Statistical Theory for Data Science

STA2212H S LEC9101

Week 2

January 13 2026

Today

1. Recap
2. Profile likelihood and likelihood ratio tests
3. Bayesian inference
4. Statistics in the News
5. HW Questions

Upcoming: Toronto Data Workshop Link

Wednesday 14 January 2026, noon (EST) on Zoom
Amy Mann, University of Oxford
“Measuring the quality of mortality data in high-income settings”

Recap Week 1

- likelihood definitions: $\ell(\theta)$ $\widehat{\theta}$ $U(\theta)$ $j(\widehat{\theta})$ $I_n(\theta)$
- maximum likelihood estimator minimizes KL-divergence from the true model to the data
- maximum likelihood estimator is consistent

- maximum likelihood estimator is asymptotically normal

see Likelihood Cheat Sheet V2

Recap Week 1

- likelihood definitions: $\ell(\theta)$ $\widehat{\theta}$ $U(\theta)$ $j(\widehat{\theta})$ $I_n(\theta)$
- maximum likelihood estimator minimizes KL-divergence from the true model to the data
- maximum likelihood estimator is consistent

- maximum likelihood estimator is asymptotically normal

see Likelihood Cheat Sheet V2

- maximum likelihood estimator is equivariant

$\hat{\tau} = g(\hat{\theta})$ when $\tau = g(\theta)$

Recap Week 1

- likelihood definitions: $\ell(\theta)$ $\widehat{\theta}$ $U(\theta)$ $j(\widehat{\theta})$ $I_n(\theta)$
- maximum likelihood estimator minimizes KL-divergence from the true model to the data
- maximum likelihood estimator is consistent
- maximum likelihood estimator is asymptotically normal
- maximum likelihood estimator is equivariant
- point estimation, interval estimation, testing

$j(\theta) = -\ell''(\theta)$

$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I_1(\theta))$

$n \to \infty$

$\xrightarrow{d} Z$ bdd r.v.

see Likelihood Cheat Sheet V2

$\hat{se} = \sqrt{I_n(\hat{\theta})^{-1}}$

$\hat{\theta} \sim N(\theta, I_n^{-1}(\hat{\theta}))$

$\hat{\theta} \sim N(\theta, j^{-1}(\hat{\theta}))$ *

$\hat{\theta} \sim N(\theta, I_n^{-1}(\theta))$

observed vs expected Fisher info

$\hat{\theta} \sim N(\theta, j^{-1}(\theta))$

arxiv

... Your friend the delta-method

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\{0, I^{-1}(\theta)\}$$

$$\sqrt{n}\{g(\hat{\theta}_n) - g(\theta)\} \xrightarrow{d} N\{0, g'(\theta)^T I^{-1}(\theta) g'(\theta)\}$$

$$g(\hat{\theta}) = g(\theta) + (\hat{\theta} - \theta) g'(\theta) + \dots$$

$$E\,g(\hat{\theta}) = g(\theta) + g'(\theta)\, E(\hat{\theta} - \theta)$$

$$\simeq g(\theta) \quad \text{bec.} \quad E\hat{\theta} \simeq \theta$$

$$g(\hat{\theta}) \xrightarrow{P} g(\theta)$$

Variance-stabilizing transformations

$$\operatorname{var} g(\hat\theta) = E\{g(\hat\theta) - g(\theta)\}^2$$

$$= E\{(\hat\theta - \theta) g'(\theta)\}^2 + \cdots$$

$$\simeq \operatorname{var}(\hat\theta)\{g'(\theta)\}^2$$

vector not: $g'(\theta)$: matrix $\quad g(\theta) = \tau \in \mathbb{R}^k$

$$g'(\theta) = \begin{pmatrix} \frac{\partial g_1}{\partial \theta_1} & \cdots & \frac{\partial g_k}{\partial \theta_1} \\ \vdots & & \vdots \\ \frac{\partial g_1}{\partial \theta_p} & & \frac{\partial g_k}{\partial \theta_p} \end{pmatrix}_{p \times k}$$

$\operatorname{var}\hat\theta = I_n^{-1}(\theta)$

$$\operatorname{var} g(\hat\theta) \simeq \underset{k\times p}{\nabla g(\theta)^T} \underset{p\times p}{I_n^{-1}(\theta)} \underset{p\times k}{\nabla g(\theta)}$$

$k \times k$

Regularity conditions: Consistency

AoS Thm 9.13, Van der Vaart Thm 5.7

- $M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i;\theta)}{f(X_i;\theta_*)}$, $M(\theta) = E_{\theta_*} \log \frac{f(X_i;\theta)}{f(X_i;\theta_*)}$

$M(\theta)$ is maxed at θ_*

- (i) : $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$, (ii) : $\sup_{\theta:|\theta-\theta_*|<\epsilon} M(\theta) < M(\theta_*)$.

Proof of consistency: Show

$$\mathrm{pr}(|\hat{\theta}_n - \theta_*| > \epsilon) \le \mathrm{pr}\{M(\hat{\theta}_n) < M(\theta_*) - \delta\}, \text{ which } \to 0.$$

$$\begin{aligned} M(\theta_*) - M(\widehat{\theta}_n) &= M_n(\theta_*) - M(\widehat{\theta}_n) + M(\theta_*) - M_n(\theta_*) \\ &\le M_n(\widehat{\theta}_n) - M(\widehat{\theta}_n) + 0 - 0 \\ &\le \sup_\theta |M_n(\theta) - M(\theta)| \xrightarrow{p} 0 \end{aligned}$$

$$\{|\widehat{\theta}_n - \theta_*| \ge \epsilon\} \subset \{M(\widehat{\theta}_n) < M(\theta_*) - \delta\} \implies \mathrm{pr}(|\widehat{\theta}_n - \theta_*| \ge \epsilon) \le \mathrm{pr}\{M(\widehat{\theta}_n) < M(\theta_*) - \delta\} \to 0$$

Regularity conditions: Consistency

VdVaart Thm 5.7

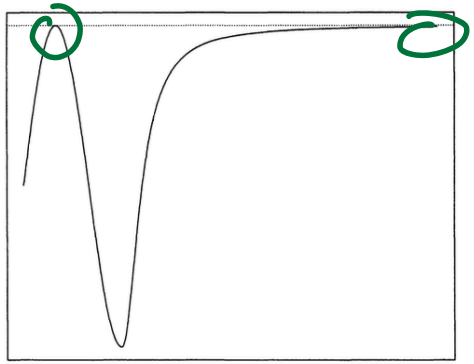


Figure 5.2. Example of a function whose point of maximum is not well separated.

