Serosurvey in Karnataka State: Summary, Design and Statistical Methodology II

Rajesh Sundaresan

IISc, on sabbatical leave at Strand Life Sciences

(Joint work with a large team – the Karnataka Serosurvey team and the IISc/ISI team)

Acknowledgements

- Strand Life Sciences

- Serosurvey project funded by NHM

- Google, Hitachi, Cisco CSR (Centre for Networked Intelligence), DST that have funded our miscellaneous COVID-19 response efforts

- Comrades-in-arms: Siva, Giri & others in the serosurvey team, Aniruddha, Minhaas, Nidhin, Nihesh, Sarath, Sharad.

Recall

- Goal 1: Estimate the total prevalence of COVID-19 in a locality
- Total burden = Past Infection + Current Infection
  - Important to do both when the active infection is high
  - Different types of tests capture different information, so it’s important to do multiple tests.

Model and ideal test outcomes

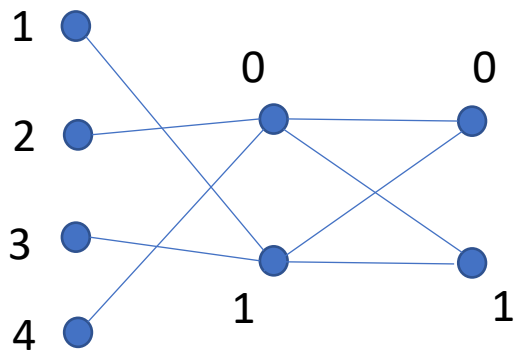
- Each individual can be in one of four different states

State	Probability	State description	RAT $j = 1$	RT-PCR test $j = 2$	IgG Antibody test $j = 3$
$s = 1$	p_1	Active infection, but no IgG	1	1	0
$s = 2$	p_2	IgG antibodies only, no active infection	0	0	1
$s = 3$	p_3	Both active infection and IgG antibodies	1	1	1
$s = 4$	p_4	Neither active infection nor IgG antibodies	0	0	0

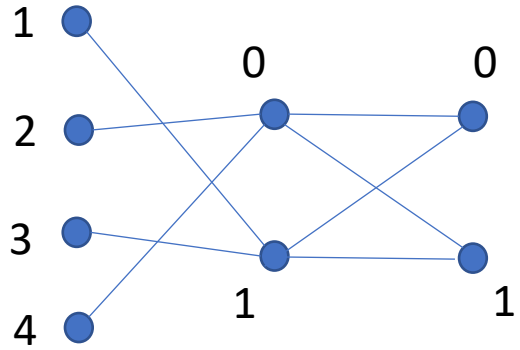
$M(s, j)$

- One model for the ideal test outcomes
- Must do IgG test to assess antibody prevalence
- Must do either RAT or RT-PCR or both for assessing active infection

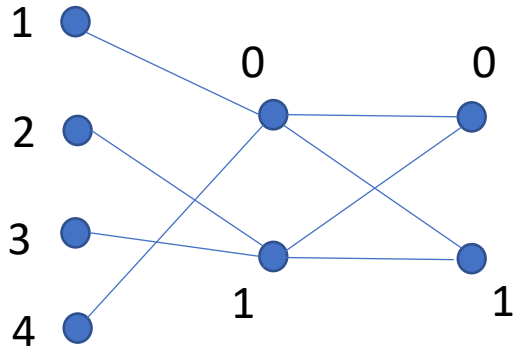
Test outcomes are noisy: tandem channels



RAT	
Sensitivity	0.5
Specificity	0.975



RTPCR	
Sensitivity	0.95
Specificity	0.97



IgG ELISA kit	
Sensitivity	0.921
Specificity	0.977

- Sensitivity = 1 – false negative rate = 1 – miss probability
- Specificity = 1 – false positive rate = 1 – false alarm probability

Protocol nuances, data issues

- Only a subset of individuals were administered the RAT
- Those who are RAT positive are not administered the RT-PCR test
- We didn’t receive RT-PCR on 1000+ samples due to delays
- IgG results from one hospital locality didn’t come
- Couldn’t match some IgG results to participants because of entry errors, duplicate SRF id issues …

Test patterns and test outcomes

Individual	RAT done	RT-PCR done	IgG done	RAT outcome	RT-PCR outcome	IgG outcome
1	1	1	1			
2	Blank = 0	1	1	NA		
3	1	Blank = 0	1	1	NA	
4	1	Blank = 0	1	0	NA	
5	1	1	Blank = 0			NA

