Un passeig per la Modelització Estadística de la Covid-19

Pere Puig

Department de Matemàtiques

Universitat Autònoma de Barcelona

Advanced Stochastic Modelling Research Group

Overview

1- El coronavirus SARS-CoV-2

2- El període d'incubació

3- Tests massius?

1- El coronavirus SARS-CoV-2

La seqüència de nucleotides es pot baixar des del National Centre for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/).

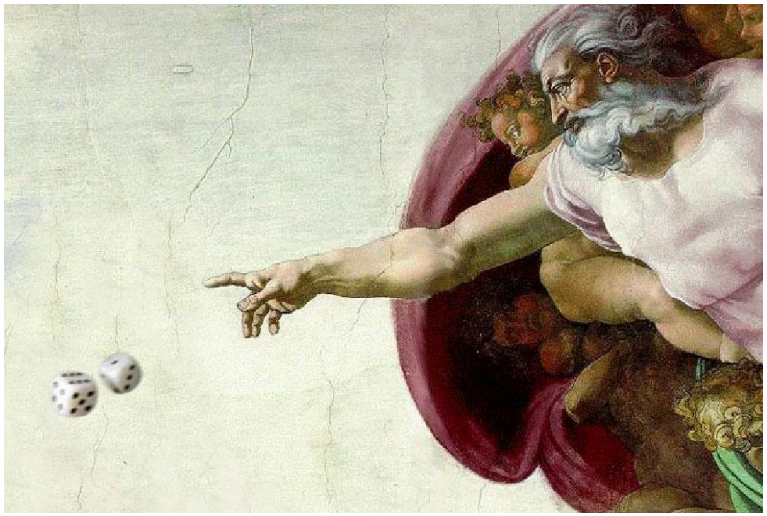
Per exemple, el genoma seqüenciat a Catalunya el 15/03/2020, accession number MT359865.

Andres,C., Garcia-Cehic,D., Pinana,M., Guerrero-Murillo,M., Rando,A., Pumarola,T., Codina,M.G., Anton,A. and Quer,J. Respiratory Viruses Unit, Microbiology Department, Vall d'Hebron University Hospital.

```
   [1] a c c t t c c c a g g t a a c a a a c c a a c c a a c t t t c g a t c t
  [37] c t t g t a g a t c t g t t c t c t a a a c g a a c t t t a a a a t c t
  [73] g t g t g g c t g t c a c t c g g c t g c a t g c t t a g t g c a c t c
 [109] a c g c a g t a t a a t t a a t a a c t a a t t a c t g t c g t t g a c
 [145] a g g a c a c g a g t a a c t c g t c t a t c t t c t g c a g g c t g c
 [181] t t a c g g t t t c g t c c g t g t t g c a g c c g a t c a t c a g c a
 [217] c a t c t a g g t t t t g t c c g g g t g t g a c c g a a a g g t a a g
 [253] a t g g a g a g c c t t g t c c c t g g t t t c a a c g a g a a a a c a
 [289] c a c g t c c a a c t c a g t t t g c c t g t t t t a c a g g t t c g c
 [325] g a c g t g c t c g t a c g t g g c t t t g g a g a c t c c g t g g a g
 [361] g a g g t c t t a t c a g a g g c a c g t c a a c a t c t t a a a g a t
 [397] g g c a c t t g t g g c t t a g t a g a a g t t g a a a a a g g c g t t
 [433] t t g c c t c a a c t t g a a c a g c c c t a t g t g t t c a t c a a a
 [469] c g t t c g g a t g c t c g a a c t g c a c c t c a t g g t c a t g t t
 [505] a t g g t t g a g c t g g t a g c a g a a c t c g a a g g c a t t c a g
 [541] t a c g g t c g t a g t g g t g a g a c a c t t g g t g t c c t t g t c
 [577] c c t c a t g t g g g c g a a a t a c c a g t g g c t t a c c g c a a g
 [613] g t t c t t c t t c g t a a g a a c g g t a a t a a a g g a g c t g g t
 [649] g g c c a t a g t t a c g g c g c c g a t c t a a a g t c a t t t g a c
 [685] t t a g g c g a c g a g c t t g g c a c t g a t c c t t a t g a a g a t
 [721] t t t c a a g a a a a c t g g a a c a c t a a a c a t a g c a g t g g t
 [757] g t t a c c c g t g a a c t c a t g c g t g a g c t t a a c g g a g g g
 [793] g c a t a c a c t c g c t a t g t c g a t a a c a a c t t c t g t g g c
 [829] c c t g a t g g c t a c c c t c t t g a g t g c a t t a a a g a c c t t
 [865] c t a g c a c g t g c t g g t a a a g c t t c a t g c a c t t t g t c c
 [901] g a a c a a c t g g a c t t t a t t g a c a c t a a g a g g g g t g t a
 [937] t a c t g c t g c c g t g a a c a t g a g c a t g a a a t t g c t t g g
 [973] t a c a c g g a a c g t t c t g a a a a g a g c t a t g a a t t g c a g
[1009] a c a c c t t t t g a a a t t a a a t t g g c a a a g a a a t t t g a c
```

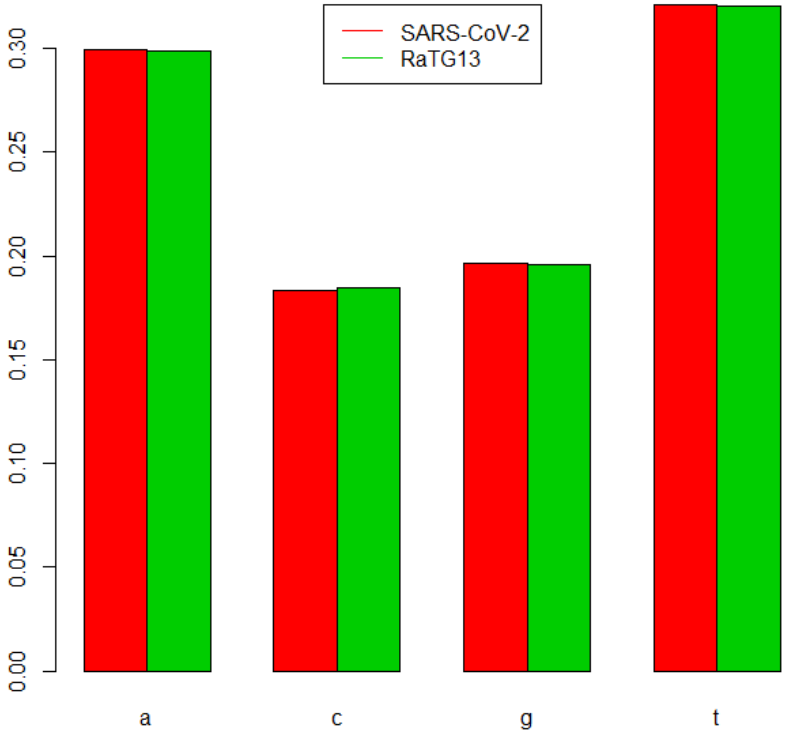
El genoma del coronavirus SARS-CoV-2 té una longitud d'una mica menys de 30.000 símbols.

