EMP Educação Matemática Pesquisa
ISSN 1983-3156
A2
http://dx.doi.org/10.23925/1983-3156.2019v21i1p433-460

Inferência informal e inferência "informal"

Informal and "Informal" Inference

MANFRED BOROVCNIK [1]

Resumen

"Inferencia Informal" es un enfoque de la inferencia estadística, que se basa en métodos de remuestreo y se vincula a bootstrap como sustituto de los intervalos de confianza y a las pruebas de aleatorización como alternativa a los contrastes estadísticos. La inferencia informal, por otro lado, es una conceptualización de la inferencia estadística mediante la simplificación de su complejidad mediante contextos que hacen que la interpretación de los conceptos desarrollados sea significativa, o mediante el establecimiento de analogías y metaconocimientos que proporcionen comprensión. Primero, ilustramos la prueba de significación mediante una prueba de rangos elemental. Segundo, construimos ideas informales sobre la inferencia, por analogía con la situación médica. En tercer lugar, destacamos el potencial de las aproximaciones informales a la inferencia estadística mediante ejemplos. En cuarto lugar, describimos el enfoque de "Inferencia Informal". Finalmente, sacamos algunas conclusiones sobre el potencial didáctico y los inconvenientes de la "Inferencia Informal". Nuestras consideraciones están significadas por el objetivo de facilitar la comprensión conceptual.

Palabras clave: *Inferencia estadística, simulación y remuestreo, comprensión conceptual, pensamiento estadístico, elementarización.*

Resumo

"A inferência informal" é uma aproximação da inferência estatística baseada na reamostragem de métodos e alça de conexões como substituição de intervalos de confiança e testes de re-randomisation como alternativa para testes estatísticos. A inferência informal, de outro lado, é uma conceptualização da inferência estatística por elementarising a complexidade cheia pelo contexto que faz a interpretação dos conceitos desenvolvidos significativa, ou estabelecendo analogias e conhecimento da Meta que fornecem o discernimento. Primeiramente, ilustramos o teste de significação por um teste de fila elementar. Em segundo lugar, construímos ideias informais sobre a inferência por uma analogia com a situação médica. Em terceiro lugar, destacamos o potencial de caminhos informais à inferência estatística por exemplos. Em quarto lugar, descrevemos a "Inferência Informal" aproximação. Finalmente, tiramos algumas conclusões sobre o potencial didático e os descontos "da Inferência Informal". As nossas considerações significam-se pela meta de facilitar a compreensão conceptual.

Palavras-chave: *Inferência estatística, simulação e reamostragem, compreensão conceptual, pensamento estatístico, elementarisation.*

[1] University of Klagenfurt, Austria - – e-mail: manfred.borovcnik@aau.at

Abstract

"Informal Inference" is an approach to statistical inference based on resampling methods and links bootstrap as replacement for confidence intervals and re-randomisation tests as alternative to statistical tests. Informal inference, on the other hand, is a conceptualisation of statistical inference by elementarising the full complexity by context that makes the interpretation of the developed concepts meaningful, or by establishing analogies and meta-knowledge that provide insight. Firstly, we illustrate the significance test by an elementary rank test. Secondly, we build informal ideas about inference by an analogy to the medical situation. Thirdly, we highlight the potential of informal ways to statistical inference by examples. Fourthly, we describe the "Informal Inference" approach. Finally, we draw some conclusions about the didactical potential and the drawbacks of "Informal Inference". Our considerations are signified by the goal of facilitating conceptual understanding.

***Keywords**: Statistical inference, simulation and resampling, conceptual understanding, statistical thinking, elementarisation.*

1. Introducción

Este artículo tiene dos objetivos principales: ilustrar las formas de elementarizar la compleja estructura de la inferencia estadística; comparar dos enfoques diferentes de la elementarización. Las dificultades en los conceptos y la adquisición de conceptos individuales en la estocástica en general y en la inferencia estadística son bien conocidas. Esto ha llevado a la búsqueda de nuevas formas de aprendizaje; a la elementalización, la idea de visualización y las Nuevas Tecnologías se han integrado en la enseñanza desde muy temprano. Los métodos estadísticos intensivos en computadoras también han servido como incentivo para las innovaciones didácticas.

La inferencia informal puede ser utilizada como una etiqueta para los esfuerzos por simplificar, visualizar o simular el modelo *hipotético* detrás de la inferencia estadística. Esto significa que el modelo estadístico en segundo plano sigue siendo el objetivo de la enseñanza y constituye el trasfondo. Esto implica que el carácter teórico de tales modelos se visualiza por medios más sencillos. La elementarización se considera una etapa transitoria para la inferencia estadística.

La "Inferencia Informal", que se remonta a los métodos intensivos en computación en estadísticas como bootstrap y re-aleatorización, es un enfoque educativo que *reduce* completamente la *inferencia* estadística *a los datos observados*, desarrollando los métodos basados únicamente en los escenarios de remuestrear de estos datos. Sólo hay hipótesis nulas naturales de ningún efecto que pueden ser probadas para determinar su significado, o los intervalos se calculan a partir de datos artificialmente simulados que imitan los intervalos de confianza.

Ilustramos ambos enfoques y presentamos una discusión detallada sobre los méritos relativos y mostramos cómo construir la comprensión conceptual mediante el metaconocimiento basado en la simplificación de la inferencia estadística.

2. Un enfoque elemental de la prueba de significación

Ilustramos la forma de pensar en una prueba de significación en una situación muy simple, que también ha sido utilizada por R. A. Fisher en su temprana justificación del método. La *tarea* es: La eficacia de un fármaco antihipertensivo debe ser corroborada por un ensayo clínico a doble ciego, aleatorio y controlado contra placebo. La *variable objetivo* es: La diferencia intra-individual de presión arterial = presión arterial sistólica al inicio menos el valor después de 4 semanas de medicación medido en mm Hg. Las *hipótesis* en la prueba: La hipótesis nula (H_0) establece que Verum (el medicamento) es igualmente efectivo que Placebo (un medicamento falso que no puede ser reconocido como tal por los pacientes ni por el médico). La hipótesis alternativa (H_1) es que Verum es mejor que Placebo.