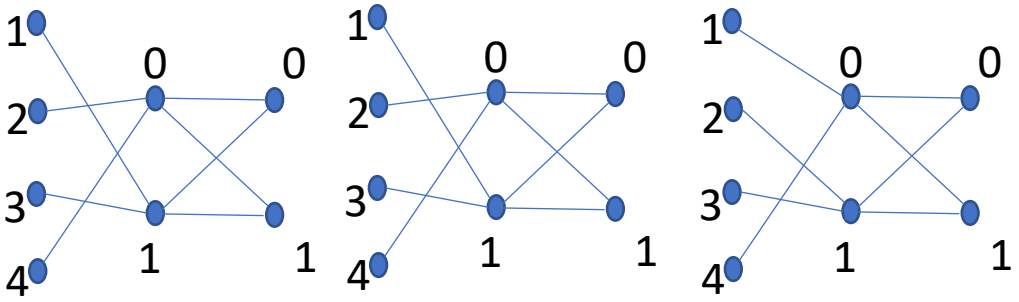
$t = (t_1, t_2, t_3) \in \{0,1\}^3$

8 possibilities

$y = (y_1, y_2, y_3) \in \{0,1, NA\}^3$

27 possibilities
If $t_i = 0, y_i = NA$

Parametric model (contd.)



- Let $p = (p_1, p_2, p_3)$ with the usual positivity and sum conditions
- The four disease state probabilities are $(p_1, p_2, p_3, p_4 = 1 - (p_1 + p_2 + p_3))$
- Assume that N individuals are sampled, N small enough that we may assume their states are iid $\sim (p_1, \ldots, p_4)$
- For each individual n, we know the set of administered tests:
  $t(n) \in \{0,1\}^3$ and the test outcomes $y(n) \in \{0,1,NA\}^3$
- Likelihood, from Siva's slides:

$$P_p(y(n)|t(n)) = \sum_{s=1}^{4} p_s \cdot q(y(n)|t(n), s)$$

$$q(y|t,s) = \prod_{j:\, t_j=1} [\sigma(M(s,j),j)]^{1\{M(s,j)=y_j\}} \cdot [1 - \sigma(M(s,j),j))]^{1-1\{M(s,j)=y_j\}}$$

Maximum likelihood estimation

- Given the test patterns (assumed independent of p), the likelihood of the tests' outcomes on N participants is:

$$L\left(p; \left(t(n), y(n)\right)_n\right) = \prod_{n=1}^{N} P_p(y(n)|t(n))$$

- Find the $\hat{p}(N)$ that best explains the test outcomes:

$$\hat{p}(N) = \arg\max_p L(p;\ (t(n), y(n))_n)$$

- Concave function of p, unique maximum, easy to identify the MLE

Consistency of the MLE and asymptotic normality

- Under some regularity conditions on the score function, which our model satisfies, the MLE is consistent as $N \to \infty$:

$$\hat{p}(N) \to p \text{ in probability}$$

- Under additional conditions, which our model once again satisfies

$$\sqrt{N}(\hat{p}(N) - p) \to Normal(0, i(p)^{-1}) \text{ in distribution}$$

- $i(p)$ is the per-sample Fisher information matrix at p.

- This suggests that the following is a good approximation:

$$\hat{p}(N) \sim Normal(p, (Ni(p))^{-1})$$

Confidence intervals

- Estimates, suppressing N,
  - Active infection: $\hat{p}_1 + \hat{p}_3$
  - IgG prevalence: $\hat{p}_2 + \hat{p}_3$
  - Total disease burden: $\widehat{\wp} = \hat{p}_1 + \hat{p}_2 + \hat{p}_3 = u^T \hat{p}$, where $u = [1, 1, 1]^T$
- Var$(\widehat{\wp})$ is approximately $u^T (Ni(p))^{-1} u$
- 95% confidence: $\widehat{\wp} \pm 1.96\sqrt{u^T (Ni(p))^{-1} u}$
- Design effect of 3 increases the variance by a factor 3 to account for sampling biases.

