Variables aleatorias multidimensionales: independencia vs. dependencia, discusión

Antonio Sala

Dept. Ing. Sistemas y Automatica (DISA)
Universitat Politècnica de València (UPV)

Video-presentación disponible en:

http://personales.upv.es/asala/YT/V/**indepdis**.html

Presentación

Motivación:
El objetivo del análisis multivariante es describir la relación entre múltiples variables aleatorias. El caso más sencillo es 2 variables, que pueden estar "relacionadas" o no, o una puede "causar" la otra.

Objetivos:
Comprender los conceptos "prácticos" asociados a la "dependencia" o "independencia" entre dos variables aleatorias y el elevado riesgo de sacar conclusiones incorrectas (falacias).

Contenidos:
Probabilidad condicional. Independencia. Discusión. Conclusiones.

Revisión de conceptos

Probabilidad condicional de y dado x. Probabilidad condicional y fórmula de Bayes:

$$f(y|x) := \frac{f(x,y)}{f_x(x)} \qquad f(y|x)f_x(x) = f(x,y) = f(x|y)f_y(y)$$

Independencia estadística: Vbles. estadísticamente independientes $\Leftrightarrow$ $f(x,y) = f_x(x)f_y(y)$.
Equiv, la "condicional" es igual a la "marginal", y no depende de la variable "medida":

$$f(y|x) = \frac{f(x,y)}{f_x(x)} = f_y(y)$$

Independencia vs. predicción

- Si dos variables son independientes, conocer una no permite estimar la otra: ($x =$"precio KWh en Vladivostok", $y =$"precio del queso Cabrales en Madrid").
- Si son "dependientes" (y se dispone de un modelo que explique esa dependencia, entonces sí se puede mejorar la predicción usando la probabilidad condicional).

Mejor predicción (puntual): Moda o media de probabilidad **condicional**.

¿Independencia/dependencia con datos experimentales?

Probar dependencia/independencia de dos variables con un modelo matemático $f(x, y)$ es fácil. Sólo necesito "factorizar", o hacer comprobaciones como en el ejemplo de los balones.

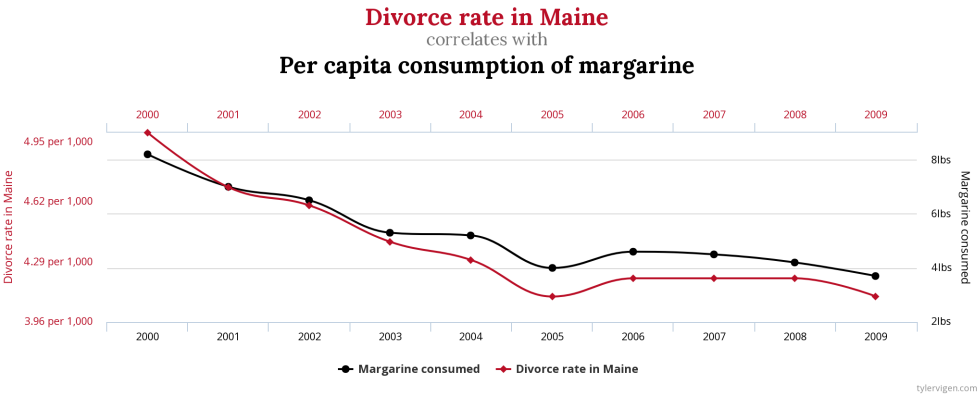
***Probar la "independencia" de variables "experimentales" es complicado...**

¿Son <precio KWh en Vladivostok> y <precio Cabrales en Utiel> independientes? ¿igual se pone de moda en Rusia el Cabrales, y incrementos de población en Vladivostok hacen subir el precio del KWh y la demanda mundial ($\rightarrow$precio) de Cabrales (correlación positiva)?

***Probar la "dependencia" también.**

si consigo una "ley de la naturaleza" o "ecuación" que predice "muchas" muestras de un experimento (mejor que el puro azar, o la media), sospecho que son dependientes... salvo que mi ley tenga 2400 parámetros ajustables y el experimento tenga 1000 datos, con lo que el resultado es "basura", o que los datos ajusten por "casualidad" (ver siguiente transp.).

Conclusiones erróneas por azar: "superstición"



El estado de Maine(US) recomienda incrementar el consumo de margarina entre las parejas casadas.

La recomendación no es cierta, la correlación sí.