5.7 *Theorem.* *Let M_n be random functions and let M be a fixed function of θ such that for every $\varepsilon > 0$*[†]

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0,$$
$$\sup_{\theta : d(\theta, \theta_0) \geq \varepsilon} M(\theta) < M(\theta_0). \tag{5.8}$$

Then any sequence of estimators $\hat{\theta}_n$ with $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$ converges in probability to θ_0.

[†] Some of the expressions in this display may be nonmeasurable. Then the probability statements are understood in terms of outer measure.

.org/10.1017/CBO9780511802256.006 Published online by Cambridge University Press

Proof. By the property of $\hat{\theta}_n$, we have $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$. Because the uniform convergence of M_n to M implies the convergence of $M_n(\theta_0) \xrightarrow{P} M(\theta_0)$, the right side equals $M(\theta_0) - o_P(1)$. It follows that $M_n(\hat{\theta}_n) \geq M(\theta_0) - o_P(1)$, whence

$\theta_0 \leftarrow \theta_*$

$$\begin{aligned} M(\theta_0) - M(\hat{\theta}_n) &\leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + o_P(1) \\ &\leq \sup_\theta |M_n - M|(\theta) + o_P(1) \xrightarrow{P} 0. \end{aligned}$$

by the first part of assumption (5.8). By the second part of assumption (5.8), there exists for every $\varepsilon > 0$ a number $\eta > 0$ such that $M(\theta) < M(\theta_0) - \eta$ for every θ with $d(\theta, \theta_0) \geq \varepsilon$. Thus, the event $\{d(\hat{\theta}_n, \theta_0) \geq \varepsilon\}$ is contained in the event $\{M(\hat{\theta}_n) < M(\theta_0) - \eta\}$. The probability of the latter event converges to 0, in view of the preceding display. ■

Regularity conditions: asymptotic normality

LaE §1.7.2; AoS Thm 9.18; MS Thm 5.4

1. The support of the density for X, i.e., the set $\{x : f(x;\theta) > 0\}$ does not depend on θ.
2. The true value θ_* is contained in an open subset of Θ, and for all θ in this open subset, the density $f(x;\theta)$ is differentiable with respect to θ for all x in the support of the density.
3. $f(x;\theta)$ is three times continuously differentiable with respect to θ for all x in the support of the density.
4. $\mathrm{E}_\theta\{u(\theta;X)\} = 0$, and $\mathrm{E}_\theta\{u(\theta;X)u^\mathsf{T}(\theta;X)\} = \mathrm{E}_\theta\{-\partial u(\theta;X)/\partial\theta^\mathsf{T}\}$, and these expectations exist and are finite for all θ in the open subset defined in 2.
5. The matrix $I_1(\theta) = \mathrm{E}_\theta\{u(\theta;X)u^\mathsf{T}(\theta;X)\}$ is positive definite for all θ in the open subset defined in 2.
6. There exist functions $M_{abc}(\cdot)$ such that

$$\left|\frac{\partial^3}{\partial\theta_a\partial\theta_b\partial\theta_c}\ell(\theta;x)\right| \le M_{abc}(x), \quad ||\theta-\theta_*|| \le \delta \quad \text{and } \mathrm{E}_{\theta_*}\{M_{abc}(X)\} < \infty.$$

$X_1, \ldots, X_n$ i.i.d. $f(x;\theta), \theta \in \Theta$

5.39 Theorem. *Suppose that the model $(P_\theta : \theta \in \Theta)$ is differentiable in quadratic mean at an inner point θ_0 of $\Theta \subset \mathbb{R}^k$. Furthermore, suppose that there exists a measurable function $\dot{\ell}$ with $P_{\theta_0}\dot{\ell}^2 < \infty$ such that, for every θ_1 and θ_2 in a neighborhood of θ_0,*

$$\left|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)\right| \leq \dot{\ell}(x)\,\|\theta_1 - \theta_2\|.$$

If the Fisher information matrix I_{θ_0} is nonsingular and $\hat{\theta}_n$ is consistent, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = I_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}}(1).$$

$o_p(1) \xrightarrow{p} 0$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $I_{\theta_0}^{-1}$.

$\xrightarrow{d} N(0, I_{\theta_0}^{-1}) \neq f(x; \theta_0)$

Today

1. Recap
2. Profile likelihood and likelihood ratio tests
3. Bayesian inference
4. Statistics in the News
5. HW Questions

... Example: logistic regression

```
Boston.glm <- glm(crim2 ~ . - crim, family = binomial,
                  data = Boston) #fit logistic regression
confint(Boston.glm)
Waiting for profiling to be done...
                    2.5 %        97.5 %
(Intercept) -47.480389822 -21.699753794
zn            -0.152359922  -0.020567540
indus         -0.149113408   0.024168460
chas          -0.646429219   2.233443233
nox           34.967619055  64.088411260
rm            -1.811639107   0.950196261
age           -0.001231256   0.046865843
dis            0.280762523   1.140619391
rad            0.376833861   0.975898274
tax           -0.012038221  -0.001324887
```

approx 95% CI for β_1 is $(-.15, -.02)$

approx 95% CI for β_4 is $(34.9, 64.1)$

...