2.1. Re-atribución y rangos

Las ideas básicas se ilustran con el test de Mann-Whitney para muestras independientes. Se utilizan rangos en lugar de las mediciones de los pacientes y un argumento de realeatorización para leer un valor de *P* para H_0 de los datos. Después de ordenar y clasificar los datos (ver Figura 1), encontramos – sorprendentemente – todos los datos del grupo Placebo en los rangos más bajos con una suma de rango de 10, mientras que el grupo Verum alcanza la suma máxima de rango de 26.

La hipótesis nula establece que no hay diferencia en el efecto de Verum o Placebo, por lo que se nos debe permitir percibir a 4 de las 8 personas como el grupo de control (Placebo) y a las otras como el grupo Verum. La ventaja de la forma actual de abordar el problema es la siguiente: La hipótesis nula tiene la obvia implicación de que *cualquiera de estas formas de reclutar un hipotético grupo de control tiene la misma justificación y, por lo tanto, la misma probabilidad*. Sólo tenemos que encontrar todas las reatribuciones de 4 personas de las 8 a un grupo de control. Hay

$$C(8,4) = \binom{8}{4} = \frac{8!}{4!\,4!} = \frac{8 \cdot 7 \cdot 6 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1} = 70 \,.$$

En la Figura 2 (izquierda), ordenamos estas posibilidades por la suma del rango (sólo unos pocos para mostrar el principio); en la Figura 2 (derecha), mostramos las posibilidades de la suma del rango mediante un gráfico de barras.

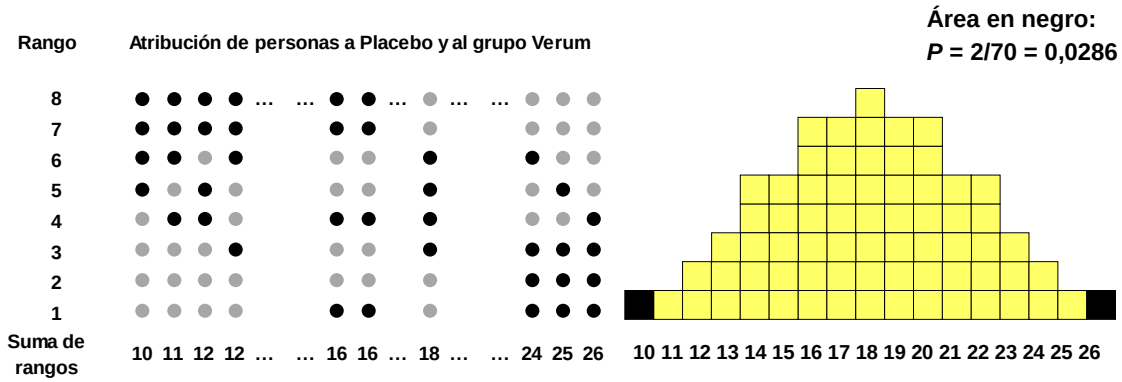
Bajo la hipótesis nula, la distribución en la Figura 2 (derecha) representa la distribución de probabilidad de la suma del rango. Como la probabilidad de obtener la suma del rango extremo de 10 para el grupo Placebo es sólo 1/70, obtenemos un valor de *P* para H_0 de 2/70 = 0,0286 < 0,05 (si la prueba se aplica por ambos lados, es decir, si una diferencia entre los dos grupos podría ser de cualquier manera). En la percepción habitual de las pruebas de significación, podemos rechazar H_0 en el nivel del 5%.

Figura 1. Datos y datos ordenados del experimento Verum contra Placebo

	Datos originales	Ordenados	Rangos	Suma de rangos
Placebo	2,5	0,9	1	
	0,9	1,8	2	Σ = 10
	1,8	2,5	3	
	3,6	3,6	4	
Verum	3,7	3,7	5	
	5,2	4,8	6	Σ = 26
	4,8	5,2	7	
	6,1	6,1	8	

Fuente. Los autores.

Figura 2. Izquierda: Todas clasificaciones ordenadas por la suma del rango –
Derecha: Posibilidades de cada una de las sumas de rango como distribución de probabilidad bajo H_0



Fuente 2. Los autores.

2.2. El valor de P: Algunas preocupaciones iniciales

Hemos calculado la probabilidad de un resultado observado si se aplica la hipótesis nula H_0 y utilizamos este valor de *P* para juzgar la credibilidad de H_0. Si *P* es menor del 5%, H_0 es rechazado; *P* es la probabilidad de una declaración de falso positivo, es decir, la prueba produce un resultado significativo si el medicamento no es efectivo:

P = *P*(la prueba estadística es significativa | la droga médica no es efectiva).

Hemos observado algo que tiene menos de 5% de probabilidad si se aplica H_0 (medicamento no efectivo). Sin embargo, sólo nos interesa la siguiente probabilidad:

P(la droga médica es efectiva | la prueba estadística es significativa).

