorado	94.2%	92.3%	96.1%	1,925	133,486	1.9%	0.8%	3.0%	17	2,669	3.4%	2.0%	4.9%	46	4,859
Fresno	92.2%	90.7%	93.8%	3,067	629,055	2.7%	1.8%	3.5%	45	18,101	3.0%	1.9%	4.1%	85	20,557
Glenn	92.8%	86.7%	98.8%	420	18,216	Supp.	Supp.	Supp.	Supp.	Supp.	3.8%	0.0%	9.1%	6	741
Humboldt	88.3%	85.9%	90.7%	2,049	86,003	3.1%	1.8%	4.4%	36	3,019	7.1%	5.1%	9.0%	129	6,886
Imperial	92.4%	90.2%	94.6%	2,587	107,618	1.6%	0.9%	2.4%	37	1,886	3.4%	1.9%	4.9%	59	3,947
Inyo	96.8%	93.5%	100.0%	246	16,430	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.
Kern	92.8%	91.2%	94.4%	2,637	541,895	2.5%	1.6%	3.4%	40	14,620	3.6%	2.3%	4.8%	68	20,780
Kings	91.6%	89.3%	93.9%	2,049	86,660	2.4%	0.9%	3.9%	45	2,252	3.7%	2.1%	5.3%	50	3,491
Lake	92.4%	90.6%	94.3%	1,933	44,126	3.1%	2.1%	4.2%	Supp.	1,487	3.4%	2.0%	4.8%	64	1,624
Lassen	94.9%	90.7%	99.1%	258	16,931	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.
Los Angeles	91.0%	90.5%	91.6%	31,218	6,770,744	3.3%	3.0%	3.6%	603	247,751	3.8%	3.4%	4.2%	1,095	279,088
Madera	94.3%	92.2%	96.4%	2,063	92,523	1.6%	0.7%	2.5%	29	1,532	2.2%	1.2%	3.3%	46	2,168
Marin	93.3%	91.5%	95.0%	2,521	174,232	2.1%	1.2%	3.0%	36	3,949	3.2%	1.8%	4.6%	77	6,000
Mariposa	95.3%	90.0%	100.0%	186	12,100	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.
Mendocino	91.4%	88.4%	94.4%	1,955	60,541	3.2%	1.7%	4.7%	34	2,106	4.0%	1.3%	6.8%	78	2,676
Merced	91.8%	89.5%	94.1%	1,981	165,437	2.2%	1.2%	3.1%	32	3,912	4.0%	2.3%	5.7%	58	7,226
Modoc	96.0%	90.3%	100.0%	137	6,195	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.
Mono	92.1%	83.4%	100.0%	142	7,520	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.
Monterey	92.9%	91.0%	94.8%	1,906	270,966	1.4%	0.4%	2.3%	40	4,036	3.2%	1.9%	4.6%	53	9,421
Napa	92.8%	89.8%	95.9%	2,004	90,123	Supp.	Supp.	Supp.	22	Supp.	4.0%	1.2%	6.7%	46	3,844
Nevada	94.5%	92.7%	96.2%	1,938	70,410	1.7%	0.7%	2.6%	Supp.	1,234	3.3%	1.9%	4.6%	61	2,455
Orange	92.7%	91.8%	93.6%	9,212	2,170,557	2.5%	1.9%	3.0%	143	57,688	3.4%	2.8%	4.1%	212	80,674

