Principios básicos de estadística Multidimensionalidad, variabilidad y causalidad

Carlos J. Gil Bellosta cgb@datanalytics.com

Mayo 2013

El error de tipo III: respuesta exacta, pregunta equivocada

Error de tipo III

Consiste en dar la respuesta exacta a la pregunta equivocada.

Error de tipo III, según Tukey

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

John Tukey, 1962

Errores de tipo I y II

- Error de tipo I: se concluye que algo sucede cuando, realmente, no sucede.
- Error de tipo II: se concluye que algo no sucede cuando, reamente, sucede.

Tukey sobre el error de tipo I

The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.

John Tukey, 1984

Mejores respuestas a mejores preguntas: por un sano escepticismo

El objetivo último de la charla es conseguir que

- aprendamos a matizar nuestras preguntas y
- resistamos la tentación a dar malas respuestas.

Para ello repasaremos tres aspectos de la realidad que la *psique* humana no gestiona adecuadamente y que *entorpecen* su quehacer cuantitativo:

- Multidimensionalidad
- Variabilidad
- Causalidad

Ejercicio mental: ¿qué casa me compro?

- Muchas variables a tener en cuenta
- ¿Las sopesamos todas adecuadamente?

La reducción de la dimensionalidad como heurística

The class of one-good-reason heuristics orders cues, finds the first one that allows a decision to be made, and then stops and ignores all other cues.

G. Gigerenzer & H. Brighton, 2008

La reducción de la dimensionalidad es un atajo mental para

- comprender (¿explicar?) la realidad y
- tomar de decisiones.

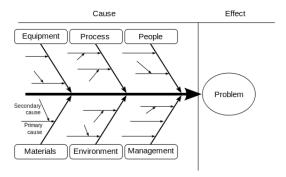
La realidad es poliédrica

- Hace un tiempo, la palabra poliédrico se puso de moda entre políticos, etc., para referirse a fenómenos complejos
- Se refiere a los muchos puntos de vista (variables) que afectan a muchos problemas, particularmente sociales
- Pero, ¿qué queremos decir con poliédrico en contextos no meramente retóricos?



Fuente: Wikipedia

Poliédrico: muchas variables involucradas



Fuente: Wikipedia (fish bone diagram)

¡Trata de publicar esto en el periódico!

```
Residuals:
   Min
           1Q Median
                          3Q
                                 Max
-4.3054 -1.5585 -0.1539 1.7687 5.3592
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)
                  1.517e+00 6.635e-01 2.287
                                              0.0227 *
disturbLT-SEL
                  3.248e-01 8.653e-01 0.375
                                              0.7076
disturbSETTLE
                1.688e+00 1.546e+00 1.092
                                              0.2754
                  2.092e-01 1.338e+00 0.156
                                              0.8758
disturbVIRGIN
elev
                  2.340e-03 5.503e-04 4.252 2.64e-05 ***
disturbLT-SEL:elev 2.112e-05 7.402e-04 0.029
                                              0.9772
disturbSETTLE:elev -2.041e-03 2.073e-03 -0.984
                                              0.3255
disturbVIRGIN:elev -4 867e-04 1 065e-03 -0 457
                                               0.6478
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
Residual standard error: 2.094 on 403 degrees of freedom
Multiple R-squared: 0.1139, Adjusted R-squared: 0.09848
F-statistic: 7.398 on 7 and 403 DF, p-value: 2.249e-08
```

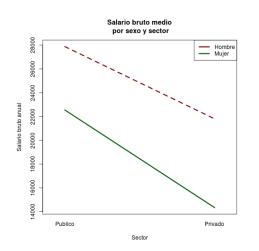
¿Estudiamos la multidimensionalidad usando estadística dura?

¿Estamos preparados para utilizar y entender expresiones como la siguiente?

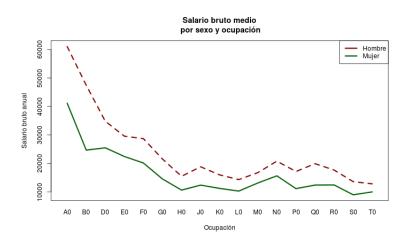
Un estudio de ese tipo aparece en la sección *Análisis multidimensional* del Observatorio Laboral de la Crisis (FEDEA, abril de 2013), http:

//www.fedea.net/observatorios/empleo/pubs/olc-19.pdf

¿Reciben las mujeres menor salario por el mismo trabajo? Visión del sector...

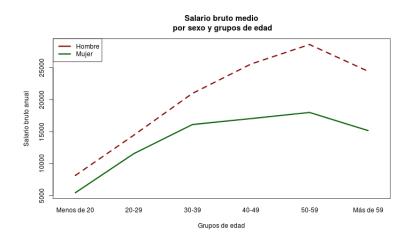


..., de la ocupación,...



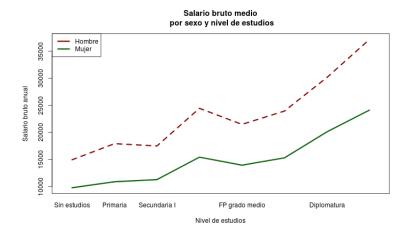
Fuente: http://rpubs.com/joscani/esalarial

... la edad,...