Es fan moltíssimes anàlisis estadístiques amb els genomes, moltes fent inferència estadística tot i que són "objectes" deterministes.



Models multinomials, z-scores per a dinucleòtids, models markovians (higher order), hiden markov chains, etc...

El virus RaTG13 dels ratpenats Rhinolophus afinis, accesion number MN996532, és molt semblant al SARS-CoV-2



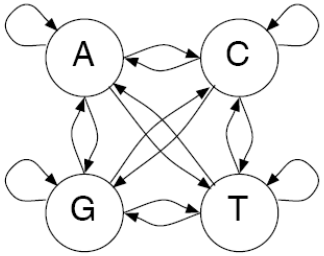
Una anàlisi (global) una mica més complexa, en aquest cas l'ajust mitjançant un model de Markov d'ordre 1 (MC1) accentua la similitud:

	a	c	g	t
a	0.32	0.23	0.19	0.26
c	0.38	0.16	0.08	0.38
g	0.27	0.20	0.19	0.34
t	0.25	0.15	0.27	0.34

SARS-CoV-2

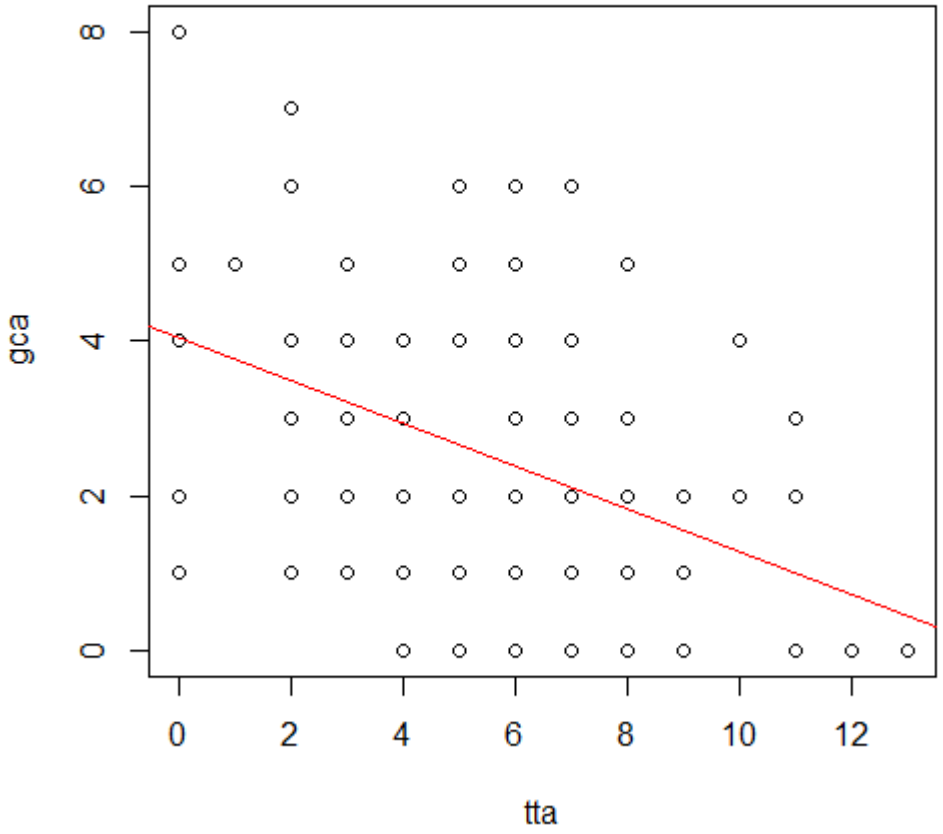
	a	c	g	t
a	0.32	0.23	0.19	0.26
c	0.37	0.16	0.08	0.39
g	0.28	0.20	0.19	0.34
t	0.25	0.15	0.27	0.33

RaTG13



States: A,C,G,T
Emissions: corresponding letter
Transitions: $a_{st} = P(x_i = t \mid x_{i-1} = s)$

Cal anar a mirar relacions més complexes. Per exemple, fem recomptes del nombre de vegades que apareix cada trinucleòtid en trossos de cadenes de longitud 200 (sliding window analysis). En total tenim 149 trossos.



```
        Pearson's product-moment correlation

data:  tta and gca
t = -5.924, df = 147, p-value = 2.137e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5602529 -0.2993324
sample estimates:
       cor
-0.4390021
```