$\hat{\beta}_1 \pm 1.96\, \hat{se}_1$

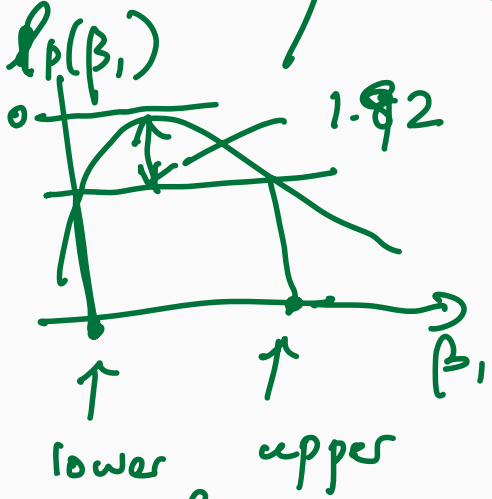
... Vector parameters

`Waiting for profiling to be done` – what's profiling?

β_1: $\ell_p(\beta_1)$ = profile log-likelihood for β_1

$= \ell(\tilde\beta_0, \beta_1, \tilde\beta_1, \dots, \tilde\beta_p)$

$\tilde\beta_0, \tilde\beta_2, \dots, \tilde\beta_p$ are f-s of β_1



$\ell_p'(\beta_1) = 0 \longrightarrow \hat\beta_1$

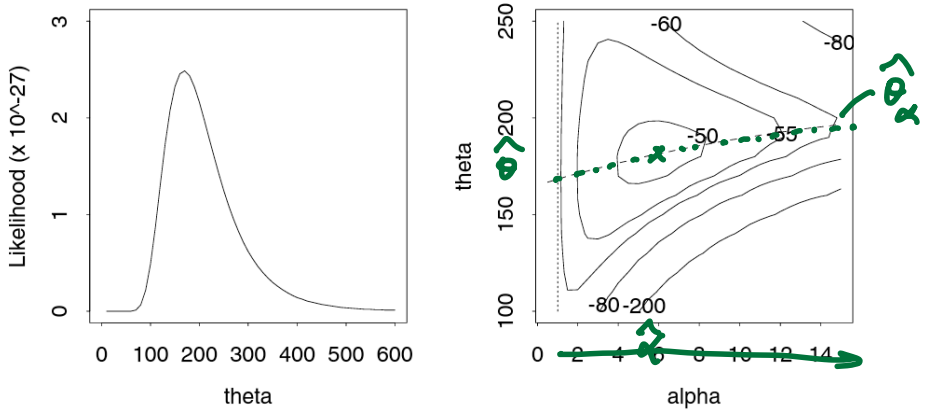
$-\ell_p''(\hat\beta_1) \doteq$ inverse variance of $\hat\beta_1$

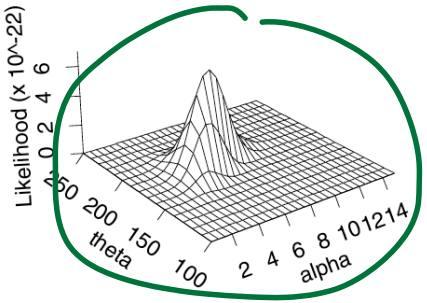
of approx 95% CI

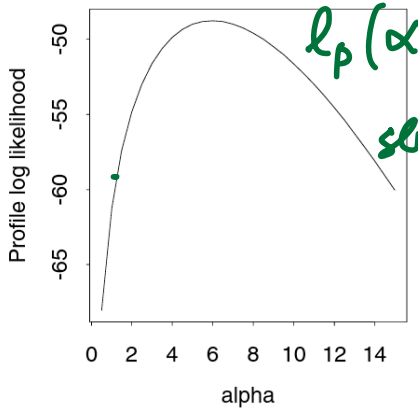
$\Pr\left(\chi^2_1 \geq \frac{1.92}{2}\right) = .05$

Profile likelihood function

Figure 4.1 Likelihoods for the spring failure data at stress 950 N/mm^2. The upper left panel is the likelihood for the exponential model, and below it is a perspective plot of the likelihood for the Weibull model. The upper right panel shows contours of the log likelihood for the Weibull model; the exponential likelihood is obtained by setting $\alpha = 1$. that is, slicing L along the vertical dotted line. The lower right panel shows the profile log likelihood for α, which corresponds to the log likelihood values along the dashed line in the panel above, plotted against α.







$\ell_p(\alpha) = \ell(\alpha, \hat{\theta}_\alpha)$

slice at ....

$\theta = (\psi, \lambda)$ $\quad \theta \in \mathbb{R}^p$

$\psi \in \mathbb{R}^d \quad \lambda \in \mathbb{R}^{p-d}$

$\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$ $\quad$ Bi $\quad \tilde{\beta}_{(-1)}(\beta)$

$$2\{\ell_p(\hat{\psi}) - \ell_p(\psi_0)\} \xrightarrow{d} \chi^2_d$$

under $f(x; \psi_0, \lambda_0)$ true (unk.)

$X_1, \ldots, X_n$ iid with jt. den $f(x; \psi_0, \lambda_0)$

i) $\ell_p'(\hat{\psi}) = 0 \qquad (\hat{\lambda}_{\hat{\psi}} = \hat{\lambda})$

$$0 = \frac{\partial}{\partial \psi} \ell_p(\psi)\Big|_{\hat{\psi}} = \frac{\partial \ell}{\partial \psi}(\psi, \hat{\lambda}_\psi)\Big|_{\hat{\psi}} = \frac{\partial \ell}{\partial \psi}(\hat{\psi}, \hat{\lambda}_{\hat{\psi}})\Big|_{\hat{\psi}} + \underbrace{\frac{\partial \ell}{\partial \lambda}(\psi, \hat{\lambda}_\psi)}_{0} \hat{\lambda}_\psi'$$

$\hat{\lambda}_{\hat{\psi}} = \hat{\lambda}$ $\quad$ bec. $\hat{\lambda}_\psi$

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{\partial \ell(\psi,\lambda)}{\partial \psi} \\ \frac{\partial \ell(\psi,\lambda)}{\partial \lambda} \end{pmatrix}\Bigg|_{\hat{\psi}, \hat{\lambda}} = 0 = \begin{pmatrix} \frac{\partial \ell(\psi,\lambda)}{\partial \psi}\Big|_{\hat{\psi}, \hat{\lambda}} \\ \frac{\partial \ell(\psi,\lambda)}{\partial \lambda}\Big|_{\hat{\psi}, \hat{\lambda}} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Example