¡Pero este número no se puede calcular a partir de los hechos! La conclusión sobre un estudio clínico se basa en métodos estadísticos. Los médicos no son expertos en estadística y no necesitan serlo. Sin embargo, deben conocer los principios de los métodos científicos. Neyman y Pearson (1933) limitan claramente el alcance de las pruebas estadísticas. Afirman que "ninguna prueba basada en una teoría de la probabilidad puede por sí sola proporcionar una evidencia valiosa de la verdad o falsedad de una hipótesis". Un diálogo entre un facultativo y un estadístico ilustra el conflicto:

> Facultativo: Intentaste explicarme la prueba estadística, pero ¿qué significa si mi prueba da un resultado significativo? ¿Puedo alegar que el medicamento es efectivo?
>
> Estadístico: No – Usted sólo puede calcular qué tan probable es el resultado de la prueba si el medicamento no es efectivo.
>
> Facultativo: La comisión de ética ha aprobado este estudio para investigar la eficacia de este medicamento. Le he preguntado si puede demostrarlo mediante una prueba estadística. Como el resultado ahora es significativo, pensé, que la probabilidad de que mi medicamento sea efectivo es del 95%, porque el valor de P es del 5%.
>
> Estadístico: Me preguntaste algo para lo que el valor de P no tiene respuesta. La probabilidad de error de su estado de cuenta es mayor, pero no puedo calcularlo.
>
> Facultativo: Puede que tengas razón, pero yo lo he hecho como todos, ¿por qué debería estar mal? El resultado de la prueba estadística es significativo y será publicado: El fármaco es eficaz ($P < 0,05$).

3. Inferencia informal – Una analogía con la situación médica

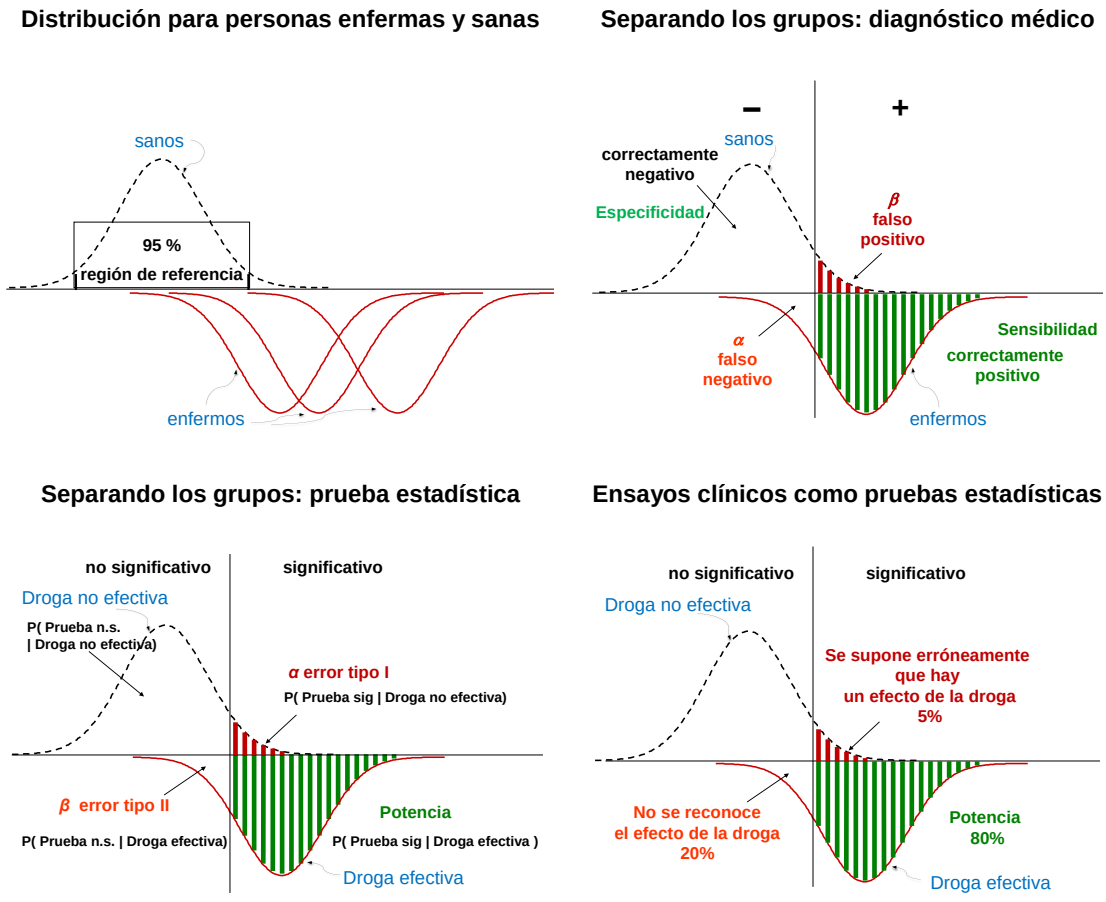
Exploramos la situación en medicina donde siempre hay una decisión que puede llevar a varios errores, sea cual sea la decisión. Una prueba de diagnóstico puede compararse con una prueba estadística. Esto sirve para entender mejor las pruebas estadísticas. También puede servir para entender e investigar mejor la decisión médica.

3.1. Separar las personas sanas y enfermas

La otra tarea en medicina es el diagnóstico de una enfermedad que se basa en los valores de una variable biométrica de un examen médico. Comparamos la tarea de separar dos grupos por los valores de una variable en los diferentes escenarios (Figura 3): la prueba diagnóstica, el ensayo clínico, la prueba estadística.

Para la prueba de drogas, ha surgido una norma con cifras mágicas: la potencia (probabilidad de detectar un efecto de la droga si existe) debe alcanzar el 80%, el error alfa (droga que erróneamente se supone que tiene un efecto) no debe ser superior al 5%.

Figura 3. Separación de dos grupos: Terminología diferente para los mismos conceptos



Fuente. Los autores.

3.2. Prueba médica como situación de decisión

Extendemos la prueba en la Sección 2 para incluir la hipótesis alternativa (cambio en la media de la variable bajo escrutinio) para permitir consideraciones de poder. El ensayo clínico, el problema del diagnóstico o el control de calidad (Sección 4.4), todos llevan la

estructura de una situación de decisión. Se debe tomar una decisión sobre H_0 o H_1 basada en los datos.

Es útil reconocer que la estructura del problema de decisión y los errores potenciales siguen siendo los mismos en los tres contextos. La analogía (Tabla 1) ilustra el significado del mismo concepto en varios contextos:

$P = P$(diagnóstico + | enfermo) Sensibilidad en el contexto del diagnóstico,

$P = P$(prueba significativa | medicamento efectivo) Potencia en pruebas estadísticas.