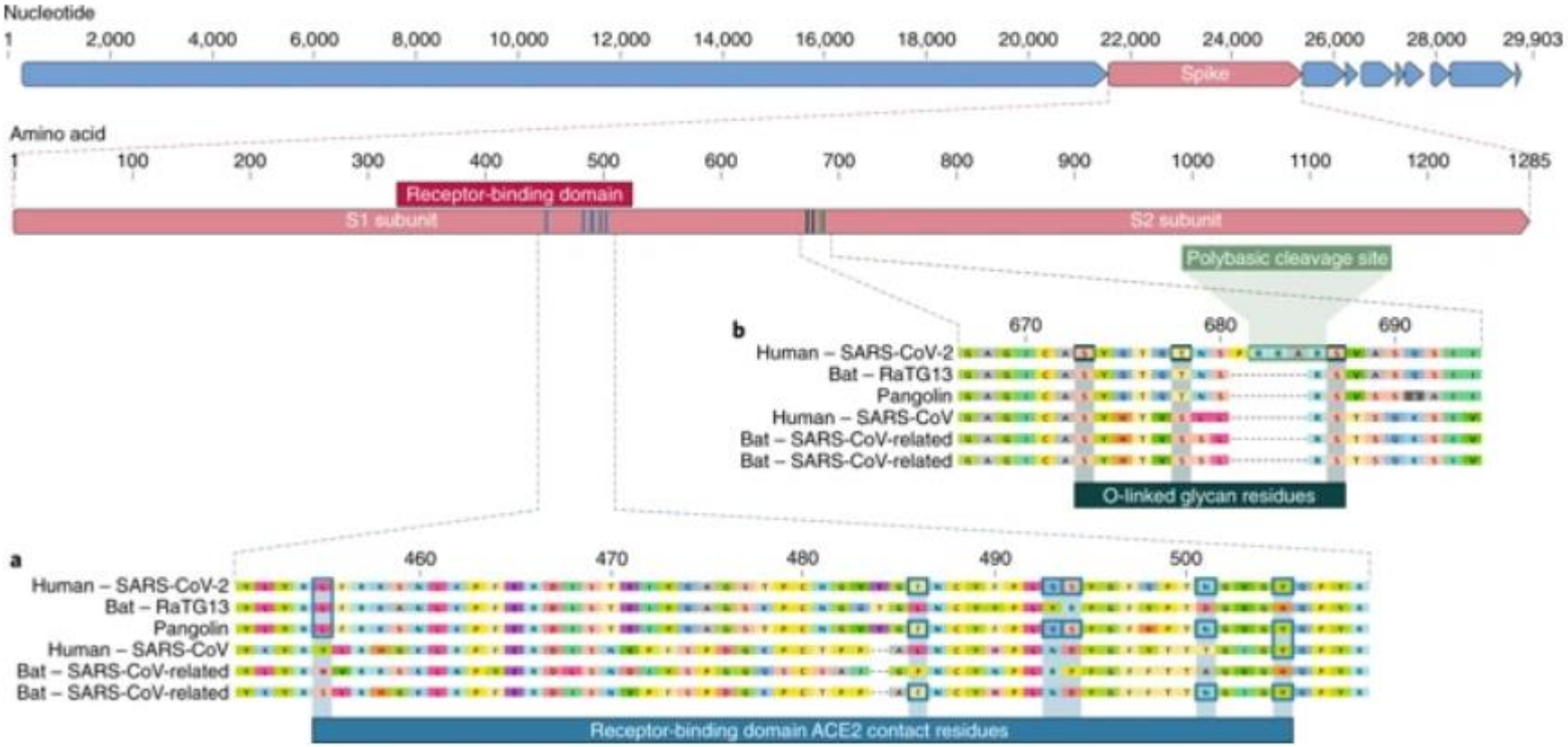
De totes les combinacions de 2 trinucleòtids ($C_{64,2}$ = 2016) del SARS-CoV-2 els que correlacionen més negativament (relació de competència) són el gca i el tta.

Què passa amb les principals relacions de competència en el RaTG13 i el SARS-CoV-2 ?

SARS-CoV-2			**RaTG13**		
gca	tta	r=-0.439	**tcg**	**tta**	**r=-0.417**
cgc	tta	r=-0.419	gca	tta	r=-0.391
aat	ctg	r=-0.414	cgc	tta	r=-0.379

nature medicine

Fig. 1: Features of the spike protein in human SARS-CoV-2 and related coronaviruses.



2- El període d'incubació

United Nations | COVID-19 Response

How long is the incubation and transmission period for COVID-19?

The "incubation period" means the time between catching the virus and beginning to have symptoms of the disease. Most estimates of the incubation period for COVID-19 range from 1-14 days, most commonly around five days.

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 MARCH 26, 2020 VOL. 382 NO. 13

Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia

Qun Li, M.Med., Xuhua Guan, Ph.D., Peng Wu, Ph.D., Xiaoye Wang, M.P.H., Lei Zhou, M.Med., Yeqing Tong, Ph.D., Ruiqi Ren, M.Med., Kathy S.M. Leung, Ph.D., Eric H.Y. Lau, Ph.D., Jessica Y. Wong, Ph.D., Xuesen Xing, Ph.D., Nijuan Xiang, M.Med., Yang Wu, M.Sc., Chao Li, M.P.H., Qi Chen, M.Sc., Dan Li, M.P.H., Tian Liu, B.Med., Jing Zhao, M.Sc., Man Liu, M.Sc., Wenxiao Tu, M.Med., Chuding Chen, M.Sc., Lianmei Jin, M.Med., Rui Yang, M.Med., Qi Wang, M.P.H., Suhua Zhou, M.Med., Rui Wang, M.D., Hui Liu, M.Med., Yinbo Luo, M.Sc., Yuan Liu, M.Med., Ge Shao, B.Med., Huan Li, M.P.H., Zhongfa Tao, M.P.H., Yang Yang, M.Med., Zhiqiang Deng, M.Med., Boxi Liu, M.P.H., Zhitao Ma, M.Med., Yanping Zhang, M.Med., Guoqing Shi, M.P.H., Tommy T.Y. Lam, Ph.D., Joseph T. Wu, Ph.D., George F. Gao, D.Phil., Benjamin J. Cowling, Ph.D., Bo Yang, M.Sc., Gabriel M. Leung, M.D., and Zijian Feng, M.Med.

RESULTS

Among the first 425 patients with confirmed NCIP, the median age was 59 years and 56% were male. The majority of cases (55%) with onset before January 1, 2020, were linked to the Huanan Seafood Wholesale Market, as compared with 8.6% of the subsequent cases. The mean incubation period was 5.2 days (95% confidence interval [CI], 4.1 to 7.0), with the 95th percentile of the distribution at 12.5 days. In its early stages, the epidemic doubled in size every 7.4 days. With a mean serial interval of 7.5 days (95% CI, 5.3 to 19), the basic reproductive number was estimated to be 2.2 (95% CI, 1.4 to 3.9).

STATISTICAL ANALYSIS

The epidemic curve was constructed by date of illness onset, and key dates relating to epidemic identification and control measures were overlaid to aid interpretation. Case characteristics were described, including demographic characteristics, exposures, and health care worker status. The incubation period distribution (i.e., the time delay from infection to illness onset) was estimated by fitting a log-normal distribution to data on exposure histories and onset dates in a subset of cases with detailed information avail-