$X_1, \ldots, X_n$ i.i.d. Gamma (α, λ) (shape, rate)

$$f(x_i; \lambda, \alpha) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x_i^{\alpha-1} \exp(-\lambda x_i)$$

$$\ell(\lambda, \alpha) = -n \log \Gamma(\alpha) + n\alpha \log \lambda + (\alpha - 1) \sum \log x_i - \lambda \sum x_i$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{n\alpha}{\lambda} - \sum x_i = 0 \Rightarrow \hat{\lambda}_\alpha = \frac{n\alpha}{\sum x_i} = \frac{\alpha}{\bar{x}}$$

Constrained mle

$$\ell_p(\alpha) = -n \log \Gamma(\alpha) + n\alpha \log(\alpha/\bar{x}) + (\alpha - 1) \sum \log x_i - \frac{\alpha n \bar{x}}{\bar{x}}$$

$$0 = \ell_p'(\hat{\alpha}) = -n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + n \log \frac{\alpha}{\bar{x}} + n \cdot \frac{1}{\alpha/\bar{x}} + \sum \log x_i - n$$

… Example

$$\frac{1}{n}\sum \log x_i - \log(\bar{x}) = \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} - \log \hat{\alpha}$$

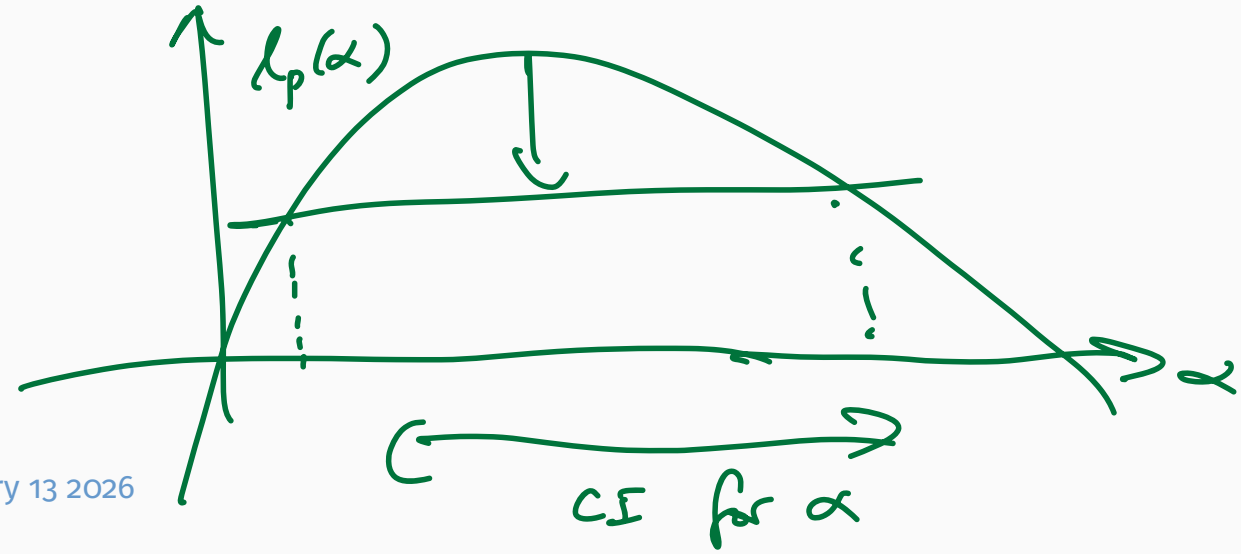
find a.var($\hat{\mu}$) via mv delta method

defines $\hat{\alpha}$

$$\overline{\log x} - \log(\bar{x}) = \psi(\hat{\alpha}) - \log \hat{\alpha}$$

$$\hat{\lambda} = \hat{\alpha}/\bar{x}$$

$$l_p(\alpha) \leq$$



Likelihood ratio statistic

SM 4.5, AoS 10.6

- model $f(\boldsymbol{x}; \theta), \quad \theta \in \mathbb{R}^p$
- likelihood and log-likelihood function $L(\theta; \boldsymbol{x}), \quad \ell(\theta; \boldsymbol{x})$
- maximum likelihood estimator $\widehat{\theta} = \widehat{\theta}(\boldsymbol{x})$

- hypothesized value θ_0 for θ
- likelihood ratio statistic $w(\theta_0) \triangleq 2\{\ell(\widehat{\theta}) - \ell(\theta_0)\}$ ← def

- Theorem: Under ... regularity conditions on the model ... if θ_0 is the true value

$$w(\theta_0) \xrightarrow{d} \chi^2_p, \quad n \to \infty,$$

- Approximation: $\{\theta : w(\theta) \geq \chi^2_p(\alpha)\}$ is a $1 - \alpha$ (approx) confidence set for θ

$$\text{pr}\{\chi^2_p \geq \chi^2_p(\alpha)\} = \alpha$$

$H_0: \theta = \theta_0$ LRT is $w(\theta_0)$ big or small? on χ^2 scale

$$W(\theta_0) = 2\{\ell(\widehat{\theta}) - \ell(\theta_0)\} \xrightarrow{d} \chi^2_p$$

$$\ell(\theta_0) = \ell(\hat\theta) + (\theta_0 - \hat\theta)^T \ell'(\hat\theta) + \frac{1}{2}(\theta_0 - \hat\theta)^T \ell''(\hat\theta)(+\ldots)$$

$1\times p$ $p\times 1$

$$\ell(\hat\theta) - \ell(\theta_0) = + \frac{1}{2}(\hat\theta - \theta_0)^2 [-\ell''(\hat\theta)]$$

$$2\{\ell(\hat\theta) - \ell(\theta_0)\} = \frac{n}{n}(\hat\theta - \theta_0)^2 I\left\{\left[-\frac{\ell''(\hat\theta)}{n}\right]/I_1\right\}$$

$\downarrow$ $\downarrow p\ 1$ ~~$I_1(\theta)$~~

$$\sqrt{n}(\hat\theta - \theta) I^{1/2}(\theta_0) \xrightarrow{d} N(0,1)$$

$$n(\hat\theta - \theta_0)^2 I(\theta) \xrightarrow{d} \chi^2_1$$

$$\Lambda_n \quad \lambda = 2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_0)\} \xrightarrow{d} \chi^2_d$$

$d = \dim(\Theta) - \dim(\Theta_0)$

$$2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_0)\} = 2\left[\ell(\hat{\theta}) - \ell(\theta_0) - \{\ell(\hat{\theta}_0) - \ell(\theta_0)\}\right]$$

$\ell(\hat{\theta}) - \ell(\theta_0) \rightarrow \chi^2_p$

$\ell(\hat{\theta}_0) - \ell(\theta_0) \sim \chi^2_{p-q}$?