More about the Fisher information matrix

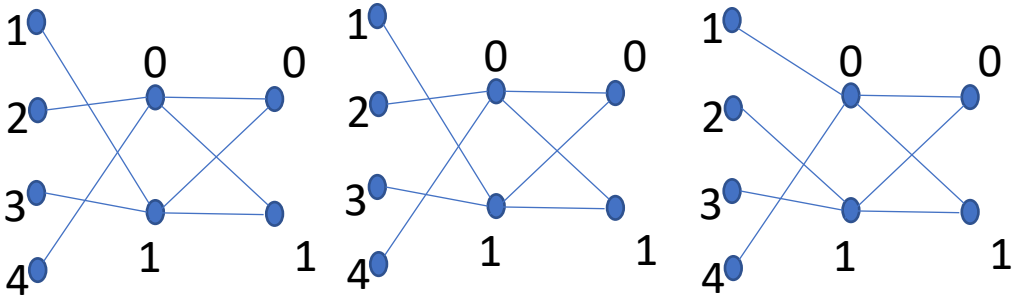
- Variance is approximately: $u^T (Ni(p))^{-1} u$, where $u = [1, 1, 1]^T$

$$Ni(p) = \sum_{t \in T} w_t \cdot i_t\,(p)$$

- Here w_t is the number of tests of test pattern t

  and $i_t(p)$ is the Fisher information per sample when test pattern is t

So, what's new?



- An honest-to-goodness assessment: Perhaps the above picture
  - Handles multiple tests on a participant naturally
  - Enhances evidence for IgG = 0 if either RAT or RTPCR is positive, and vice-versa
  - Naturally handles noisy observations, e.g., RAT sensitivity is 50%
  - Naturally handles partial data
- Once the model is identified, it's standard fare all the way
- If only IgG antibody test is done, there's a closed form expression for the MLE given by the so-called Rogan-Gladen formula

$$\left[\frac{Crude\ estimate(IgG)\ +\ Specificity(IgG)\ -1}{Sensitivity(IgG)\ +\ Specificity(IgG)\ -1}\right]_0^1$$

Improving the Karnataka survey

- In the serosurvey, we lived with whatever $(w_t, t \in T)$ we got
- Could we have done better for the money we spent?
- Test costs (approximate)
  - RAT – Rs. 450
  - RT-PCR – Rs. 1200
  - IgG – Rs. 300
- 11000 RAT + 16500 RT-PCR + 16500 IgG: Cost = Rs. 3 Crores

A design problem

- Given budget C, cost c_t for test pattern t, how many participants should be administered test pattern t?

- Relevant question. Why?
  - If test pattern $t = (1,0,1)$, cost is Rs. 750
  - If test pattern $t' = (0,1,1)$, cost is Rs. 1500
  - For the same cost, I could administer the first pattern to two individuals