Falta cualquier información sobre el valor predictivo positivo / negativo (VPP o VPN):

P(enfermo | diagnóstico +) o P(medicamento efectivo | prueba significativa) y

P(sano | diagnóstico –) o P(medicamento no efectivo | prueba no significativa).

Esta probabilidad describe la calidad del procedimiento de decisión. No sólo que no lo sepamos, sino que también depende del predominio de la enfermedad o de la calidad de las hipótesis de investigación (tanto en las pruebas de drogas como en las pruebas estadísticas).

Tabla 1. El ensayo clínico como situación de toma de decisiones con los diferentes errores [1]

			Realidad		
			El medicamento no es efectivo El paciente está sano La calidad del lote es buena		El medicamento es efectivo El paciente está enfermo La calidad del lote es mala
Decisión de la prueba			**Hipótesis nula H_0**		**Hipótesis alternativa H_1**
“Medicamento no es efectivo” **–** “ El paciente está sano” Lote es aceptado	**No rechazar H_0**	Correcto $1-\alpha$	Ok Especificidad Ok	Error β	La falta de ese medicamento es efectivo Diagnóstico falso negativo Falsa aceptación – riesgo del consumidor
“Medicamento es efectivo” **+** “El paciente está enfermo” Lote es rechazado	**Rechazar H_0**	Error α	Falsa decisión sobre el medicamento Diagnóstico falso positivo Falso rechazo – riesgo del productor	Correcto $1-\beta$ potencia	Potencia Sensibilidad Ok

[1] ¡El orden de H_0 y H_1 así como las decisiones es diferente al anterior!

Utilizamos los datos de la mamografía en la clínica radiológica y en un programa de detección en la Tabla 2, que muestra los números absolutos de las diversas combinaciones de enfermedad y diagnóstico. Si leemos las columnas, obtenemos sensibilidad y

especificidad. La tabla permite también calcular las proporciones en filas, que son las cifras más interesantes de arriba, es decir, el VPP y el VPN. Sorprendentemente, el VPP – la probabilidad de que una persona tenga un carcinoma después de un diagnóstico positivo – depende del contexto y del predominio de la enfermedad bajo escrutinio (lo mismo ocurre con el VPN).

Gigerenzer (2002) ha abogado por reformular las probabilidades (la sensibilidad y la especificidad son usualmente conocidas) por valores esperados, que él llama frecuencias naturales. Permiten una orientación rápida sobre las probabilidades relevantes (ver Batanero & Borovcnik, 2016).

Tabla 2. Valores esperados de estado (carcinoma o no carcinoma) y diagnóstico (positivo o negativo) en clínica radiológica y en un programa de detección

	Ca	No Ca	Todos
+	80 Sensibilidad ↑	4 Falso pos. ↑	84
–	20 Falso neg. ↑	96 Especificidad ↑	116
Todos	100	100	200

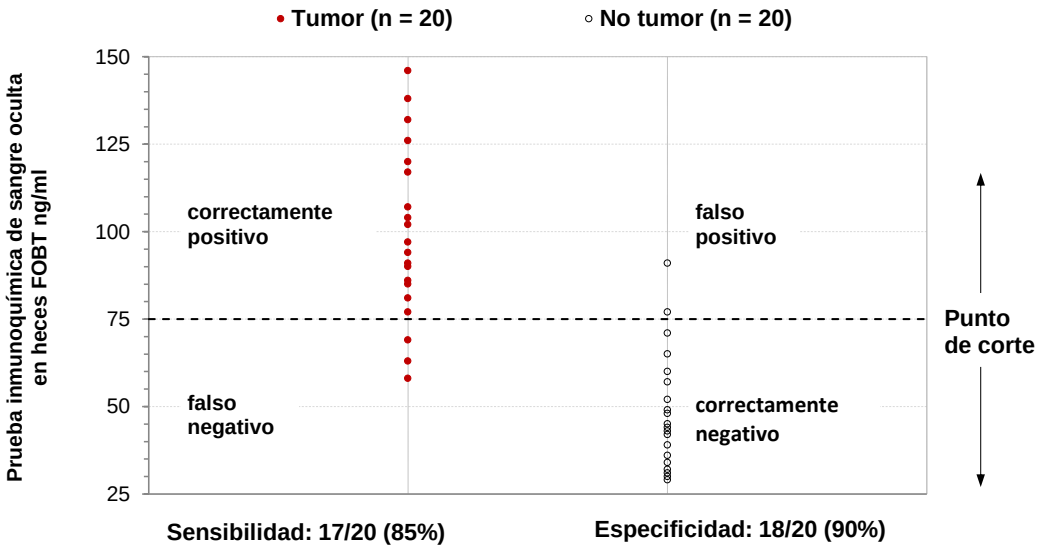
	Ca	No Ca	Todos
+	640 PPV →	3.968	4.608
–	160	95.232 NPV →	95.392
Todos	800	99.200	100.000

Predominio	**Clínico**	**50%**	**Programa de detección**	**0,8%**
Sensibilidad ↑	80,0%	= 80/100	80,0%	P(+|Ca)
Especificidad ↑	96,0%	= 96/100	96,0%	P(–|No Ca)
VPP →	95,2%	= 80/84	13,9%	P(Ca|+)
VPN →	82,8%	= 96/116	99,8%	P(No Ca|–)

3.3. Puntos de corte para separar los grupos de sanos y enfermos

Demostramos la dificultad de separar entre los grupos de pacientes tumorales y los pacientes libres de tumor mediante la introducción de un punto de corte. El examen de sangre oculta en heces (FOBT) se utiliza para detectar el cáncer de colon. Se utilizan datos de 20 pacientes en cada grupo (Figura 4).

