Summary of the results of Experiment 5, State and National tables.

1 Count tables

Tables PH2, PH3, PH4, PH6, PH7

The noisy measurements, Z, are draws from a discrete Gaussian distribution with unknown mean, Y, and known variance D, so that

$$Z = Y + \varepsilon,$$

where $\varepsilon \sim DG(Y, D)$. Our goal is estimation of the true Census count, Y. For simplicity, in our modeling work, we replace the discrete Gaussian distribution with a continuous Gaussian distribution.

Because Y is a count, we have the logical constraint that $Y \geq 0$. We can incorporate this information into the model through use of a prior distribution. We use the improper prior

$$g(Y) \propto I\left(Y \geq 0\right).$$

This results in the posterior distribution

$$\pi\left(Y \mid Z\right) \propto \exp\left\{-\frac{1}{2D}\left(Y - Z\right)^2\right\} I\left(Y \geq 0\right),$$

which is a truncated normal distribution. The mean of this distribution has a closed form representation, but the quantiles need to be estimated numerically.

The advantages of this methodology are

- Simplicity
- Speed
- Does not require additional auxiliary information
- No possibility of model misspecification
- Produces non-negative estimates
- Gives accurate interval estimates

The disadvantages of this methodology are

- Does not make use of auxiliary information
- Predictions are not meaningfully different from noisy measurements when the true counts are large

	min	max	p_neg	RMSE	COV	LEN	RMSE_s	COV_s	LEN_s
PH2 NM	1478	36320973	0.00	158.02	90.95	399.97	107.74	88.24	399.97
PH2 PRED	1478	36320972	0.00	157.90	91.10	402.95	107.62	88.24	399.66
PH3 NM	-64	9033149	0.63	96.61	89.80	56.56	14.16	90.59	40.00
PH3 PRED	4	9033149	0.00	96.52	91.35	55.97	14.40	94.12	38.23
PH4 NM	-281	29788110	0.20	139.95	90.72	399.97	150.64	85.71	399.97
PH4 PRED	48	29788113	0.00	139.36	91.18	402.17	139.21	88.31	398.73
PH6 NM	158	7830063	0.00	203.56	88.42	399.96	125.88	85.19	399.96
PH6 PRED	179	7830059	0.00	203.57	88.70	401.13	125.16	85.19	399.22
PH7 NM	-38	36320845	0.20	92.28	89.56	135.99	45.72	91.18	135.99
PH7 PRED	23	36320844	0.00	92.25	89.95	137.83	44.46	91.18	135.24

2 Ratio tables

Tables PH1, PH5, PH8

The methodology for the ratio tables is similar to that for the count tables. We extend what was done for the count tables by modeling the numerator and the denominator of the ratios jointly. Let the subscripts *num* and *den* denote the numerator and denominator, respectively. We then assume

$$Z_{num} \sim N\left(Y_{num}, D_{num}\right)$$

and

$$Z_{den} \sim N\left(Y_{den}, D_{den}\right).$$

We also have the logical constrints

- $Y_{den} \geq 1$ (no areas with zero housing units)
- Either $Y_{num}/Y_{den} \geq 1$ if the universe is households (no vacant housing units) or $Y_{num}/Y_{den} \geq 2$ if the universe is families
- $Y_{num}/Y_{den} \leq \kappa$, where κ is the truncation level (taken here to be 10, based on the configuration file)

These constraints can be summarized with an appropriatey defined matrix $\boldsymbol{D}$ and vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, so that

$$\boldsymbol{a} \leq \boldsymbol{D}\boldsymbol{Y} \leq \boldsymbol{b},$$

where $\boldsymbol{Y} = (Y_{num}, Y_{den})^{\intercal}$. For example, for the household ratio tables, we would set

$$\boldsymbol{a} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \ \boldsymbol{b} = \begin{bmatrix} \infty \\ \infty \\ \infty \end{bmatrix}, \ \boldsymbol{D} = \begin{bmatrix} 0 & 1 \\ 1 & -1 \\ -1 & \kappa \end{bmatrix}$$

We use the improper prior distribution

$$g(\boldsymbol{Y}) \propto I\left(\boldsymbol{a} \leq \boldsymbol{D}\boldsymbol{Y} \leq \boldsymbol{b}\right),$$

which results in a posterior distribution

$$\pi\left(\boldsymbol{Y} \mid \boldsymbol{Z}\right) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{Y}-\boldsymbol{Z}\right)^{\intercal}\boldsymbol{D}\left(\boldsymbol{Y}-\boldsymbol{Z}\right)\right\} I\left(\boldsymbol{a} \leq \boldsymbol{D}\boldsymbol{Y} \leq \boldsymbol{b}\right),$$

which is a truncated multivariate normal distribution. We generate samples from this distribution using the `R` package `tmvmixnrom`.

The advantages of this methodology are

- Simplicity
- Speed
- Does not require additional auxiliary information
- Produces ratios that are bounded between 1 (or 2 for family tables) and the truncation level
- Gives accurate interval estimates for *all* ratios

The disadvantages of this methodology are

- Does not make use of auxiliary information
- Predictions are not meaningfully different from noisy measurements when both the true numerator and denominator are large
- Intervals are not symmetric around the point estimates (although this is not unique to this methodology)

	min	max	p_bad	RMSE	COV	LEN
PH1 NM	-3.67	11.92	0.78	0.51	88.50	0.45
PH1 PRED	0.17	6.30	0.00	0.24	90.20	0.26
PH5 NM	-5.17	17.21	0.72	0.69	87.52	1.18
PH5 PRED	0.50	6.20	0.00	0.20	89.28	0.34
PH8 NM	-26.25	16.12	0.26	1.31	89.93	11.30
PH8 PRED	1.44	6.07	0.00	0.23	90.65	0.25