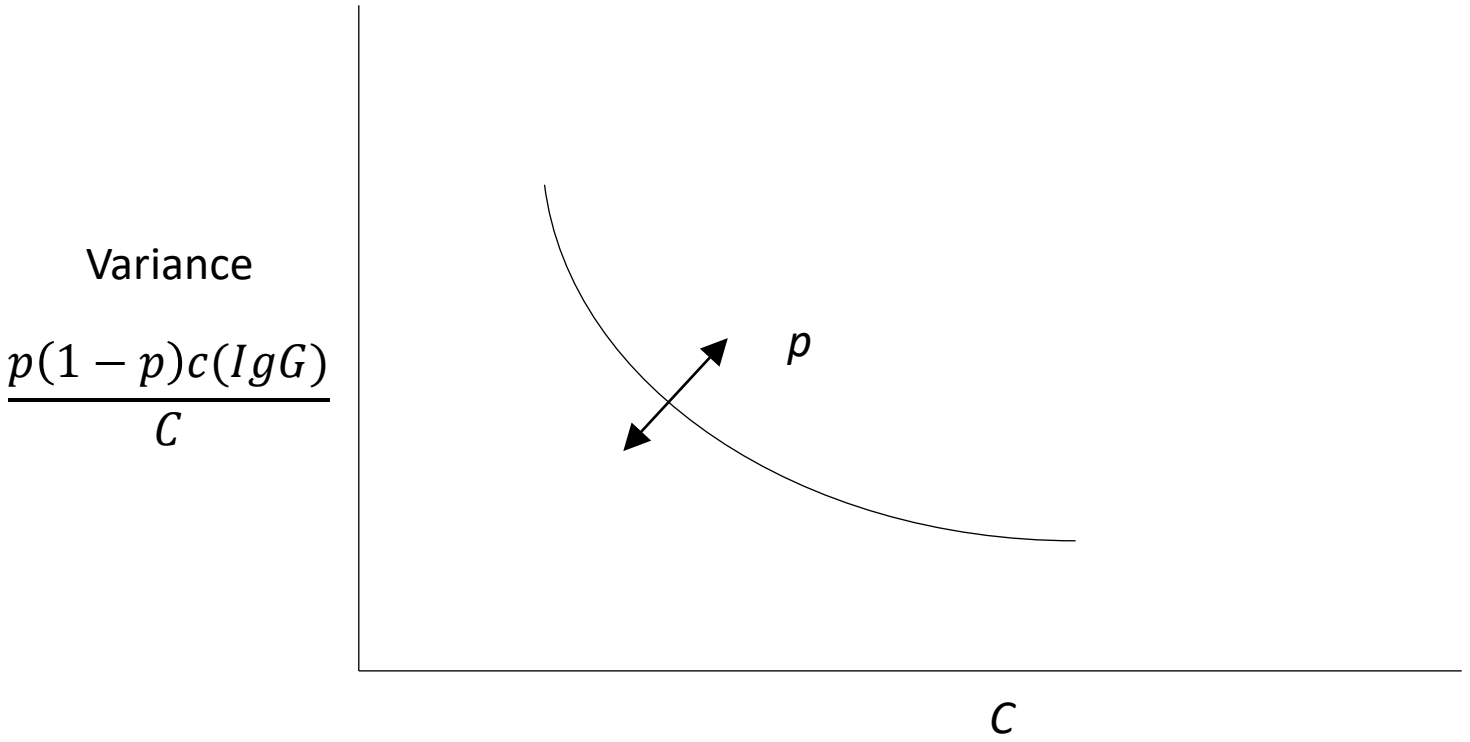
- How should the field epidemiologist allocate resources? What's the epidemiologist's goal?

An instructive look at the simplest case

- Budget C. Allow only one test, the IgG test. It's cost is $c(IgG)$
- If N tests are administered, the standard estimator's variance is $\frac{p(1-p)}{N}$
- Cost of N tests $Nc(IgG) \leq C$, or $N \leq \frac{C}{c(IgG)}$
- To minimise variance, need N as large as possible, so $N = \frac{C}{c(IgG)}$
- Thus the minimum variance is $\frac{p(1-p)c(IgG)}{C}$
- Worst case design: $\frac{c(IgG)/4}{C}$

  or if you have some side information about p, find the worst case within a range

How much accuracy can the budget buy?



Back to the design problem

- Goal 2: Given budget C, cost c_t for test pattern t, how many participants should be administered test pattern t in order to minimise the variance of the total disease burden $\widehat{\wp}$

- Mathematical formulation:

$$\min_{w} \quad u^T \left(\sum_t w_t \, i_t(p) \right)^{-1} u$$

$$subject\ to \sum_t w_t c_t \leq C \, , w_t \geq 0 \ \forall t$$

The c-optimal design

Theorem:

Let the vector v^* optimise

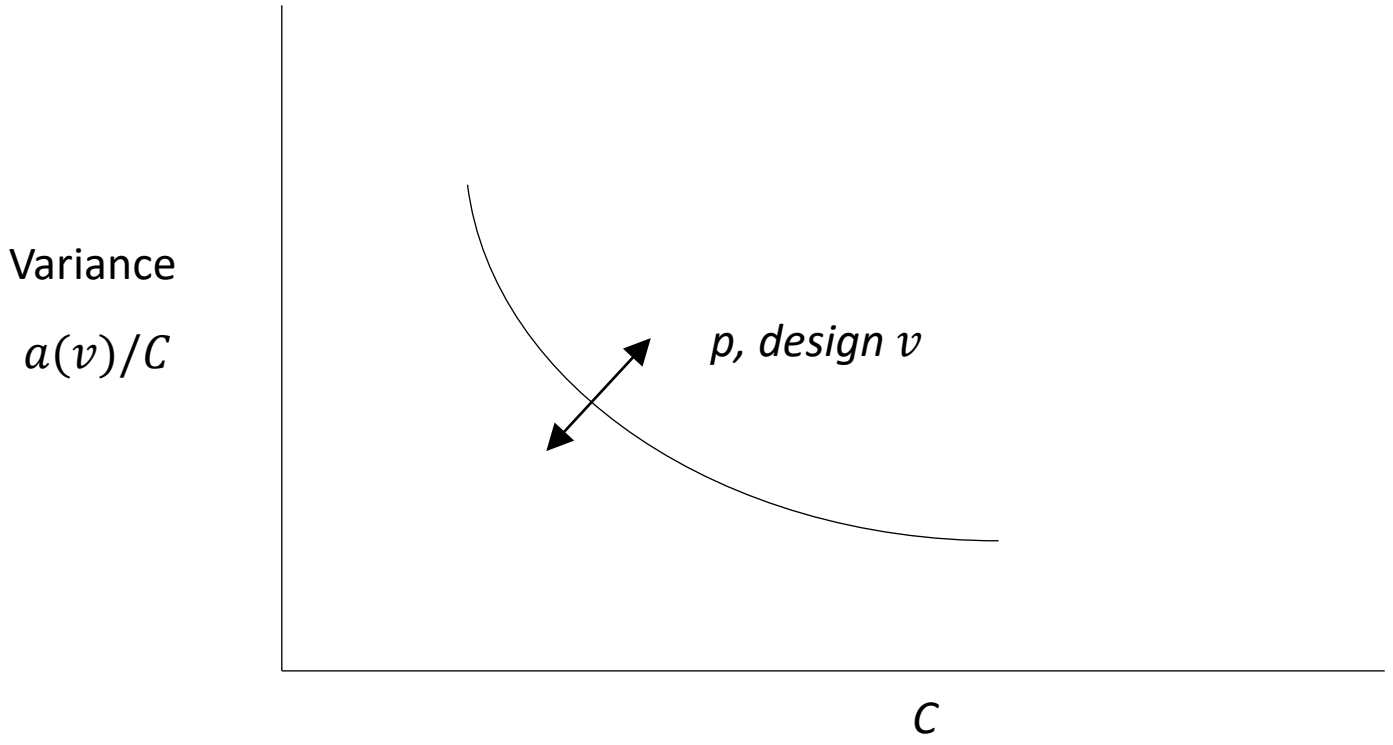
$$\min_{v} \quad u^T \left(\sum_t v_t \, i_t(p)/c_t \right)^{-1} u$$

$$subject\ to \quad \sum_t v_t \leq 1\,, \qquad v_t \geq 0 \ \forall t$$

Then the optimal allocation w^* satisfies $w_t^* = (v_t^*/c_t)\, C$.

The minimum variance is $\frac{a(v^*)}{C}$, where $a(v^*)$ is the value of the above optimisation problem.

How much accuracy can the budget buy?



Numerical examples

- Test costs (approximate)
  - RAT – Rs. 450
  - RT-PCR – Rs. 1200
  - IgG – Rs. 300

- If RT-PCR cost is Rs. 1200

  (0,0,IgG) : (RAT,0,IgG) = 1:24

- If RT-PCR cost reduces to Rs. 1000

  (0,0,IgG) : (0,RT-PCR,IgG) = 4:3

Extension 1: Worst case design

$$(v, p) \mapsto u^T \left(\sum_t v_t \, i_t(p)/c_t \right)^{-1} u$$

- For a fixed p, a convex function of v
- For a fixed v, a concave function of p

Theorem: If the sets for v and p are compact and convex, the "game" has a value.

Extension 2: Handling observables

- RAT is 68% sensitive on symptomatics versus 47% on asymptomatics
- If r(0) fraction of the population is asymptomatic and r(1) population is symptomatic, what's the optimal allocation policy knowing symptom presentation?
- $\min_{w(0),w(1)} \sum_x r(x) u^T (\sum_t w_t(x)\ i_t(p,x))^{-1} u$, subject to budget constraints
- Can solve this also quite easily.
  - Increase use of RAT on symptomatics
  - Consider budget subhead C_0 and C_1 for asymptomatics and symptomatics.
  - For each of these, $v_t^*(0)$ and $v_t^*(1)$ are independent of C_0 and C_1.
  - Then optimise optimise over C_0 and C_1.

Summary

- We demonstrated how to optimally allocate test patterns to minimise the variance: c-optimal design
- We found the accuracy that your budget can buy. The test proportions don’t change
- Chernoff 1953, Trace (Inverse Fisher Information)
- Note the goal – minimise the variance disease burden.
- In practice, there may be additional goals that may warrant the use of RAT on account of its PoC usability