Figura 4. La elección del punto de corte conduce a una separación de los dos grupos con una calidad distinta, medida por los conceptos de sensibilidad y especificidad

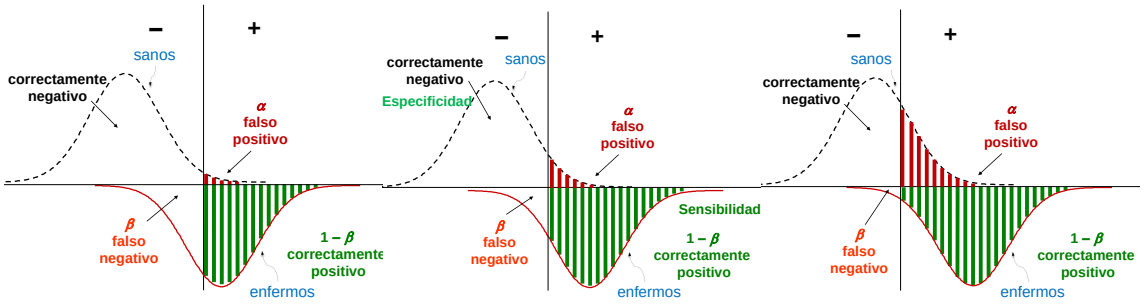


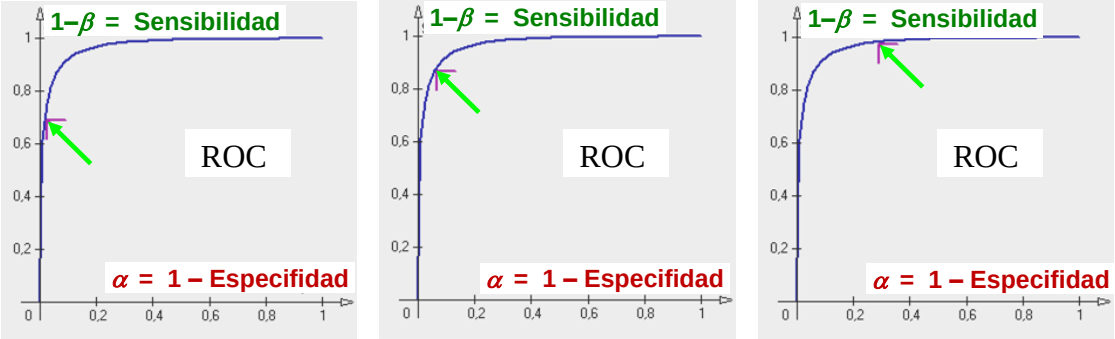
Fuente. Los autores.

Si diagnosticamos a un paciente como positivo si la FOBT supera los 75 y negativo en los demás casos, vemos que en el grupo tumoral tres personas están falsamente mal clasificadas como negativas, lo que equivale a una sensibilidad de 17/20 = 85%. Por otro lado, este punto de corte conduce a dos casos de diagnóstico falso positivo en el grupo libre de tumor, lo que corresponde a una especificidad de 18/20 = 90%.

¿Qué punto de corte debe utilizarse para el diagnóstico? Si variamos el punto de corte, generamos varios procedimientos para el diagnóstico, todos con diferentes propiedades (Figura 5).

Figura 5. Todos los puntos de corte conducen a métodos de diagnóstico con un punto en la curva ROC – puntos en la esquina superior izquierda reflejan buenos diagnósticos





1 – β = Sensibilidad:	**0,699**	**0,876**	**0,985**
1 – α = Especificidad:	**0,985**	**0,943**	**0,718**

Fuente. Los autores.

Es habitual ilustrar la calidad del diagnóstico mediante la llamada curva ROC, que muestra a cada punto de corte el punto correspondiente (α, 1 – β), es decir, el error de tipo I en la primera coordenada y la sensibilidad en la segunda. Esto significa que un punto a la izquierda y arriba está ligado a un diagnóstico con buenas propiedades para separar los dos grupos.

Reconocemos que cuanto más alto sea el punto de corte elegido, más a la izquierda (bueno para el diagnóstico) y más abajo (malo para el diagnóstico) estará el punto correspondiente en la curva ROC. Nos enfrentamos a consecuencias antagónicas al desplazar el punto de corte. Tenemos que encontrar un compromiso entre los dos objetivos para conseguir un pequeño error de tipo I y una gran potencia. Varias enfermedades tienen diferentes distribuciones para la variable objetivo y corresponden a diferentes curvas ROC. La enfermedad con la distribución punteada (azul), que es la misma que la distribución entre las personas sanas, convierte el diagnóstico en un experimento de lanzamiento de monedas, es decir, una simple conjetura. El ROC correspondiente es la diagonal, todos los puntos están lejos de la esquina superior izquierda que alberga los puntos para el diagnóstico de procedimientos con buenas propiedades.

Figura 6. Cuanto más cercanas sean las distribuciones de enfermos y sanos, más difícil será el diagnóstico

Fuente. Los autores.

3.4. Algunas conclusiones de la analogía con la medicina

De la analogía con la medicina, aprendemos que normalmente nos enfrentamos a un problema de decisión, que puede ser descrito – para simplificar – por dos escenarios. El diagnóstico de enfermedades es un problema de decisión, que compara las distribuciones bajo el escenario de personas sanas y enfermas. Decidamos lo que decidamos, somos propensos a cometer un error. Siempre hay – al menos – dos errores divergentes en el juego:

- Diagnosticar la enfermedad cuando la persona está sana.
- No reconocer la enfermedad a pesar de que la persona la tiene.