County	%	LB	UB	Freq.	Total	%	LB	UB	Freq.	Total	%	LB	UB	Freq.	Total
Placer	93.1%	91.2%	95.1%	1,974	270,091	1.8%	0.6%	3.1%	20	5,318	4.1%	2.5%	5.7%	52	11,936
Plumas	97.5%	96.1%	98.9%	233	12,933	Supp.*	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.
Riverside	92.0%	91.0%	93.0%	6,838	1,549,012	3.8%	3.1%	4.4%	113	63,585	2.7%	2.1%	3.3%	192	45,306
Sacramento	90.8%	89.4%	92.2%	4,937	997,239	2.9%	2.1%	3.7%	82	31,705	4.6%	3.7%	5.5%	197	50,733
San Benito	88.4%	82.5%	94.3%	2,016	36,718	1.5%	0.6%	2.5%	29	643	8.5%	2.5%	14.4%	59	3,522
San Bernardino	92.9%	91.9%	93.9%	5,813	1,398,038	2.0%	1.4%	2.5%	106	29,740	3.2%	2.5%	3.9%	159	48,188
San Diego	91.2%	90.5%	92.0%	16,963	2,124,209	3.2%	2.7%	3.7%	266	74,681	4.1%	3.6%	4.7%	537	96,402
San Francisco	83.2%	80.9%	85.5%	3,410	572,356	9.8%	8.1%	11.6%	76	67,768	5.0%	3.6%	6.5%	189	34,720
San Joaquin	91.6%	89.8%	93.4%	2,102	489,808	2.5%	1.4%	3.6%	25	13,340	4.2%	2.8%	5.5%	67	22,309
San Luis Obispo	95.1%	93.5%	96.7%	1,993	181,417	1.7%	0.7%	2.6%	14	3,155	2.9%	1.6%	4.2%	49	5,508
San Mateo	93.8%	92.3%	95.3%	2,474	542,994	2.5%	1.6%	3.4%	33	14,682	2.2%	1.3%	3.2%	67	12,956
Santa Barbara	92.6%	90.7%	94.5%	1,938	299,787	2.6%	1.5%	3.7%	26	8,389	3.3%	2.2%	4.4%	56	10,633
Santa Clara	93.1%	91.9%	94.4%	6,006	1,307,595	2.2%	1.7%	2.8%	110	31,251	2.8%	1.9%	3.8%	161	39,787
Santa Cruz	87.9%	85.1%	90.7%	1,923	173,599	3.3%	1.9%	4.7%	31	6,546	7.3%	5.1%	9.5%	102	14,453
Shasta	92.0%	89.8%	94.2%	2,003	118,430	1.9%	1.0%	2.9%	37	2,507	5.0%	3.1%	6.9%	64	6,414
Sierra	91.3%	74.3%	100.0%	47	1,061	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.	Supp.
Siskiyou	95.1%	92.9%	97.3%	771	29,686	Supp.	Supp.	Supp.	Supp.	Supp.	1.9%	0.7%	3.1%	20	607
Solano	91.4%	88.7%	94.1%	2,065	280,694	2.8%	1.4%	4.2%	24	8,607	5.1%	2.8%	7.5%	69	15,808
Sonoma	91.5%	89.4%	93.5%	2,260	340,562	3.0%	2.1%	3.9%	26	11,337	4.3%	2.7%	6.0%	83	16,165
Stanislaus	90.9%	88.5%	93.3%	2,023	350,718	2.4%	1.3%	3.5%	34	9,256	4.8%	2.7%	6.9%	63	18,651
Sutter	92.6%	90.4%	94.8%	2,130	65,848	2.2%	1.3%	3.2%	38	1,590	3.1%	1.6%	4.6%	44	2,198
Tehama	94.8%	93.0%	96.7%	1,011	45,967	1.1%	0.2%	1.9%	18	525	2.8%	1.4%	4.2%	17	1,375
Trinity	91.2%	77.0%	100.0%	203	8,779	Supp.	Supp.	Supp.	Supp.	Supp.	7.7%	0.0%	21.9%	6	741
Tulare	94.7%	93.1%	96.2%	2,090	295,595	1.4%	0.6%	2.1%	30	4,292	2.5%	1.3%	3.8%	51	7,846
Tuolumne	96.5%	94.7%	98.4%	633	39,407	Supp.	Supp.	Supp.	Supp.	Supp.	1.6%	0.3%	2.8%	12	638
Ventura	93.1%	91.5%	94.7%	2,615	575,376	2.3%	1.3%	3.2%	30	13,992	3.4%	2.2%	4.7%	73	21,290
Yolo	88.8%	85.4%	92.3%	2,006	142,508	4.1%	1.5%	6.7%	34	6,528	5.6%	3.3%	7.9%	83	8,988
Yuba	92.2%	90.0%	94.3%	1,988	50,736	2.1%	1.2%	3.0%	25	1,151	4.6%	2.9%	6.3%	77	2,552

% = proportion of county estimated directly from survey (with no spatial smoothing or modelling)

LB/UB = lower and upper bounds of 95% confidence interval for proportion

Freq. = number of survey respondents in the SO category, summed over 2014-2022 (9 yrs.)

Total = number of Californian adults in SO category, as estimated by CHIS survey weights (technically, expansion factors)

* Supp. = suppressed; the survey frequency falls below the threshold for reporting, per UCLA data de-identification guidelines

Table 2. Proportion of Gay Men by County (with 95% Confidence Interval)
(California Health Interview Survey, CHIS)