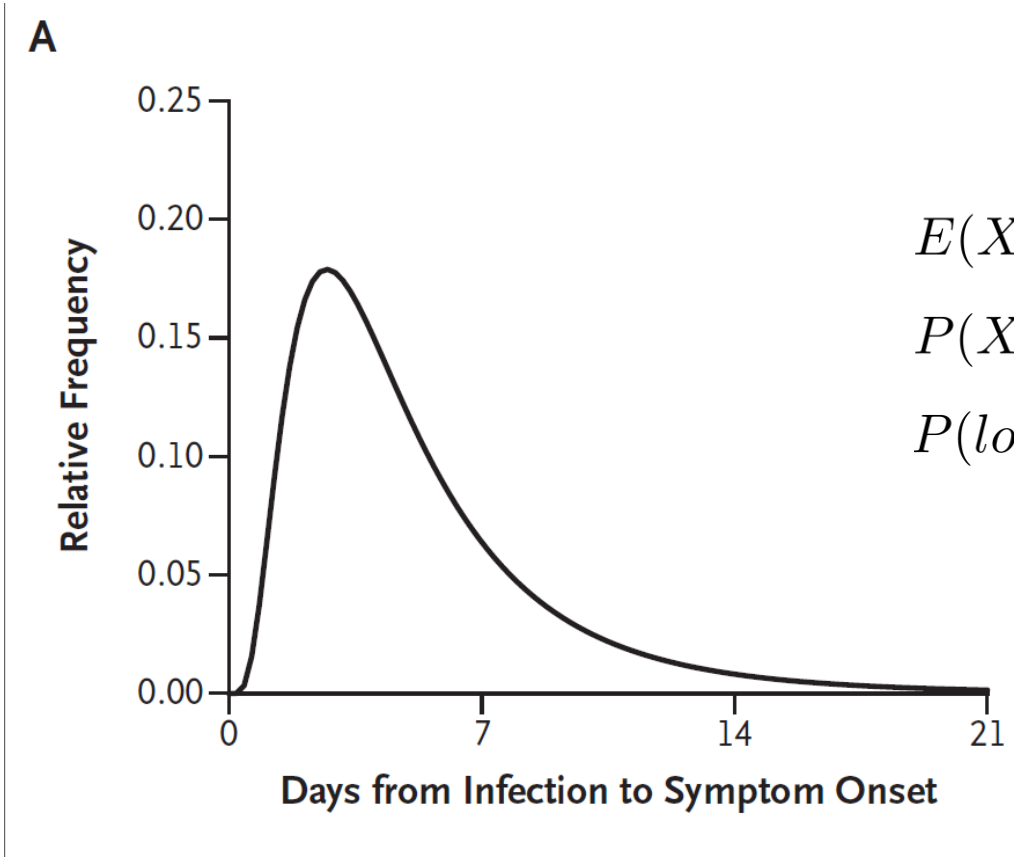
Pere Puig
@PerePuigUAB

Exercici pels nostres estudiants:
Si tenim una distribució lognormal amb mitjana (esperança) igual 5.2, i el percentil 95 és 12.5, quin és el percetil 99?
La resposta és una mica inquietant.

Translate Tweet

4:20 PM · Mar 5, 2020 · Twitter Web App

Direm que la variable aleatòria X segueix una distribució lognormal si el seu logaritme Y=log(X) segueix una distribució normal.



$$\log(X) \sim N(\mu, \sigma^2)$$

$$E(X) = \exp(\mu + \tfrac{\sigma^2}{2}) = 5.2$$

$$P(X \leq 12.5) = 0.95$$

$$P(log(X) \leq log(12.5)) = 0.95$$

$$\mu + 1.645\sigma = \log(12.5)$$

$$\mu = 1.42,\ \sigma = 0.67$$

Les conclusions serien que,

- El percentil 99: $\exp(\mu + 2.326\sigma) = 19.6$
- L'eficàcia d'una quarantena de 14 dies:

$$P(Z \leq \frac{\log(14) - \mu}{\sigma}) = 0.966$$

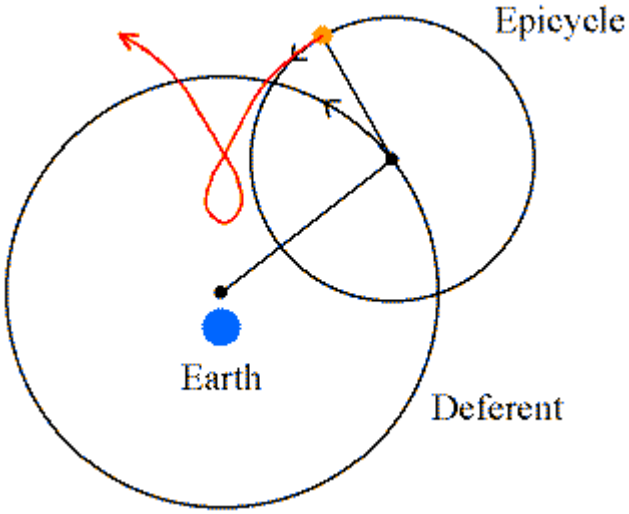
- 2-14 days represents the current official estimated range for the novel coronavirus COVID-19.
- However, a case with an incubation period of **27 days** has been reported by Hubei Province local government on Feb. 22 [12]
- In addition, a case with an incubation period of **19 days** was observed in a JAMA study of 5 cases published on Feb. 21. [13]
- An outlier of a **24 days incubation period** had been for the first time observed in a Feb. 9 study.[11] WHO said at the time that this could actually reflect a second exposure rather than a long incubation period, and that it wasn't going to change its recommendations.
- Period can **vary greatly** among patients.

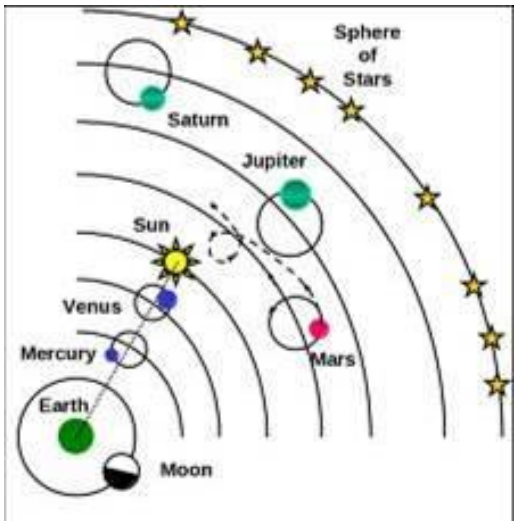
Per què la distribució lognormal?

- És una pràctica habitual en Epidemiologia

- Altres distribucions habituals per aquests tipus de dades: Weibull, Gamma

- De moment no hi ha raons “físiques” o “biològiques” que justifiquin cap d’aquestes distribucions. És una decisió empírica.