Dondequiera que introduzcamos el punto de corte entre los dos grupos (escenarios), los errores son influenciados por esa elección y debemos orientarnos sobre su magnitud. Varios puntos de corte para separar lo sano de lo enfermo implican diferentes tamaños de estos errores. Hay enfermedades que son fáciles de diagnosticar. Reduciendo la complejidad de la situación de decisión, se utiliza el valor de *P*, pero no es fácil interpretar este número de una manera práctica y significativa. Además, hay un tercer error: Si la decisión es buena o no, no sólo depende de los puntos de corte, sino también del predominio de la enfermedad. En resumen, en muchos casos no se obtienen coeficientes de calidad de las decisiones que puedan interpretarse correctamente.

4. Formas informales de inferencia estadística

Ilustramos varias maneras informales de explorar conceptos estadísticos clave. Uno de los principales problemas es resaltar la relevancia y el significado de la distribución

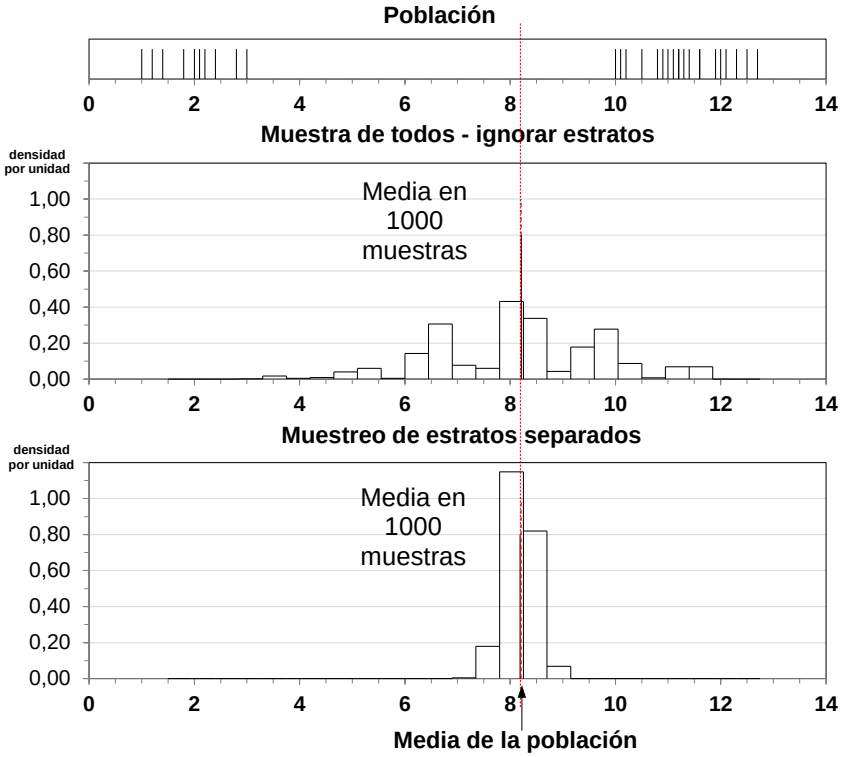
muestral de las estadísticas que estiman un parámetro de la población. Otra idea es reducir la complejidad de las pruebas estadísticas a una comparación de dos distribuciones que tenga sentido en el contexto, de modo que las decisiones y sus implicaciones se conviertan en un tema natural a discutir, al igual que en la analogía con la situación médica (de diagnosticar o probar drogas). Las exploraciones sirven para aprender sobre las características clave, también mediante el establecimiento de metaconocimientos sobre el método más allá de las matemáticas. El objetivo es reducir la complejidad de la situación, pero mantener abierto el camino hacia la situación general.

4.1. Dos métodos diferentes para estimar la media

Los valores de la población están marcados por una barra (ver Figura 7). Dos estratos homogéneos son visibles. Si se conoce tal caso de estratos, es aconsejable tenerlo en cuenta en el muestreo. Se comparan dos métodos: Método 1: Muestra aleatoria de 6 elementos de toda la población ignorando los estratos; Método 2: Muestra aleatoria de 2 del estrato 1 y 4 del estrato 2. Para ambos métodos, podemos ver en el escenario de simulación (Figura 7) que la media de los datos simulados es aproximadamente igual a la media de la población (estimador insesgado). Vemos también que el muestreo de estratos (Método 2) proporciona resultados mucho más precisos. La mejora de la estimación mediante el muestreo de los estratos en comparación con el muestreo sin tener en cuenta los estratos es estable en la repetición del escenario.

Se puede visualizar a las personas que son muestreadas y sus datos y mostrar – como en un video – cómo éstos cambian al renovar la muestra para obtener una impresión sobre la variabilidad del muestreo y el error cambiante en la estimación de la media de la población por la media de la muestra. Se aclarará inmediatamente que el Método 2 (muestreo estratificado) conduce a errores menores en general. Es importante ver cómo se comportan las muestras individuales antes de resumir el resultado de muchas muestras mediante la distribución de la media. Esto es importante ya que en la práctica sólo tenemos una (!) muestra. El resultado de este proceso (con 1000 muestras y su media cada una) se muestra en la Figura 7: Lo que ha sido una impresión se corrobora ahora en el escenario de la simulación. Por el Método 1 (muestreo sin restricciones), el error es en general muy grande con medias entre 2 y 12, mientras que para el Método 2 (muestreo de estratos) el error tiende a ser pequeño con medias entre 7,5 y 9,5.

Figura 7. Muestreo sin restricciones y de estratos – distribución del muestreo de la media



Fuente. Los autores.