	Model 1: Direct Estimate (2014-2022)*				Model 2: Multi-level, Random Intercepts (2014-2022)**				Model 3: Random Intercepts w/ "Year" as Predictor (2022 Only)†			
County	%	Denom‡	LB	UB	%	Denom	LB	UB	%	Denom	LB	UB
Alameda	2.0%	26,625	1.5%	2.6%	2.5%	33,558	2.1%	2.9%	3.3%	42,948	2.8%	3.8%
Alpine	1.0%	10	0.1%	8.9%	1.7%	17	0.7%	4.2%	2.2%	22	0.9%	5.4%
Amador	0.9%	305	0.3%	2.3%	1.5%	519	0.8%	2.9%	2.0%	679	1.1%	3.9%
Butte	1.4%	2,327	0.6%	3.3%	0.9%	1,503	0.6%	1.3%	1.2%	1,972	0.8%	1.8%
Calaveras	1.9%	692	1.0%	3.5%	1.0%	370	0.5%	1.9%	1.4%	526	0.8%	2.7%
Colusa	1.0%	161	0.1%	8.3%	1.2%	196	0.6%	2.5%	1.7%	260	0.8%	3.4%
Contra Costa	1.4%	12,212	0.9%	2.1%	1.4%	12,717	1.1%	1.8%	1.9%	16,514	1.4%	2.4%
Del Norte	0.8%	174	0.3%	2.1%	0.8%	182	0.4%	1.8%	1.1%	236	0.5%	2.4%
El Dorado	0.7%	1,114	0.3%	1.7%	0.7%	1,122	0.5%	1.2%	1.0%	1,459	0.6%	1.5%
Fresno	1.6%	11,413	1.0%	2.4%	1.4%	10,163	1.1%	1.8%	1.8%	13,243	1.4%	2.4%
Glenn	0.3%	66	0.1%	1.4%	0.9%	199	0.5%	1.9%	1.2%	257	0.6%	2.4%
Humboldt	2.5%	2,799	1.5%	4.2%	1.7%	1,830	1.2%	2.3%	2.2%	2,382	1.6%	3.0%
Imperial	1.1%	1,474	0.6%	2.0%	0.8%	1,013	0.5%	1.1%	1.0%	1,354	0.7%	1.5%
Inyo	0.2%	24	0.0%	5.0%	1.0%	157	0.5%	2.2%	1.4%	202	0.6%	2.9%
Kern	1.3%	8,453	0.8%	2.1%	1.1%	6,887	0.8%	1.5%	1.4%	8,986	1.0%	2.0%
Kings	1.2%	1,295	0.4%	3.1%	1.2%	1,379	0.9%	1.7%	1.6%	1,798	1.1%	2.3%
Lake	1.0%	541	0.4%	2.5%	1.0%	566	0.7%	1.5%	1.4%	732	0.9%	2.0%
Lassen	1.5%	421	0.4%	5.7%	1.2%	319	0.6%	2.5%	1.6%	417	0.7%	3.2%
Los Angeles	2.4%	187,132	2.1%	2.6%	2.6%	206,962	2.4%	2.8%	3.4%	265,088	3.1%	3.7%
Madera	1.1%	1,218	0.5%	2.4%	0.9%	1,047	0.6%	1.4%	1.2%	1,362	0.8%	1.8%
Marin	1.3%	2,799	0.8%	2.3%	1.5%	3,051	1.1%	2.0%	2.0%	4,152	1.5%	2.7%
Mariposa	0.2%	26	0.0%	0.8%	1.0%	138	0.4%	2.2%	1.3%	180	0.6%	2.9%
Mendocino	1.6%	1,199	0.7%	3.8%	1.2%	890	0.9%	1.7%	1.6%	1,163	1.1%	2.3%
Merced	1.5%	3,066	0.9%	2.6%	1.1%	2,218	0.8%	1.6%	1.5%	2,887	1.0%	2.1%
Modoc	0.0%	-	0.0%	100.0%	0.9%	61	0.4%	2.1%	1.2%	79	0.5%	2.8%
Mono	1.5%	162	0.3%	6.7%	1.1%	112	0.4%	2.5%	1.4%	143	0.6%	3.2%
Monterey	0.9%	2,911	0.4%	2.1%	1.2%	3,893	0.8%	1.7%	1.6%	5,048	1.1%	2.3%
Napa	0.9%	1,044	0.5%	1.6%	1.4%	1,587	1.0%	2.0%	1.9%	2,062	1.4%	2.6%
Nevada	0.7%	627	0.3%	1.8%	0.8%	683	0.5%	1.3%	1.1%	888	0.7%	1.7%
Orange	1.7%	43,961	1.3%	2.3%	1.3%	33,965	1.1%	1.6%	1.8%	43,672	1.5%	2.1%
Placer	0.9%	2,949	0.4%	2.4%	0.9%	2,878	0.6%	1.4%	1.2%	3,724	0.8%	1.8%
Plumas	0.3%	46	0.0%	100.0%	0.9%	147	0.4%	2.0%	1.2%	193	0.5%	2.7%
Riverside	2.8%	51,615	2.3%	3.4%	4.1%	74,651	3.7%	4.6%	5.4%	96,661	4.8%	6.0%
Sacramento	1.9%	22,711	1.3%	2.7%	2.1%	25,512	1.8%	2.5%	2.8%	33,123	2.3%	3.3%
San Benito	1.0%	470	0.5%	1.9%	1.0%	469	0.7%	1.4%	1.3%	610	0.9%	1.9%