Dependencia vs. Causalidad

Dependencia (o "correlación") estadística **NO** implica **causalidad** (relación causa-efecto):

- El evento A ocurre antes en el tiempo que B y, de datos pasados, afirmamos que $p(B|A) > P(B|\neg A)$... ¿decimos que A causa B? ¿Cuando vemos "A", apostamos por "B"? ufff... ¡cuidado!
- El concepto de "causalidad" en sí es complicado de definir (desde Aristóteles llevan los filósofos discutiendo el tema).

Leer: https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

Hay que ser cuidadosos con las conclusiones

- Estudio detecta correlación significativa entre "precio KWh Vladivostok" y "precio Cabrales en Utiel". Tres posibles conclusiones:
  1. ¡No me lo creo! Esto es una pura casualidad, que más datos desmentirán.
  2. Hay una tercera variable que "causa" ambas: lanzamos hipótesis de que están "causados" por el incremento de población en Rusia (y que el Cabrales les encanta a los rusos). Habrá que "probar" o "desmentir" la hipótesis.
  3. ¿Superstición o visionario?: si sube el KWh en Vladivostok, eso causa que suba el Cabrales. Vista la tendencia alcista del KWh, voy a comprar acciones de queserías asturianas.

- Estudio estadístico detecta correlación positiva entre "buena salud" y "consumo moderado de alcohol". Dos posibles conclusiones[1] :
  1. El alcohol es saludable: bebe vino y tendrás salud y felicidad.
  2. El alcohol es perjudicial: las personas con mala salud deben evitar el consumo de alcohol (los enfermos no beben, por prescripción médica).

[1] Depende de quien pague el estudio ("asociación de productores de vino" *versus* "dinero público noruego" (esto es, de país cuyo gobierno no oculte resultados malos para un sector económico importante)), saldrá en prensa una u otra conclusión. Los datos no mienten, las personas que los seleccionan o interpretan sí.

Hay que ser cuidadosos con las conclusiones

- Estudio detecta correlación significativa entre "precio KWh Vladivostok" y "precio Cabrales en Utiel". Tres posibles conclusiones:
  1. ¡No me lo creo! Esto es una pura casualidad, que más datos desmentirán.
  2. Hay una tercera variable que "causa" ambas: lanzamos hipótesis de que están "causados" por el incremento de población en Rusia (y que el Cabrales les encanta a los rusos). Habrá que "probar" o "desmentir" la hipótesis.
  3. ¿Superstición o visionario?: si sube el KWh en Vladivostok, eso causa que suba el Cabrales. Vista la tendencia alcista del KWh, voy a comprar acciones de queserías asturianas.

- Estudio estadístico detecta correlación positiva entre "buena salud" y "consumo moderado de alcohol". Dos posibles conclusiones[1] :
  1. El alcohol es saludable: bebe vino y tendrás salud y felicidad.
  2. El alcohol es perjudicial: las personas con mala salud deben evitar el consumo de alcohol (los enfermos no beben, por prescripción médica).

[1] Depende de quien pague el estudio ("asociación de productores de vino" *versus* "dinero público noruego") (esto es, de país cuyo gobierno no oculte resultados malos para un sector económico importante), saldrá en prensa una u otra conclusión. Los datos no mienten, las personas que los seleccionan o interpretan sí.

Resumen: life is hard

- Podemos inferir conclusiones erróneas (superstición, relación causa→efecto) a partir de unos datos por "casualidad"... dependencia afirmada que es falsa.
- Datos futuros pueden desmentir una dependencia que se suponía cierta (siguientes años de estadística de divorcios en Mayne). Sobre todo en Medicina, Sociología, Psicología. En Física, datos más "precisos" o con más "energía" o "espectro" desmienten teorías.
- Dos variables pueden ser "dependientes" pero no haber encontrado el modo de demostrarlo por falta de "tiempo" (–a– No hemos hecho la estadística Vladivostok/Cabrales) o de "inteligencia" (–b– a nadie antes de Einstein se le había ocurrido $E = mc^2$).

La ciencia progresa por "Matemáticas" –b– o "coleccionismo de sellos" –a–.

Conclusiones

- Variables **independientes**: saber una variable no sirve para decir nada de la otra, **condicional = marginal**.
- Dependencia estadística / correlación NO implica "causa".
- Dependencia/independencia es difícil de probar con muestras ruidosas. La estadística "bien hecha" intenta cuantificar la probabilidad de error.
- Es fácil equivocarse extrayendo conclusiones erróneas de "causalidad" en estudios estadísticos.