$\max_{\theta \in \Theta_0} \ell(\theta)$

$\ell(\psi_0, \hat{\lambda}_{\psi_0}) - \ell(\psi_0, \lambda_0)$

χ^2_q

$\hat{\theta} \sim N(\theta, I_n^{-1}(\hat{\theta}))$

$2\{\ell(\hat{\theta}) - \ell(\theta_0)\} \sim \chi^2_p$

Calculating maximum likelihood estimators

AoS 9.13.4

- Newton Raphson iteration

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + \{-\ell''(\hat{\theta}^{(k)})\}^{-1}\ell'(\hat{\theta}^{(k)}) = \hat{\theta}^{(k)} + \frac{S(\hat{\theta}^{(k)})}{H(\hat{\theta}^{(k)})}$$

- Fisher scoring replaces $-\ell''(\cdot)$ by its expected value $J(\cdot)$
- Quasi-Newton: approximate $j(\hat{\theta}^{(t)})$ with something easy to invert and use information from $j(\hat{\theta}^{(t)})$ to compute $j(\hat{\theta}^{(t+1)})$

- EM algorithm

- Notes on optimization: Tibshirani, Pena, Kolter CO 10-725 CMU

Misspecified models

- model assumption $X_1, \ldots, X_n$ i.i.d. $f(x; \theta), \theta \in \Theta$
- true distribution $X_1, \ldots, X_n$ i.i.d. $F(x)$

notation

- maximum likelihood estimator based on model:

$$\sum_{i=1}^{n} \ell'(\hat{\theta}_n; X_i) = 0$$

- what is $\hat{\theta}_n$ estimating ?

Misspecified models

- model assumption $X_1, \ldots, X_n$ i.i.d. $f(x;\theta), \theta \in \Theta$
- true distribution $X_1, \ldots, X_n$ i.i.d. $F(x)$ notation
- maximum likelihood estimator based on model:

$$\sum_{i=1}^{n} \ell'(\hat{\theta}_n; X_i) = 0$$

- what is $\hat{\theta}_n$ estimating ?
- define the parameter $\theta(F)$ by

$$\int_{-\infty}^{\infty} \ell'\{\theta(F); x\} dF(x) = 0$$

- $$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N(0, \sigma^2)$$

- $$\sigma^2 = \frac{\int [\ell'\{\theta(F); x\}]^2 dF(x)}{(\int [\ell''\{\theta(F); x\}]^2 dF(x))^2}$$

Misspecified models

- $$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N(0, \sigma^2)$$
- $$\sigma^2 = \frac{\int [\ell'\{\theta(F); x\}]^2 dF(x)}{(\int [\ell''\{\theta(F); x\}]^2 dF(x))^2}$$
- more generally, for $\theta \in \mathbb{R}^p$,
$$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N_p\{0, G^{-1}(F)\}$$
- $$G(F) = J(F)I^{-1}(F)J(F),$$
- $$J(F) = \int -\ell''\{\theta(F); x_i\} dF(x_i), \quad I(F) = \int \{\ell'(\theta(F); x_i)\}\{\ell'(\theta(F); x_i)\}^T dF(x_i)$$

Godambe information

sandwich variance

model $\{ f(x|\theta) ; \theta \in \Theta \}$ as per usual

prior $\pi(\theta)$ density f= for θ ← new

posterior $\pi(\theta|x) = f(x|\theta)\pi(\theta) / \int f(x|\theta)\pi(\theta)d\theta$

"marg'l for x

$f(x)$

$m(x)$

sample $x_1, \cdots, x_n$

$$\pi(\theta|\underline{x}) = \frac{\prod_{i=1}^{n} f(x_i;\theta)\cdot\pi(\theta)}{\int \prod f(x_i;\theta)\pi(\theta)d\theta} = \frac{L(\theta;x)\pi(\theta)}{m(x)}$$

Frequentist:

θ_0 θ_x

- There is a fixed parameter (unknown) we are trying to learn
- Our methods are evaluated using probabilities based on $f(x;\theta)$

95% CI for β_1 is $(.7, 1.2)$

Bayesian:

- The parameter can be treated as a random variable
- We model its distribution $\pi(\theta)$ with a prob
- Combine this with a model $f(x \mid \theta)$
- Update prior belief on the basis of the data

$\pi(\theta \mid \underset{\sim}{x}) = \cdots$

$P_\pi(\theta \geq 1.5 \mid \underset{\sim}{x})$

$= 0.72$

MAP $\arg\sup_\theta \pi(\theta \mid x)$

Example: Binomial

MS 5.26; AoS Ex.11.2

$X_1, \ldots, X_n$ i.i.d. Bernoulli (θ) $\quad \pi(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, 0 < \theta < 1$ Beta(α, β)

posterior mean, mode

$$\pi(\theta|\underset{\sim}{x}) = \overbrace{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}^{L(\theta;\underset{\sim}{x})} \cdot \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \Big/ \int \cdots = d\theta$$

$$= \theta^{\sum x_i + \alpha - 1}(1-\theta)^{n-\sum x_i + \beta - 1} / B(\alpha+s, \beta+n-s)$$

$\sim$ Beta$(s+\alpha, n-s+\beta)$

$$E(\theta|\underset{\sim}{x}) = \frac{s+\alpha}{n-s+\beta+s+\alpha} = \frac{s+\alpha}{n+\alpha+\beta}$$

