Comments on a Class of Covariance Structures Derived from Discrete Bivariate Distributions

Marlos Viana

University of Illinois at Chicago
viana@uic.edu
https://symmetrystudies.blogspot.com
https://sites.google.com/view/magviana/home

Celebrating 40 years of the Greek Statistical Institute
(1981-2021)
March 26-28, 2021
(slightly annotated revision)

Celebrating 40 years of the Greek Statistical Institute (1981-2021)

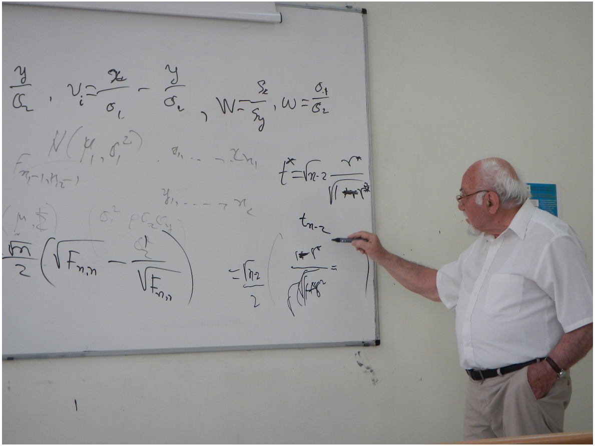
This is a special reunion celebrating the Institute's 40 years of activities promoting statistical research and education - in a few of which I was honored to have participated, going back to the Kastoria 2006 meeting.

However, it is the lasting friendship that I was able to develop with so many of you over the years that I must never forget to remark.

I must also say a special thank you to Prof. Alexis Karagrigoriou for the invitation and for organizing the event in collaboration with his team and the Institute's office. I thank you all.

Prof. Theophilos Cacoullos

Therefore, we are also celebrating the life of our dear friend Prof. Cacoullos whose contributions to statistical science set the standard of excellence for the Institute through his leadership and continued dedication.



Comments on a Class of Covariance Structures Derived from Discrete Bivariate Distributions

These comments are related to a 2020 seminar about statistical assessments of biomedical tests[1] at the end of which an experienced epidemiologist/cardiologist raised the question- in his own words- of whether **the sensitivity of a screening test was dependent on the disease prevalence**.

The heated ensuing discussion confirmed once again that *a person's knowledge is (socially) effective only to the extent that it can be communicated* to those who may benefit from it.

Let me illustrate a possible cause of our inability to communicate with each other, a necessary component of the usefulness of our profession:

[1] The slides are linked here.

Context: Statistical Assessment of Screening Tests

The **clinician** consulting the CDC *Principles of Epidemiology* would understand:

Sensitivity as the ability of surveillance [test] to detect the health problem that it is intended to detect; A/(A+C)

Prevalence as the proportion of persons in a population who have a particular disease. (A+C)/N.

	$T+$	$T-$	
$D+$	A	C	A+C
$D-$	B	D	B+D
	A+B	C+D	N

Context: Statistical Assessment of Screening Tests

Whereas for the **statistician**:

$$\eta_1 = P(T+ \mid D+), \;\; \theta_1 = P(T- \mid D-), \;\; \pi_1 = P(D+)$$

$$\eta_2 = P(D+ \mid T+), \;\; \theta_2 = P(D- \mid T-), \;\; \pi_2 = P(T+)$$

these are **conditional and marginal** probabilities.

Annotation: η_1 is the test sentitivity, θ_1 its specificity, π_1 the condition prevalence. Replacing T with D in the formulas above we get η_2 is the test predictive value positive, θ_2 its predictive value negative, and π_2 the test positivity.

The Parts of the Problem:

Owing to the statistician:

- formulating;
- deriving;

the joint **statistical dependence** structure of

$$\tau \equiv ((\underline{\eta_1}, \theta_1, \underline{\pi_1}), (\eta_2, \theta_2, \pi_2)) \equiv (\tau_1, \tau_2)$$

Owing to all scientists:

- Explaining the results to all who can benefit from them.

A Multinomial / Dirichlet Formulation:

The joint frequency counts and underlying probabilities by:

$$x = \begin{bmatrix} x_0 & x_1 \\ x_2 & x_3 \end{bmatrix}, \quad q = \begin{bmatrix} q_0 & q_1 \\ q_2 & q_3 \end{bmatrix},$$

The underlying multinomial and Dirichlet models:

$$x \mid q \sim M_3(q, n), \quad q \sim D_3(a)$$

The corresponding estimates:

$$(\frac{x_0}{x_0 + x_1}, \frac{x_3}{x_2 + x_3}, \frac{x_0 + x_1}{n}) \mapsto (\frac{q_0}{q_0 + q_1}, \frac{q_3}{q_2 + q_3}, q_0 + q_1) = \tau_1.$$

Since τ_1 differs from τ_2 only by a transposition (12) we can write $\tau = (\eta, \theta, \pi)$ when no distinction is needed. This is equivalent to replacing T with D in the original tables, or replacing rows with columns.