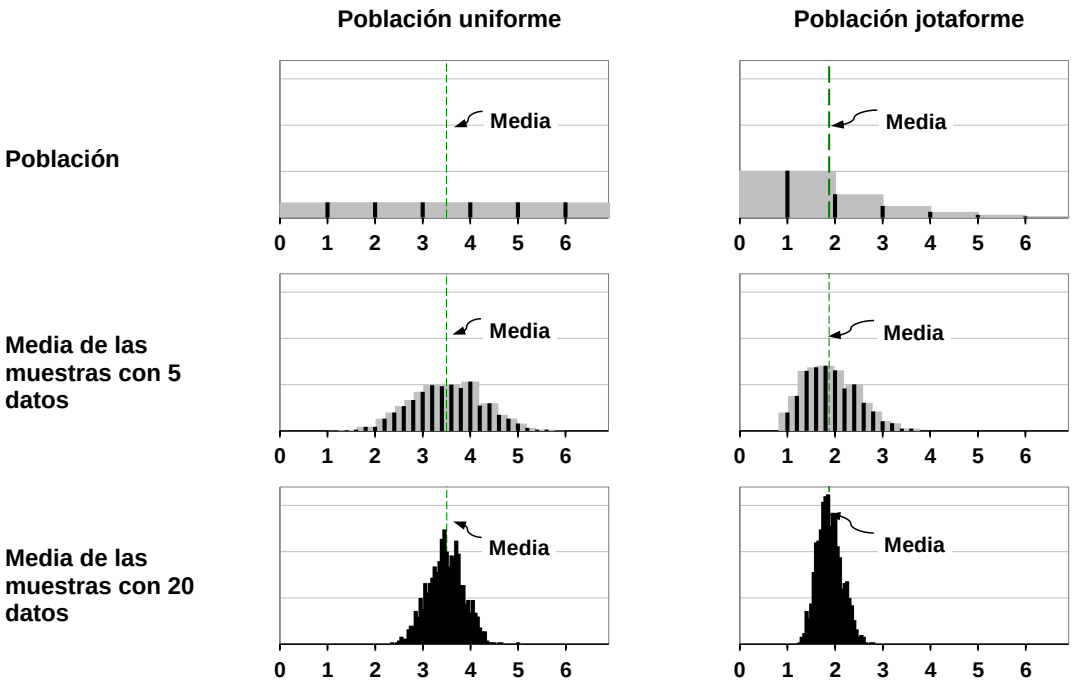
4.2. La distribución del muestreo de la media es artificial

En el laboratorio estadístico, podemos simular muestras de cualquier población. La distribución del muestreo de la estimación de cualquier parámetro indica cómo varía la estimación de una muestra a otra. Por lo general, sólo tenemos una muestra y por lo tanto parece contra-intuitivo hablar de la variación de la estimación. Sin embargo, en un experimento de pensamiento, podemos repetir la muestra muy a menudo para ilustrar las propiedades de la estimación. ¿Tenemos suerte en una muestra de tener una estimación cercana al parámetro correspondiente de la población, o podemos confiar en el hecho de que el riesgo general de obtener grandes desviaciones del parámetro de la población es pequeño?

Podemos simular dos poblaciones completamente diferentes – una con una distribución uniforme y otra con una distribución en forma de J sobre la población – para resaltar el concepto de distribución del muestreo e ilustrar sus características clave (ver Figura 8). Independientemente de la población parental, la distribución del muestreo de la media (como de muchos otros parámetros) se restringe al punto de la media de la población (el parámetro de interés) y se asemeja cada vez más a una distribución normal con un tamaño

de muestra creciente. Véase también Batanero y Borovcnik (2016) para un escenario que muestre estas propiedades de la distribución del muestreo de la media; la Figura 8 ilustra la evolución de 5 a 20 muestras. Nótese que para la media, el ancho de la distribución (medido por el error estándar) se reduce a la mitad si tomamos una muestra 4 veces más grande que antes. Es instructivo ver que la forma de la distribución del muestreo es casi invariable con la repetición.

Figura 8. Distribución de la muestra de la media de una muestra de una población uniforme (izquierda) y una población en forma de J (derecha)



Fuente. Los autores.

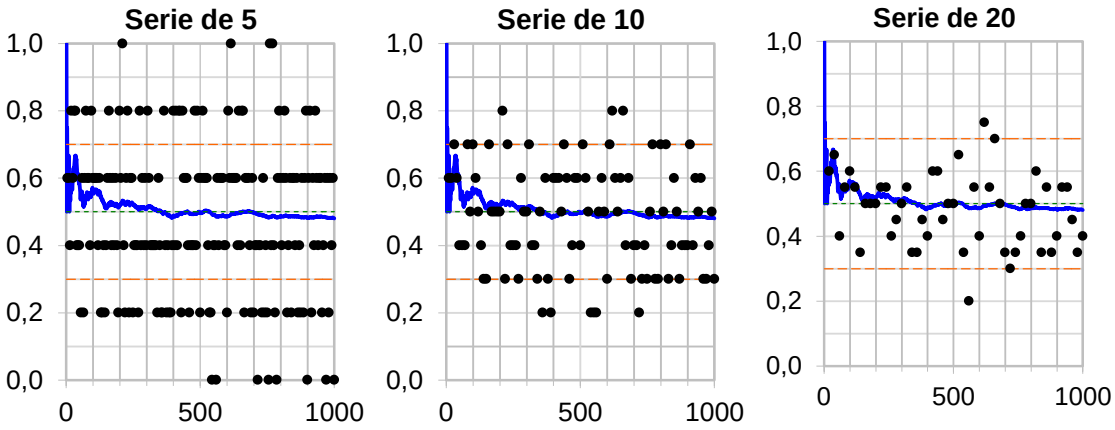
4.3. Medición de una probabilidad desconocida

En el siguiente experimento (lanzamiento de una moneda), se investigan las frecuencias relativas según una idea expresada en Batanero y Borovcnik (2016). En lugar de mostrar cómo convergen las frecuencias relativas (¿qué debería significar y hacia dónde deberían converger?), se plantea la tarea de estimar la probabilidad desconocida. La estimación puede basarse en muestras (bloques) de 5 ensayos (5 lanzamientos de la moneda) o 10 ó 20. En la Figura 9 se muestra cómo convergen las frecuencias relativas con el número de ensayos y se observa el desarrollo hasta que se realizan 1000 ensayos. La serie actual (en la Figura 9) no puede fluctuar mucho debido a los últimos 1000 valores. La curva sugiere una gran precisión de menos de 0,5 puntos porcentuales de fluctuación. Sin embargo, un nuevo experimento muestra – como en un video – otra curva con otro “punto límite”;

dentro de +/– 3 % de puntos; una repetición de la serie de 1000 pruebas también "convergerá"; aún en un punto diferente.