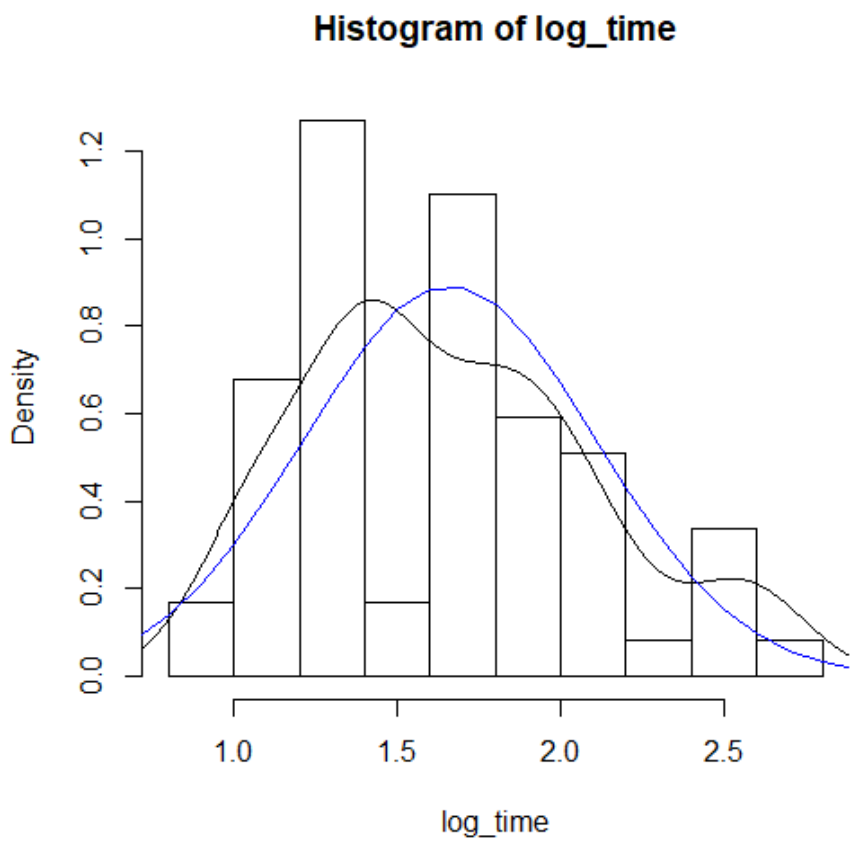
Un model matemàtic (estadístic) no té perquè ser veritable per a ser útil…però millor si ho és!

Estimate the incubation period of coronavirus 2019 (COVID-19)

Ke Men[1], Xia Wang[2], Yihao, Li[3], Guangwei Zhang[1], Jingjing Hu[1], Yanyan Gao[1], Henry Han*[4]

1. Institute for Research on Health Information and Technology, School of Public Health, Xi'an Medical University, Xi'an, Shaanxi 710021, China
2. The Air Force Military Medical University, Xi'an, Shaanxi 710032, China
3. Business Analytics, Fordham University, Lincoln Center, New York, NY 10023, USA
4. Computer and Information Science, Fordham University, Lincoln Center, New York, NY 10023, USA

Els autors publiquen els temps d'incubació de 59 pacients (enquestes epidemiològiques) en 10 regions a la Xina, des del Desembre de 2019 fins el Febrer de 2020.



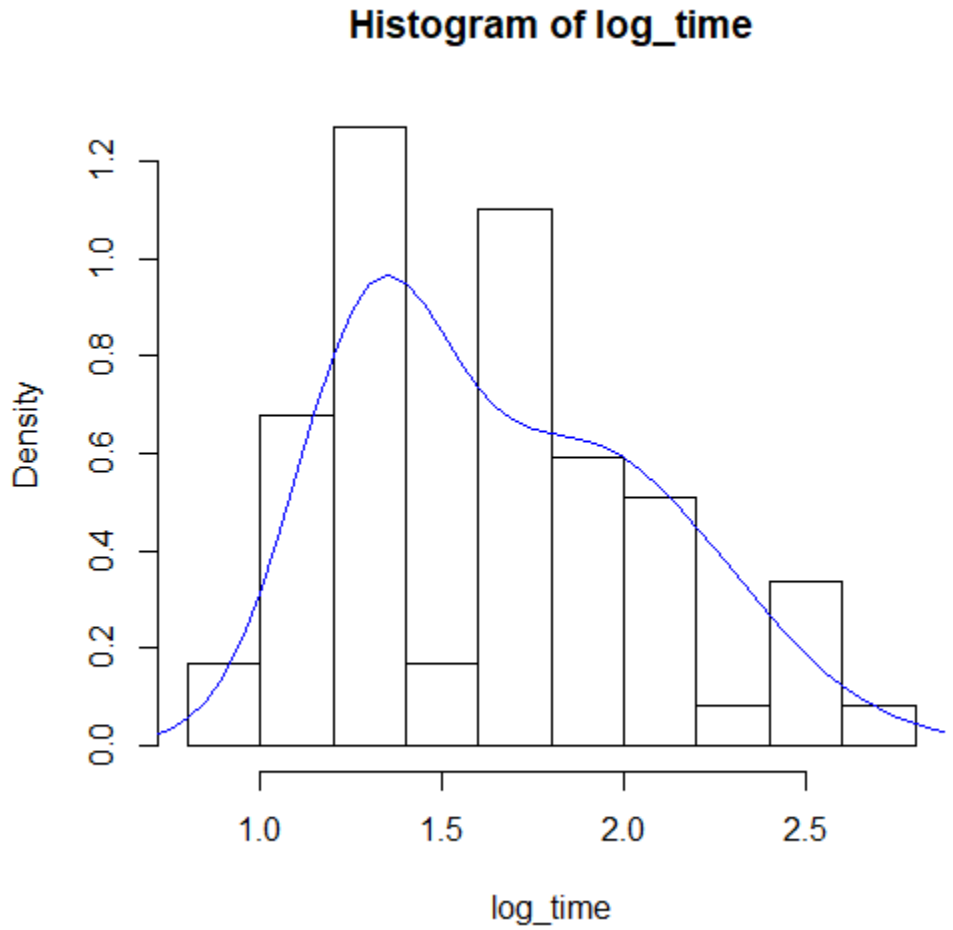

```
        Shapiro-Wilk normality test

data:  log(data$Incubation)
W = 0.94824, p-value = 0.01395
```

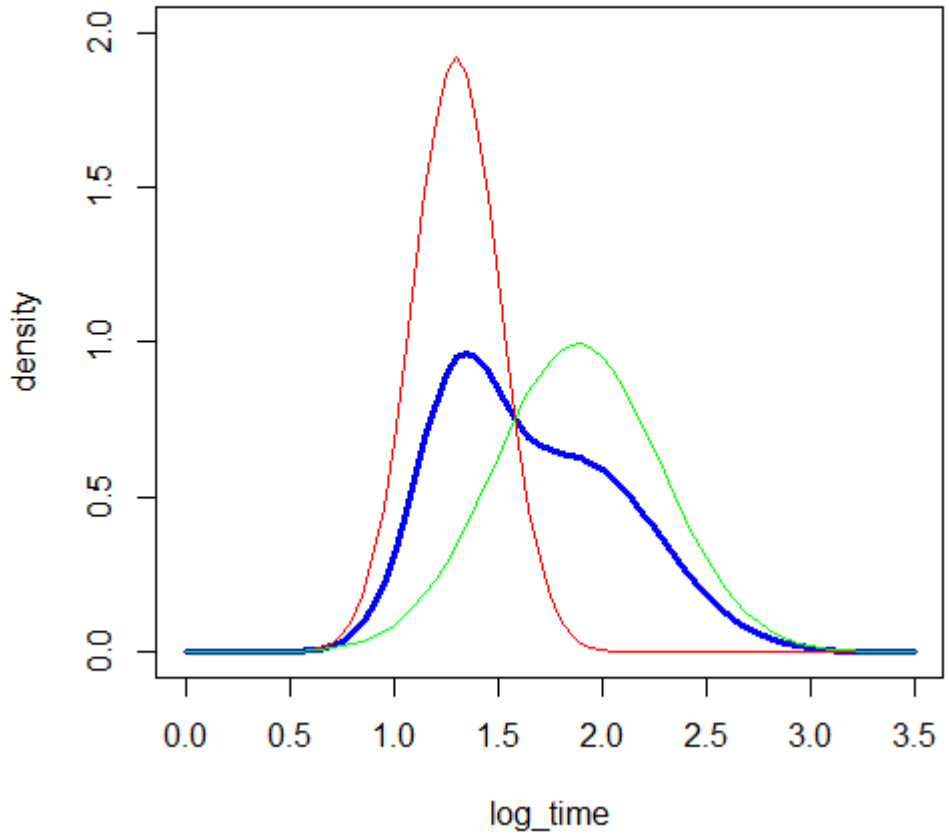
Una mixtura o barreja de distribucions normals podria funcionar millor.