A Multinomial / Dirichlet Formulation:

Under $q \sim D_3(a)$ it is well-known[2] that there is a 1-1 mapping ϕ taking q in the simplex $\sum q_j = 1$ to $\phi(q) = (\eta, \theta, \pi)$ in $[0,1]^3$ and that its Dirichlet distribution factors into the corresponding beta (more generally Dirichlet) marginals:

$$(\eta, \theta, \pi) \sim D_1^{\eta}(\phi_\eta(a)) D_1^{\theta}(\phi_\theta(a)) D_1^{\pi}(\phi_\pi(a))$$

so that we obtain the stochastic independence of (η, θ, π) or of the their corresponding odds,

$$\boxed{\eta \perp \theta \perp \pi} \quad \mathcal{O}_\eta \perp \mathcal{O}_\theta \perp \mathcal{O}_\pi$$

and since $q \mid x \sim D_3(x + a)$ we also obtain their posterior stochastic independence

$$\boxed{(\eta \mid x) \perp (\theta \mid x) \perp (\pi \mid x)}$$

[2] e.g., K. Fang, S. Kotz and K. Ng Symmetric Multivariate and Related Distributions, or M. Bishop, S. Fienberg, P. Holland Discrete Multivariate Analysis.

Cross-Covariances:

The covariance between each component of τ_1 and each component of τ_2 follows directly from the factorization of τ or $\tau \mid x$. Its is sometimes easier to work with the odds instead of the original probabilities, so that, for example,

$$\mathcal{O}_{\eta_1} = \frac{\eta_1}{1-\eta_1}, \;\; \mathcal{O}_{\eta_2} = \frac{\eta_1 \pi_1}{(1-\theta_1)(1-\pi_1)}$$

so that

$$\begin{aligned} E(\mathcal{O}_{\eta_1}\mathcal{O}_{\eta_2}) &= E_{D_1(\eta)}[\eta^2(1-\eta)^{-1}] \\ &\times E_{D_1(\pi)}[\pi(1-\pi)^{-1}] \\ &\times E_{D_1(\theta)}[(1-\theta)^{-1}] \\ &= \text{product of quotients of beta functions.} \end{aligned}$$

etc ...and similarly for the cross-covariances of $\mathcal{O} \mid x$.

Numerically:

The following are Monte Carlo estimates for the joint correlation structure of $(\eta_1, \theta_1, \pi_1, \eta_2, \theta_2, \pi_2)$ under the Dirichlet formulation for a test data

	$T+$	$T-$	
$D+$	80	20	100
$D-$	10	30	40
	90	50	140

The size of the MC is 10^6.

Numerically:

Cor $(\eta_1, \theta_1, \pi_1, \eta_2, \theta_2, \pi_2 \mid x)$ with $a = (1, 1, 1, 1)$:

	η_1	θ_1	π_1	η_2	θ_2	π_2
η_1	1.00	0.00	0.00	0.16	0.68	0.71
θ_1	0.00	1.00	0.00	0.80	0.33	-0.49
π_1	0.00	0.00	1.00	0.57	-0.64	0.50
η_2	0.16	0.80	0.57	1.00	0.00	0.00
θ_2	0.68	0.33	-0.64	0.00	1.00	0.00
π_2	0.71	-0.49	0.50	0.00	0.00	1.00

Note that since sensitivity η_1 and positivity $\pi_2 = P(T+)$ are correlated, perhaps this is the dependence that the clinician *experienced* in his decades of practice and understood as the dependence between sensitivity and prevalence $\pi_1 = P(D+)$. If is also opportune to observe that the test positivity is also referred in some texts as the *apparent* prevalence.

The Challenging Part:

How to *explain* any (prior) stochastic independence

$$\boxed{\eta \perp \theta \perp \pi} \quad \mathcal{O}_\eta \perp \mathcal{O}_\theta \perp \mathcal{O}_\pi,$$

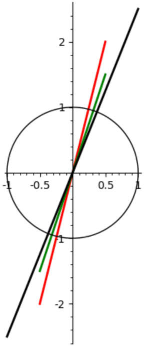
their (posterior) conditional stochastic independence

$$\boxed{(\eta \mid x) \perp (\theta \mid x) \perp (\pi \mid x)},$$

or their corresponding cross-dependencies to those who may benefit from that knowledge?

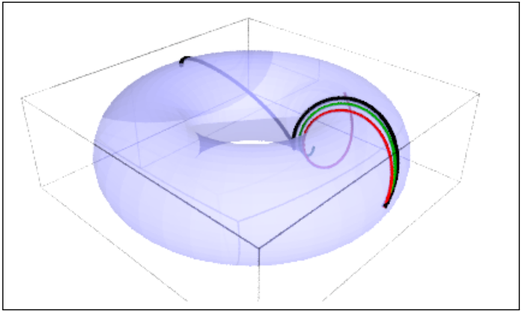
A Quick Detour: $\mathbb{P}^1 \mapsto S^1$ Representations.