$\tilde{\theta}_B = \frac{s+\alpha}{n+\alpha+\beta}$ $\quad \hat{\theta} = \frac{s}{n}$ $\quad \alpha = \beta = 1$ $\quad \tilde{\theta}_B = \frac{s+1}{n+2}$

Example: censored exponential

$X_1, \ldots, X_n$ i.i.d. Exponential (λ) $\qquad\qquad$ $\pi(\lambda) \sim \text{Exp}(\alpha)$

censored at r smallest x; let $Y_i = X_{(i)}, i = 1, \ldots, r$

$$f(\boldsymbol{y} \mid \lambda) = \prod_{i=1}^{r} \lambda^r \exp(-\lambda y_i) \prod_{i=r+1}^{n} \exp(-\lambda y_r) = \lambda^r \exp[-\lambda\{\Sigma_{i=1}^{r} y_i + (n-r)y_r\}]$$

… Example: censored exponential

$$f(\boldsymbol{y} \mid \lambda) = \prod_{i=1}^{r} \lambda^r \exp(-\lambda y_i) \prod_{i=r+1}^{n} \exp(-\lambda y_r) = \lambda^r \exp\{-\lambda \Sigma_{i=1}^{r} y_i + (n-r) y_r\}, \quad \pi(\lambda) = \alpha \exp(-\alpha\lambda)$$

$\pi(\lambda \mid \boldsymbol{y})$

posterior mean and and mode

$$\hat{\theta}_B = \frac{s+\alpha}{n+\alpha+\beta}$$

$$= w_n \left(\frac{s}{n}\right) + (1-w_n)\left(\frac{\alpha}{\alpha+\beta}\right)$$

$\hat{\theta}$ — prior $E(\theta)$

Example: log-odds ratio

AoS Ex 11.3

- $X_1, \ldots, X_n$ i.i.d. Bernoulli (p); $\pi(p) \propto 1$
- posterior $\pi(p \mid \boldsymbol{x}) = Beta(s+1, n-s+1)$
- $\psi = \log\{p/(1-p)\}$

many mistakes!

$s = \Sigma x_i$

$\pi(\theta \mid \underset{\sim}{x})$

change of variables

$\pi(\tau \mid \underset{\sim}{x})$

$\tau = g(\theta)$

$$P_n(\Psi < \psi \mid \underset{\sim}{x}) = \int_{-\infty}^{\psi} \pi(\psi \mid \underset{\sim}{x})\, d\psi$$

$$= \int_{-\infty}^{e^{\psi}/(1+e^{\psi})} \pi(p \mid \underset{\sim}{x})\, dp$$

$$\pi(\psi \mid \underset{\sim}{x}) = \frac{d}{d\psi}\{\ \}$$

$$\frac{e^{\psi}(1+e^{\psi}) - e^{\psi}e^{\psi}}{(1+e^{\psi})^2} = \frac{e^{\psi}}{(1+e^{\psi})^2}$$

$$\{\ \} = \left(\frac{e^{\psi}}{1+e^{\psi}}\right)^{s}\left(\frac{1}{1+e^{\psi}}\right)^{n-s}\left(\frac{e^{\psi}}{1+e^{\psi}}\right)' \cdot \frac{1}{B(s+1, n-s+1)}$$

$$= \left(\frac{e^{\psi}}{1+e^{\psi}}\right)^{s}\left(\frac{1}{1+e^{\psi}}\right)^{n-s}\frac{e^{\psi}}{(1+e^{\psi})^2} \cdot \frac{1}{B(s+1, n-s+1)}$$

$= \left(\frac{e^\psi}{1+e^\psi}\right)^{s+1}\left(\frac{1}{1+e^\psi}\right)^{n-s+1}$.

Let $a_n = \hat{\theta}_n + a I_n^{-1/2}(\hat{\theta})$

$b_n = \hat{\theta}_n + b I_n^{-1/2}(\hat{\theta})$

$\theta \in \mathbb{R}$

$$\Rightarrow \int_{a_n}^{b_n} \pi(\theta | \underline{x}) d\theta \xrightarrow[n\to\infty]{} \Phi(b) - \Phi(a)$$

Bernstein-von Mises

under model $X_1, \ldots, X_n = \underline{X}$ iid $f(x; \theta_0)$

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx'$$

$\pi(\theta | \underline{x}) \dot{\sim} N(\hat{\theta}, I_n^{-1}(\hat{\theta}))$

We now describe a limit theorem for the posterior distribution, analogous to those in (1.46)–(1.48). For $a < b$, define

$$a_n = \widehat{\theta} + a\{J(\widehat{\theta})\}^{-1/2}, \quad b_n = \widehat{\theta} + b\{J(\widehat{\theta})\}^{-1/2},$$

where the maximum likelihood estimate $\widehat{\theta}_n$ and the observed Fisher information $J(\widehat{\theta})$ also depend on n. The limit result is

$$\int_{a_n}^{b_n} \pi(\theta \mid y)\, \mathrm{d}\theta \xrightarrow{p} \int_a^b \phi(t)\, \mathrm{d}t, \tag{1.65}$$

where $\phi(\cdot)$ is the density of the standard normal distribution. The associated approximation is expressed as

$$\pi(\theta \mid y) \dot{\sim} N\{\widehat{\theta}, J^{-1}(\widehat{\theta})\}, \tag{1.66}$$

showing that pivotal quantities from Bayesian posteriors are asymptotically equivalent to those based on (1.53), the standardized maximum likelihood estimator. This is sometimes expressed by saying "the data swamps the prior".