$$X = \begin{cases} X_1 & \text{amb probabilitat} \quad \omega \\ X_2 & \text{amb probabilitat} \quad 1-\omega \end{cases}$$

$$X_i \sim N(\mu_i, \sigma_i^2)$$

A més, pot ser molt interpretable!


Histogram of log_time
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.0
1.5
2.0
2.5
log_time



$\hat{\mu}_1 = 1.30,\ \sigma_1 = 0.21$
$\hat{\mu}_2 = 1.88,\ \sigma_2 = 0.40$

$\omega = 0.38$
$1 - \omega = 0.62$

Conseqüències

- Hi ha dos grups entre la població, on el temps d'incubació del virus és comporta de manera diferent.

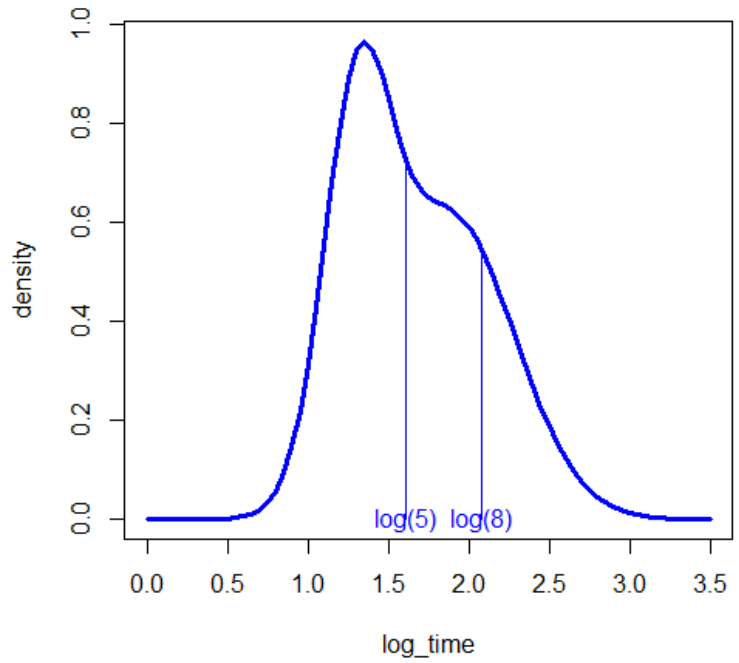
- En el 38% de la població, el temps mitjà d'incubació és de 3.7 dies i en el 62% és de 7.1 dies.

- El temps de quarantena hauria de ser diferent per els dos grups: 5.4 dies pel primer grup i 14 pel segon (marge del 3%).

Problema: cóm detectar a priori als incubadors ràpids?

Dades amb censura d'interval

Data collected for this study included region, age, gender, exposure history, and illness onset. For those cases whose incubation periods locate in an interval $[x_1, x_2]$, we use its midpoint $\delta = \frac{x_1+x_2}{2}$ to represent its incubation period. For example, Case no. 2 in our dataset went on a business trip in Wuhan on Jan 12th, 2020 and returned to Shaanxi on Jan 15th, 2020, but had fever symptoms on Jan 20th 2020. The incubation will be calculated as $\delta = \frac{(20-12)+(20-15)}{2} = 6.5$ days. More details about the dataset can be found in the



Per a estimar bé els paràmetres cal utilitzar tècniques de l'Anàlisis de Supervivència

3- Tests massius?

COVID-19: nuevo método de Israel para pruebas múltiples

Científicos del Instituto Technion y el Centro Médico Rambam presentaron una novedosa solución para evaluar a más de 60 pacientes al mismo tiempo.

El profesor Roy Kishony, jefe del grupo de investigación de biología del Technion, afirmó que el nuevo procedimiento agrupa múltiples muestras en un solo tubo de ensayo. Estas muestras se examinan luego con el sistema de análisis de PCR común, algo que tarda varias horas.

Governor Ricketts: State working to expand coronavirus testing

"Instead of taking each one of those test tubes and testing them separately, we'll take five of those test tubes and put them into one and test that one test tube; and if that one sample comes back negative we know all five of those were negative, and so we have just saved four tests," the governor said.

Ricketts said, if the mixed sample tests positive, they'll have to retest each sample. However, he said, with tests more likely to be negative, the strategy is saving time and resources.

Aquest és l'anomenat "Dorfman two-stage group testing procedure" que es va aplicar durant la segona guerra mundial i que es recollit en el famós llibre de probabilitats del Feller.

Dorfman, Robert (1943). The Detection of Defective Members of Large Populations, *The Annals of Mathematical Statistics*, **14**(4), p. 436–440

Suposem que barrejem n (n>1) mostres biològiques i que la prevalença de la malaltia és p. Sigui X la variable aleatòria que ens indica el número total de PCRs que haurem de fer:

$$X = \begin{cases} 1 \text{ si tots estan sans} \\ n+1 \text{ si almenys un està malalt} \end{cases}$$

$$P(X = 1) = (1-p)^n$$

$$P(X = n+1) = 1 - (1-p)^n$$

Quin serà el nombre esperat de PCRs?