Fuente: http://rpubs.com/joscani/esalarial

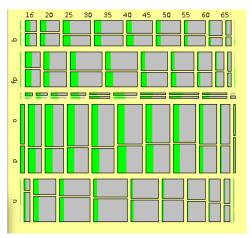
... y el nivel de estudios.



Fuente: http://rpubs.com/joscani/esalarial

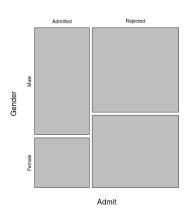
Pero, ¿cómo es la población subyacente?

Radiografía de la población activa española por edad, nivel educativo y sexo.

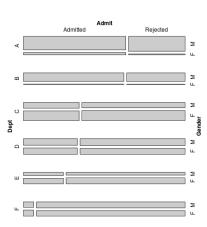


¿Estaban sesgadas las admisiones a Berkeley en 1973?

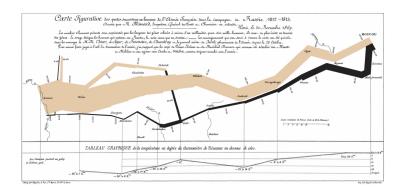
Visión 2D



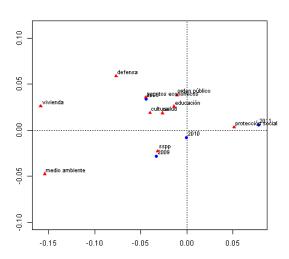
Visión 3D



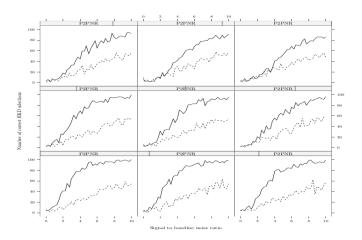
¿Cuántas variables aparecen?



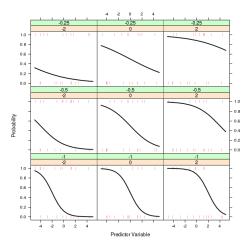
Diez partidas presupuestarias (2008-11)



Celosías (*small multiples*) para partir gráficos por una variable...



... o más de una variable



Fuente: http://casoilresource.lawr.ucdavis.edu/drupal/book/export/html/510

Variabilidad: una cita de N. Taleb

[...] but we are committing crimes against life, the living, science, and wisdom, for the sake of eliminating volatility and variation.

N. Taleb, Antifragility, 2012

¿Qué vemos en la foto?



Vemos España...



... a través de una encuesta



Encuestas y variabilidad

- La variabilidad es consecuencia de la extrapolación (de una muestra a la población entera)
- ¿Quién se comió la varianza? ¿El gato de la redacción?
- ¡Son creíbles los resultados de encuestas con tantas cifras significativas?

Encuestas y variabilidad: el caso de la EPA

A pesar de que la EPA es la encuesta más grande...

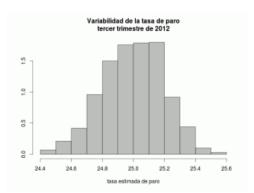
INE, 2012-10-26

El número de parados crece en 85.000 personas y alcanza la cifra de 5.778.100. La tasa de paro se incrementa 38 centésimas hasta el 25,02%.

El País, 2013-04-25

La EPA cifra en 6.202.700 el número de desempleados y la tasa de paro en el 27,16%

Encuestas y variabilidad: el caso de la EPA



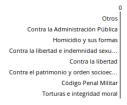
Posibles tasas de paro (tercer trimestre 2012) compatibles con la EPA a nivel nacional (oficialmente, 25.02 %).

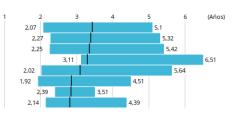
Variabilidad bien entendida

Edad:	(45 años)	Sexo: MUJER	19 julio 2012 09:31	Referencia:	44818980
-------	-----------	-------------	---------------------	-------------	----------

Hematimetría								
Recuento								
Prueba	Resultado		Unidades	Valores de referencia				
Número total leucocitos		7.00	x 10° µ1	(3.5-11)				
Numero total hematies		4.3	x 10^6 µl	(3.5 - 5.8)				
Hemoglobina		13.5	g/dl	(12 - 15)				
Hematocrito		39.3	8	(36 - 43)				
Volumen corpuscular medio sangre		91.4	fl	(80 - 100)				
Hemoglobina corpuscular media		31.2	pg	(27 - 32)				

Variabilidad bien entendida

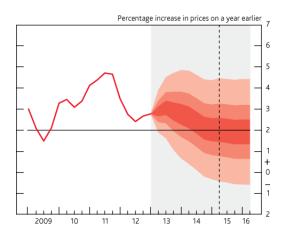




Fuente: El Indultómetro,

http://www.elindultometro.es/2013/02/27/los-mas-rapidos-a-este-lado-de-los-pirineos.html

Banco de Inglaterra: proyecciones de inflación



Fuente: Banco de Inglaterra, Overview of the Inflation Report May 2013

Variabilidad y proyecciones

Proyecciones de la población española para los próximos diez años a lo largo de los últimos cuatro (INE):

	2018	2019	2020	2021
2008	49,0			
2009	46,8	46,9		
2010	47,0	47,1	47,2	
2011	45,8	45,7	45,6	45,5

Prediction is very difficult, especially about the future.