County	%	Denom‡	LB	UB	%	Denom	LB	UB	%	Denom	LB	UB
San Bernardino	1.0%	16,368	0.7%	1.4%	1.2%	19,918	1.0%	1.5%	1.6%	25,656	1.3%	2.0%
San Diego	2.2%	58,098	1.8%	2.7%	2.3%	60,486	2.1%	2.5%	3.0%	77,634	2.7%	3.3%
San Francisco	8.5%	64,537	7.0%	10.3%	9.6%	72,968	8.7%	10.5%	12.3%	92,124	11.1%	13.6%
San Joaquin	1.4%	7,846	0.6%	2.9%	0.9%	4,976	0.6%	1.3%	1.1%	6,429	0.8%	1.7%
San Luis Obispo	0.7%	1,533	0.2%	1.9%	1.0%	2,209	0.6%	1.4%	1.3%	2,867	0.8%	1.9%
San Mateo	1.6%	9,732	1.0%	2.6%	1.6%	9,882	1.2%	2.2%	2.1%	12,830	1.6%	2.8%
Santa Barbara	1.6%	5,770	1.0%	2.8%	1.5%	5,291	1.1%	2.1%	2.0%	6,822	1.4%	2.7%
Santa Clara	1.5%	22,267	1.1%	1.9%	1.5%	23,168	1.2%	1.8%	2.0%	29,989	1.6%	2.4%
Santa Cruz	1.4%	3,175	0.8%	2.5%	1.7%	3,768	1.3%	2.3%	2.3%	4,893	1.7%	3.1%
Shasta	1.1%	1,617	0.6%	2.0%	1.0%	1,459	0.7%	1.5%	1.3%	1,902	0.9%	2.0%
Sierra	0.0%	-	0.0%	100.0%	1.1%	29	0.4%	2.8%	1.4%	38	0.6%	3.6%
Siskiyou	1.0%	367	0.3%	4.1%	1.0%	346	0.5%	1.7%	1.4%	477	0.8%	2.4%
Solano	1.7%	6,032	0.8%	3.5%	1.7%	5,904	1.2%	2.3%	2.2%	7,666	1.6%	3.0%
Sonoma	1.7%	6,681	1.2%	2.5%	2.3%	8,988	1.8%	2.9%	3.1%	12,099	2.4%	4.0%
Stanislaus	1.8%	7,166	1.0%	3.0%	1.2%	4,922	0.8%	1.7%	1.6%	6,436	1.1%	2.3%
Sutter	0.7%	538	0.3%	1.5%	0.9%	641	0.6%	1.3%	1.1%	843	0.8%	1.7%
Tehama	0.6%	305	0.2%	1.7%	0.8%	404	0.5%	1.4%	1.1%	527	0.6%	1.9%
Trinity	0.0%	-	0.0%	100.0%	0.9%	127	0.4%	2.1%	1.3%	165	0.6%	2.8%
Tulare	1.0%	3,418	0.5%	1.9%	1.0%	3,286	0.7%	1.5%	1.3%	4,278	0.9%	1.9%
Tuolumne	0.6%	257	0.1%	3.2%	0.8%	384	0.4%	1.6%	1.2%	526	0.6%	2.2%
Ventura	1.9%	12,272	1.1%	3.2%	1.1%	7,439	0.8%	1.6%	1.5%	9,690	1.1%	2.1%
Yolo	2.6%	4,375	0.9%	7.3%	1.4%	2,403	1.0%	2.0%	1.8%	3,116	1.3%	2.6%
Yuba	1.0%	606	0.6%	1.9%	1.1%	654	0.8%	1.6%	1.5%	857	1.0%	2.1%

* Direct estimate = estimate based only on the respondents in that county. This estimate pools 2014-2022 data for Models 1 & 2.

** Random intercepts = only counties, with no predictors, are used in the model. The intercept (i.e., estimate of the county proportion) "shrinks," or smooths, estimated proportions for small counties toward the statewide mean.

† Same as Model 2, but with "year" added as a predictor to the model. This allows us to account for change over time. Here, we take only estimates for 2022 (the last year in the data) on the assumption that estimates from more recent years are closer to the true current population value than older years. Note that, since the proportion of gay men grew over time, estimates from the last year alone are higher than for the previous two models, which average over all years.

‡ Denominator = these are estimated population totals that could be used as denominators. They are the total county adult population (not shown) multiplied by estimated proportion (first column under each model, "%"). To express the 95% confidence intervals (LB=lower bound, UB=upper bound) in population terms, we would multiply them by the total population