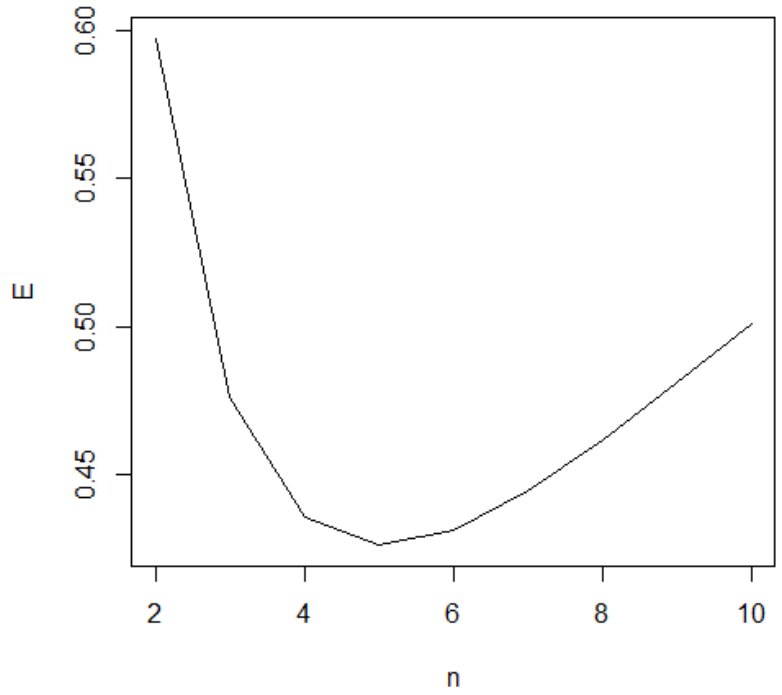
$$E(X) = (1-p)^n + (n+1)(1-(1-p)^n)$$
$$= 1 + n(1-(1-p)^n)$$

Si volem fer tests a un gran grup d'individus (mida N), els dividirem en grupets de mida n i el nombre esperat de PCRs que haurem de fer és,

$$\frac{N}{n}E(X) = N(\frac{1}{n} + 1 - (1-p)^n)$$

Donada la prevalença p, quin valor de n minimitza el nombre esperat de PCRs?

Per exemple, si p=0.05 aquests són els valors esperats per a diversos valors de n:



Per n=5 el valor esperat és de E=0.43

Per tant, per cada 100 individus ens estalviem 57 PCRs!!

TABLE I

Optimum Group Sizes and Relative Testing Costs for Selected Prevalence Rates

Prevalence Rate (per cent)	Optimum Group Size	Relative Testing Cost	Percent Saving Attainable
1	11	20	80
2	8	27	73
3	6	33	67
4	6	38	62
5	5	43	57
6	5	47	53
7	5	50	50
8	4	53	47
9	4	56	44
10	4	59	41
12	4	65	35
13	3	67	33
15	3	72	28
20	3	82	18

Dorfman, Robert (1943). The Detection of Defective Members of Large Populations, *The Annals of Mathematical Statistics*, **14**(4), p. 436–440

nature

NEWS · 10 JULY 2020

The mathematical strategy that could transform coronavirus testing

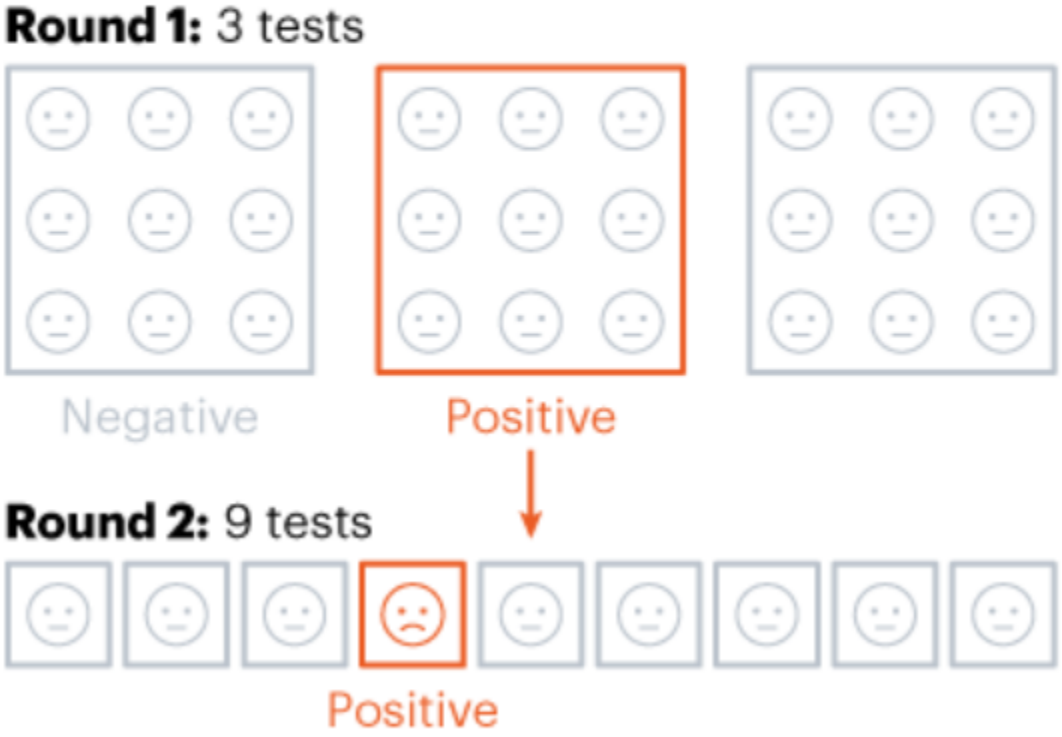
Four charts show how pooling samples from many people can save time or resources.

GROUP TESTING

Countries can save time and money by testing many people at once. Researchers are trialling various methods for group testing.