Calculations similar to the Taylor series expansions of Section 1.7.2 show that an equivalent approximation is

Choosing priors

MS p.287 ff

- conjugate priors
- non-informative priors — flat, “ignorance”
- convenience priors
- minimally/weakly informative priors
- hierarchical priors

Flat priors

MS p.290

- if parameter space is closed (interval), e.g. $\Theta = [a, b]$, then $\pi(\theta) \sim U(a, b)$ represents 'indifference'
- example: Beta (1,1) prior for Bernoulli probability
- example 5.34: $X \sim N(\mu, 1), \pi(\mu) \propto 1$

Flat priors

MS p.290

- if parameter space is closed (interval), e.g. $\Theta = [a, b]$, then $\pi(\theta) \sim U(a, b)$ represents 'indifference'
- example: Beta (1,1) prior for Bernoulli probability
- example 5.34: $X \sim N(\mu, 1), \pi(\mu) \propto 1$
- improper priors can lead to proper posteriors

ntbc

- priors flat in one parameterization are not flat in another

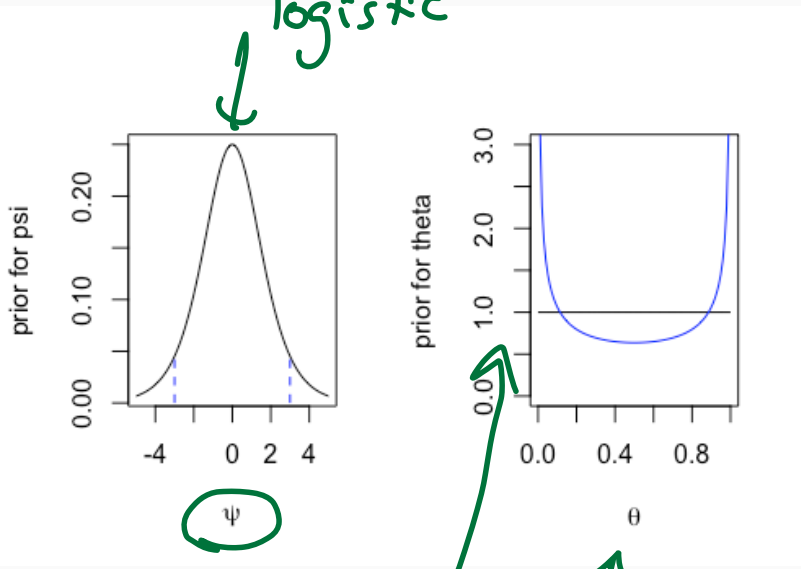
... Flat priors

- Example: $X \sim Bin(n, \theta), 0 < \theta < 1; \theta \sim U(0, 1)$
- log-odds ratio $\psi = \psi(\theta) = \log\{\theta/(1-\theta)\}$
- $\pi(\psi) = \dfrac{e^{\psi}}{(1+e^{\psi})^2}, -\infty < \psi < \infty$
- prior probability $-3 < \psi < 3 \approx 0.9$
- an invariant prior: $\pi(\theta) \propto I^{1/2}(\theta)$

$\{\theta(1-\theta)\}^{1/2} = \pi_J(\theta)$

logistic

$\pi(\theta) = 1, \quad 0 \le \theta \le 1$



$\pi_J(\psi) = \{I_1(\psi)\}^{1/2}$

$\pi(\psi) \propto 1$

$\pi(\theta)$?

Jeffreys' prior

MS p 291

- $\pi(\theta) \propto I^{1/2}(\theta)$ invariant under 1-1 transf.
- Example: $X \sim Bin(n, \theta)$ $I(\theta) = n/\{\theta(1-\theta)\}, \quad 0 < \theta < 1$
- Example 5.35: $X \sim Poisson(\lambda)$, $I(\lambda) = 1/\lambda, \quad \lambda > 0$ posterior proper?
- Jeffreys' prior for multiparameter θ: $\pi(\theta) \propto |I(\theta)|^{1/2}$ not recommended even by Jeffreys
- Example: $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$ $I(\mu, \sigma^2) =$

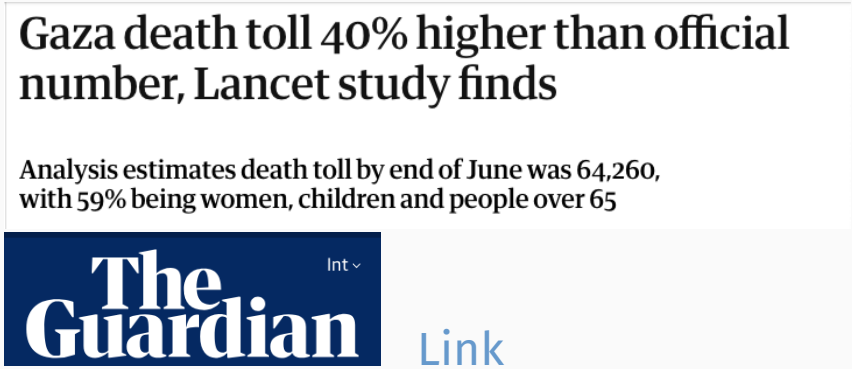
Marginalization

- Bayes posterior carries all the information about $\boldsymbol{\theta}$, given $\boldsymbol{x}$ by definition
- probabilities for any set A computed using the posterior distribution
- $\mathrm{pr}(\boldsymbol{\Theta} \in A \mid \boldsymbol{x}) =$
- if $\boldsymbol{\theta} = (\psi, \boldsymbol{\lambda})$, ...
- or, if $\psi = \psi(\boldsymbol{\theta})$
- in this context, 'flat' priors can have a large influence on the marginal posterior

Example: many normal means

Stein, 1959

Statistics in the News



Link

- "The peer-reviewed statistical analysis was conducted by academics at the London School of Hygiene & Tropical Medicine, Yale University and other institutions, using a statistical method called capture-recapture analysis"
- "The study used death toll data from the health ministry, an online survey launched by the ministry for Palestinians to report relatives' deaths, and social media obituaries"
- "Patrick Ball, a statistician at the US-based Human Rights Data Analysis Group not involved in the research, has used capture-recapture methods to estimate death tolls for conflicts in Guatemala, Kosovo, Peru and Colombia.

Traumatic injury mortality in the Gaza Strip from Oct 7, 2023, to June 30, 2024: a capture–recapture analysis

Zeina Jamaluddine, Hanan Abukmail, Sarah Aly, Oona M R Campbell, Francesco Checchi

Summary

Background Accurate mortality estimates help quantify and memorialise the impact of war. We used multiple data sources to estimate deaths due to traumatic injury in the Gaza Strip between Oct 7, 2023, and June 30, 2024.