N. Bohr, aparentemente

¿Cómo no hacer predicciones?

Ilustración 17-Distribución de frecuencias de desvíos en las previsiones (peaje España⁴⁵)



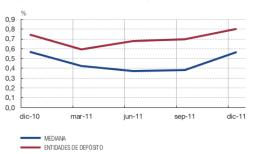
Escenarios, escenarios (y fatalismo)

- Al proyectar se formulan hipótesis: fórmulas, parámetros, etc.
- Una proyección equivale a hacer una única selección de hipótesis
- ¡Hay que crear escenarios distintos bajo juegos distintos de hipótesis! Como mínimo, tres.
- P.e., escenario optimista, pesimista, neutral.
- Además, así se lucha contra el fatalismo
- Y más, ¿cómo hablar (p.e., en la prensa) sobre proyecciones?
 ¿Usaríamos el condicional?

¿Quién se le comió la varianza al BdE?

PÉRDIDAS POR DETERIORO DE ACTIVOS

A. EVOLUCIÓN DE LAS PÉRDIDAS POR DETERIORO DE ACTIVOS (PROVISIONES ESPECÍFICAS Y GENÉRICAS) SOBRE ATM



Fuente: Informe de estabilidad financiera, Banco de España, abril de 2012

Causalidad: otra cita de N. Taleb

In the complex world, the notion of "cause" itself is suspect; it is either nearly impossible to detect or not really defined —another reason to ignore newspapers, with their constant supply of causes for things.

N. Taleb, Antifragility, 2012

El reloj, el gato y Madagascar

- Reloj: lo podemos armar, desarmar y vuelve a funcionar
- Gato: lo podemos desarmar, pero una vez desarmado no podemos hacerlo volver a funcionar
- Madagascar: ni siquiera podemos desarmarlo

Fuente: J.L. Sampedro. *El reloj, el gato y Madagascar*, Revista de estudios andaluces (1983).

Ciencias reloj: ingeniería, física, química,...

- Pueden probar distintos efectos sobre conjuntos similares de piezas
- Si se rompen, ¡no pasa nada!
- Permiten indentificar causas: $A \Rightarrow B$
- Más aún, permiten predecir: ¿a qué presión cederá la viga a tal temperatura?

Ciencias gato: medicina, biología,...

- Abundan los estudios observacionales (diseño no controlado por el experimentador)
- Los estudios experimentales tienen limitaciones:
 - Se pueden hacer con animales, pero...
 - Incluso en ese caso, a menudo, no se pueden hacer iteraciones
 - En los estudios con personas existen consideraciones que impiden que puedan hacerse *propiamente*
- Resultado: la causalidad es problemática

Ciencias Madagascar: economía, sociología,...

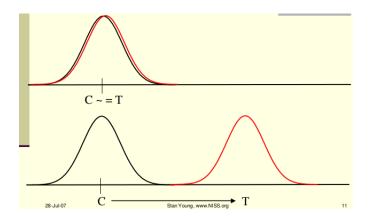
- Prácticamente, no existen los estudios experimentales (sólo observacionales)
- Resultado: la causalidad es muy, muy problemática

If you want to know the effect of one variable, then you need to keep everything else constant and manipulate that one variable.

Cita aproximada de un autor que no recuerdo

Y eso no se puede hacer con Madagascar

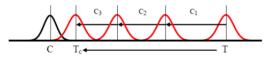
Estudios experimentales: prueban la causalidad



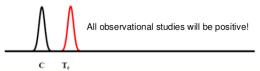
Fuente: S. Stanley Young. Everything is Dangerous: A Controversy, 2008

Estudios experimentales: ¿prueban la causalidad?

(a) Use confounding variables to reduce bias.



(b) As n get large the standard error of the mean gets small.



Fuente: S. Stanley Young. Everything is Dangerous: A Controversy, 2008

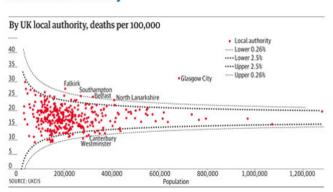
¿Cuáles son los indicios de la causalidad?

Austin Bradford Hill, en *The Environmental Disease: Association or Causation?* (1965) proporciona algunos indicios de causalidad:

- La fuerza de la relación
- La consistencia
- La especificidad
- La temporalidad: ¿qué ocurre antes?
- El gradiente: ¿más A implica más B?
- La plausibilidad
- La analogía: si sabemos que $A \Rightarrow B$ y $A \approx C...$

Variabilidad + causalidad espúrea

Bowel cancer mortality



Muchas gracias y...

¿preguntas?