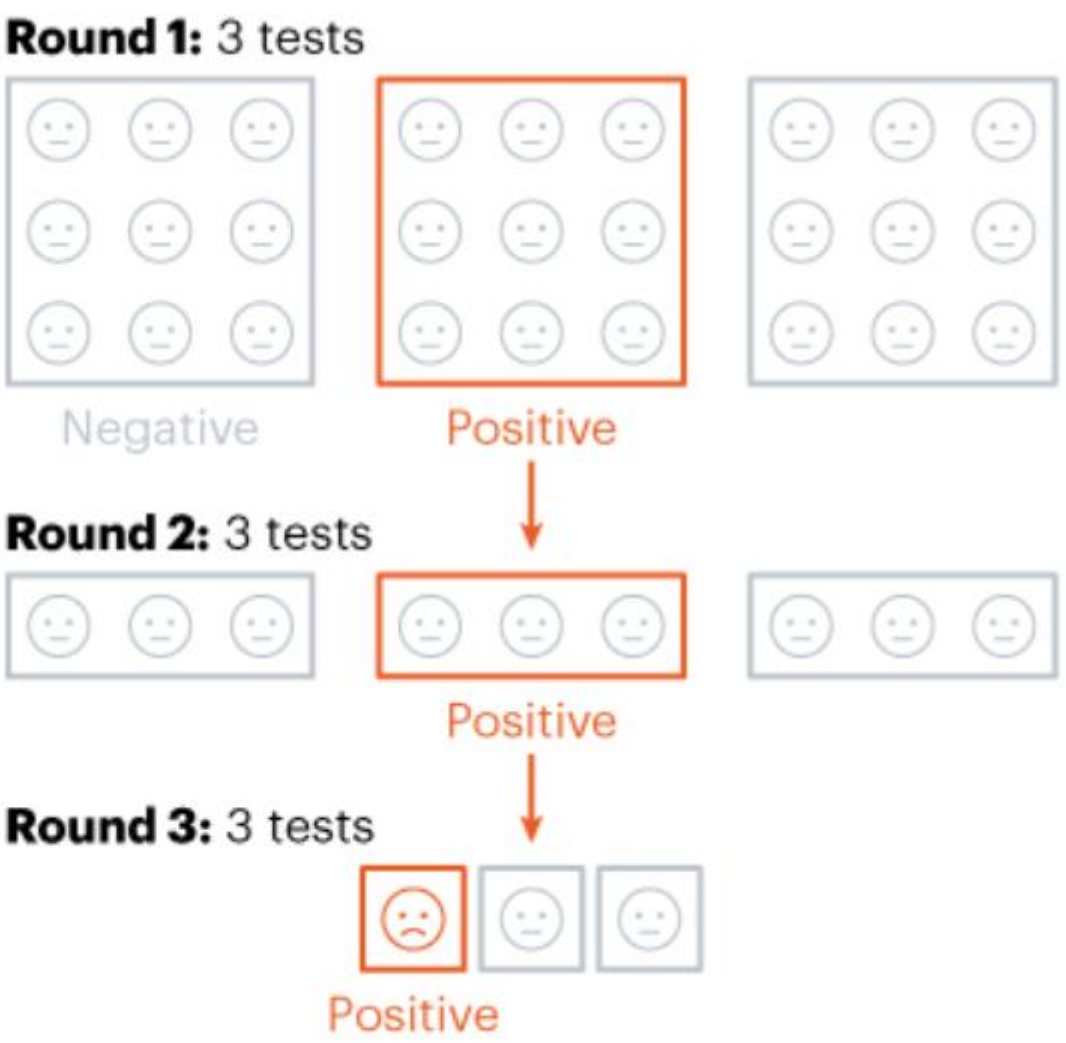
Method 1

Samples are mixed together in equal-sized groups and tested. If a group tests positive, every sample is retested individually.



Method 2

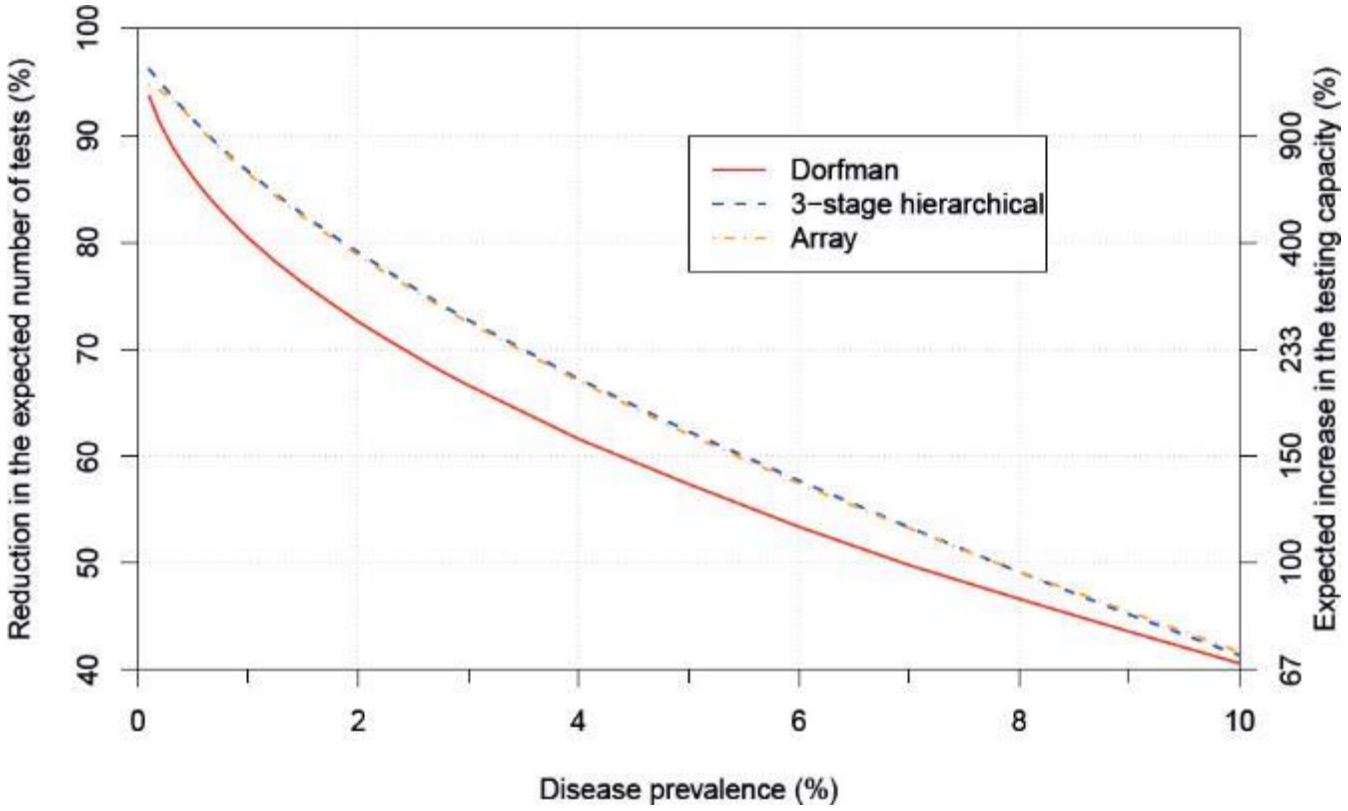
This strategy adds extra rounds of group testing to method 1, reducing the total number of tests needed.



Tests in short supply? Try group testing

Christopher R. Bilder, Peter C. Iwen, Baha Abdalhamid, Joshua M. Tebbs, Christopher S. McMahan

First published: 27 May 2020 | https://doi.org/10.1111/1740-9713.01399



Estem treballant per a poder utilitzar aquesta metodologia a l'Aragó.

Dr. Ivan Galindo
Dra. Gema Marín

Dra. Anna Alba (CRESA)

Com que hi ha tanta grip, han hagut de clausurar la Universitat. D'ençà d'aquest fet, el meu germà i jo vivim a casa, a Palafrugell, amb la família. Som dos estudiants desvagats…

El Quadern Gris, Josep Pla

Gràcies per la vostra atenció !