Methods We used a three-list capture–recapture analysis using data from Palestinian Ministry of Health (MoH) hospital lists, an MoH online survey, and social media obituaries. After imputing missing values, we fitted alternative generalised linear models to the three lists' overlap structure, with each model representing different possible dependencies among lists and including covariates predictive of the probability of being listed; we averaged the models to estimate the true number of deaths in the analysis period (Oct 7, 2023, to June 30, 2024). Resulting annualised age-specific and sex-specific mortality rates were compared with mortality in 2022.

Findings We estimated 64 260 deaths (95% CI 55 298–78 525) due to traumatic injury during the study period, suggesting the Palestinian MoH under-reported mortality by 41%. The annualised crude death rate was 39·3 per 1000 people (95% CI 35·7–49·4), representing a rate ratio of 14·0 (95% CI 12·8–17·6) compared with all-cause mortality in 2022, even when ignoring non-injury excess mortality. Women, children (aged <18 years), and older people (aged ≥65 years) accounted for 16 699 (59·1%) of the 28 257 deaths for which age and sex data were available.

Interpretation Our findings show an exceptionally high mortality rate in the Gaza Strip during the period studied. These results underscore the urgent need for interventions to prevent further loss of life and illuminate important

Published Online
January 9, 2025
https://doi.org/10.1016/S0140-6736(24)02678-3

Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK (Z Jamaluddine PhD, H Abukmail MD, S Aly DO, Prof O M R Campbell PhD, Prof F Checchi PhD); **School of Tropical Medicine and Global Health, Nagasaki University, Nagasaki, Japan** (Z Jamaluddine); **International Health System Research Group, Department of Engineering, University of Cambridge, Cambridge, UK** (H Abukmail); **Department of Emergency Medicine, Yale School of**

- finite population of size N, unknown
- collect a sample of size n, tag each item, return to population
- collect a second sample of size m, record number of tags y
- estimate of N:

$$\frac{n}{N} \approx \frac{y}{m} \implies \widehat{N} = \frac{nm}{y}$$

$$\operatorname{var} \hat{N} = (nm)^2 \cdot \operatorname{var}\left(\frac{1}{y}\right)$$

$$= n^2 \operatorname{var}\left(\frac{1}{y/m}\right)$$

- assumptions? variability?

- finite population of size N, unknown
- collect a sample of size n, tag each item, return to population
- collect a second sample of size m, record number of tags y
- estimate of N:

$$\frac{n}{N} \approx \frac{y}{m} \implies \widehat{N} = \frac{nm}{y}$$

- assumptions? variability?

- List 1: Ministry of Health data from hospitals
- List 2: Ministry of Health rolling survey of mortality and missing persons
- List 3: Social media (obituary pages)
- remove duplicates

using IDs and/or probabilistic record linkage

... Capture-recapture analysis

- 3 Lists, duplicates removed
- match people on the list
- use log-linear model fitted to probabilities of $\{001, 010, 100, 011, 101, 110, 111\}$ ntbc
- use fitted model to estimate probability of 000 many other details

… Capture-recapture analysis

$\hat{\theta} \pm 1.96\,\hat{se}$ Wald interval

- 3 Lists, duplicates removed
- match people on the list
- use log-linear model fitted to probabilities of $\{001, 010, 100, 011, 101, 110, 111\}$ ntbc
- use fitted model to estimate probability of 000 many other details

List 1	22347 unique records
List 2	7581 unique records
List 3	3190 unique records

- adding estimate of 000 yields total estimate of $64,260$ 95% CI $[55,298 - 78,525]$
- "we computed Wald confidence intervals"

  "while recognizing that bootstrap or profile intervals might have better coverage"
- see also Zivot et al. (2025) (Letter to the editor) and Jamaluddine et al. (2025) (Authors' reply)

… Capture-recapture analysis

Jamaluddine et al. 25

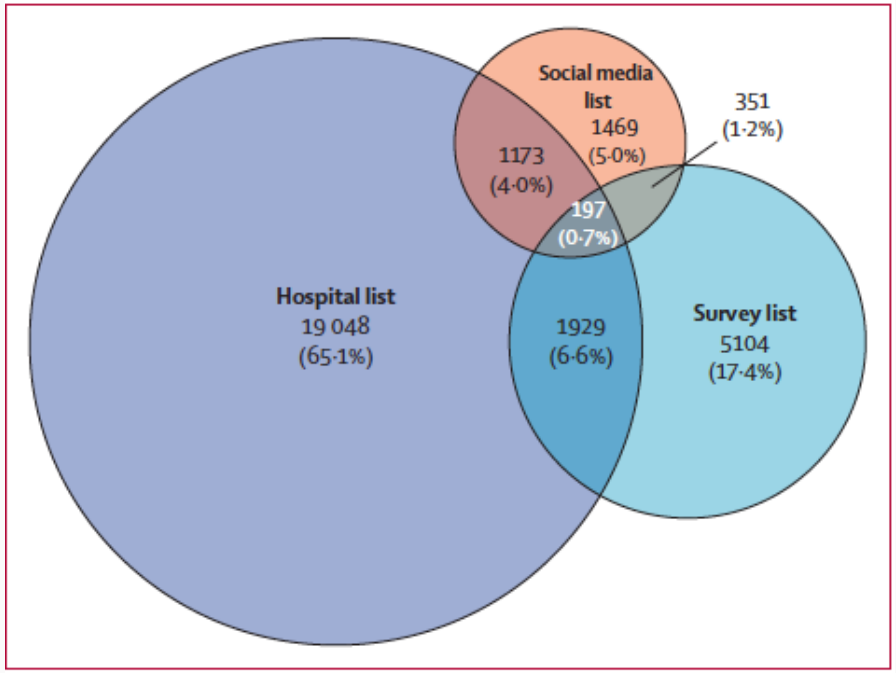


Figure 3: Overlap of decedents among the three lists (hospital, survey, and social media)