The odds-probability correspondence allows for understanding the rows (x_0, x_1), (x_2, x_3) and row sums $(x_0 + x_1, x_2 + x_3)$ as homogeneous coordinates and for their representation $v \in \mathbb{P}^1 \mapsto v/||v|| \in S^1$: $(80, 20) \simeq (4 : 1)$ in red, $(30, 10) \simeq (3 : 1)$ in green and $(100, 40) \simeq (5/2 : 1)$ in black.



$\mathbb{P}^1 \simeq S^1/\mathbb{Z}_2$ and more generally $\mathbb{P}^{n-1} \simeq S^{n-1}/\mathbb{Z}_2$ the projective space of lines through the origin of $\mathbb{R}^n$.

$\mathbb{P}^1 \mapsto S^1 \times S^1$ Representations.



There are several forces that pull us **into** and **away** from such directions. The remaining questions is for whose benefit.

The Multinomial-Dirichlet Formulation:

Under $x \mid q \sim M_3(q)$ it is also known that

$$(\frac{x_0}{x_0+x_1}, \frac{x_3}{x_2+x_3}, \frac{x_0+x_1}{n}) \mid \sigma \sim Bin(\eta)Bin(\theta)Bin(\pi)$$

so that (in short notation) under $q \sim Beta(\eta)Beta(\theta)Beta(\pi)$ we obtain the stochastically independent predictive beta-binomials

$$\boxed{\frac{x_0}{x_0+x_1} \perp \frac{x_3}{x_2+x_3} \perp \frac{x_0+x_1}{n}}$$

for the test's sensitivity and specificity and the condition prevalence, and symmetrically after the action of (12):

$$\boxed{\frac{x_0}{x_0+x_2} \perp \frac{x_3}{x_1+x_3} \perp \frac{x_0+x_2}{n}}$$

for the test's pvp, pvn and positivity.

The Challenging Part:

How to *explain*

$$\boxed{\frac{x_0}{x_1} \perp \frac{x_3}{x_2} \perp \frac{x_0 + x_1}{x_2 + x_3}}$$

$$\boxed{\frac{x_0}{x_2} \perp \frac{x_3}{x_1} \perp \frac{x_0 + x_2}{x_1 + x_3}}$$

and their cross-dependencies to those who may benefit from this knowledge?

The Few Outstanding Questions:

- What does my colleague cardiologist know that I could not grasp?
- Does he have a perception of *dependence* between events that our mathematics failed to apprehend?
- Did his decades of contact with the events of disease and testing block/expand his understanding of the processes involved?
- Are we to blame the mutual difficulty in explanation/understanding to the years of medical/statistics training we impose to our students?

Some Precedents:

Dennis Lindley once said[3] that *we statisticians face the risk of becoming like geometers who study space and never measure it*.

If statistics is an integral part of the scientific method, then we must be aware that it cannot adequately be discussed if it is divided from the science to which it applies[4].

Again, it is by reaching out to the unity of science (the collective of statistical reasoning) that we may better see the natural dependency among statistics, probability and experimentation[5].

[3] Plenary address to the 4th Valencia International Meeting on Bayesian Statistics, Spain, 1991.

[4] Gower, B. (1997), Scientific Method - An Historical and Philosophical Introduction, Routledge, London, U.K

[5] Abstracted from Viana,M and Jovanovic, B LIAISON Vol.14 No.3, 40-44, Statistics Society of Canada, 2000.

By now I am sure that Prof. Cacoullos would have brought us all together debating his views of these challenges....

...only to shortly after invite us to go shopping with him ...

...where he would cheerfully question the origin of each single item on the stand !!!

At the end of the day, however, he would keep us all united with his spirited presence !!!

Prof. Theophilos Cacoullos (1932-2020)

And we would be ready for another day !!!

Thank you very much !!

Last but not Least:

(this slide was reserved for potential follow up questions and was not presented)

> *Science is a social activity, and the means by which it is pursued are a matter for negotiation between scientists, and between them and those other members of their society who have some control over what they do. Accordingly, the methods of science are grounded in the needs and interests of the particular and different societies within which scientists work; they are not, as the philosophers would have us believe, grounded in universal requirements of rationality. It is, in short, not universal reason but the conventions agreed in particular societies which determine the legitimacy of the reasoning used by scientists in that society* [1].

See also [2].

References

[1] B. Gower, *Scientific method - an historical and philosophical introduction*, Routledge, London, U.K., 1997.

[2] Alfred W. Crosby, *The measure of reality*, Cambridge University Press, Cambridge, 1997. Quantification and Western society, 1250–1600. MR1679885