La ley de los grandes números establece que las frecuencias relativas teóricas (no empíricas) "convergen" hacia la probabilidad desconocida. Esta "convergencia" en un experimento real oculta que los resultados actuales siguen siendo propensos al azar. ¿Qué tal si cambiamos la tarea y medimos la probabilidad desconocida por series cortas e investigamos la precisión de tal medición? Después de cada bloque de 5 (10 ó 20), la muestra se resume y se utiliza para estimar la probabilidad desconocida. Las estimaciones pueden ser 0,0, 0,2, ..., 0,8 y 1,0 (según 0, 1, ..., 5 cabezas). En la Figura 9 (izquierda), vemos cómo fluctúan estas estimaciones; muchas están más allá de las líneas punteadas (rojas) con un error de estimación mayor que 0,2. Por supuesto, la muestra es muy pequeña, el error debe ser grande.

Figura 9. Medición de una probabilidad desconocida – Estudio de la precisión



Fuente. Los autores

Figura 10. Distribución de las mediciones repetidas

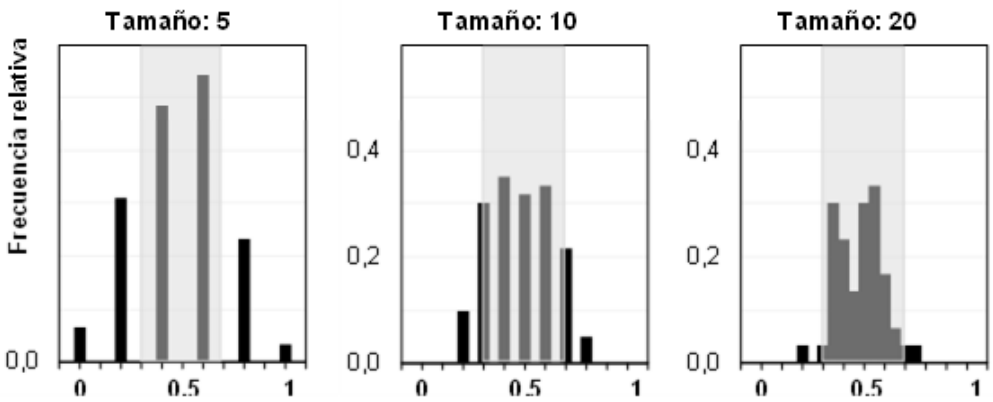


Figura 10. Los autores.

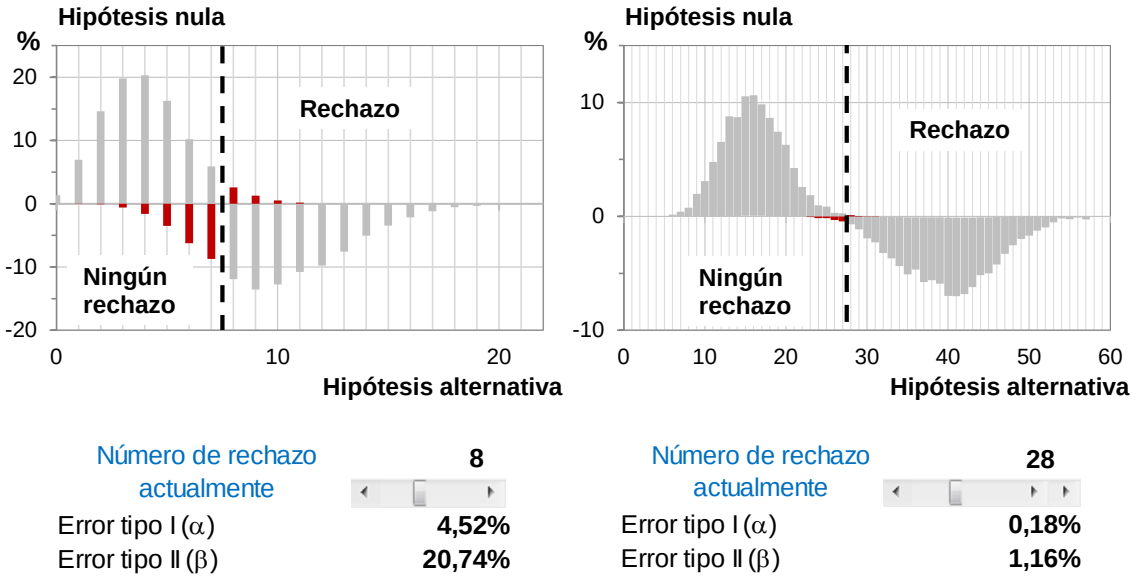
En el diagrama central de la Figura 9, se muestran los resultados de la medición de la probabilidad por muestras de 10 y la estimación fluctúa mucho menos; aún menor es la variación de la estimación (y los errores son menores) en el diagrama derecho de la Figura 9, que muestra las estimaciones de la probabilidad basadas en muestras de 20. Vemos sólo dos estimaciones más allá de las líneas discontinuas (rojas). Mediante un experimento de pensamiento, se puede concluir que la precisión de la estimación mejora cuanto mayor sea la muestra y su distribución se restringirá a la probabilidad (desconocida). Esta constricción de la distribución de las estimaciones, que se representa en la Figura 10, también puede observarse en el estudio de la distribución del muestreo de la media (en la Figura 8). El riesgo de cometer un error superior a 0,2 disminuye con el tamaño de la serie (muestra) en la que se basa la medición.

4.4. Separar buena y mala calidad – Ejercicios informales en pruebas

El siguiente ejemplo se remonta a Batanero y Borovcnik (2016). La tarea es juzgar si la producción actual (o un lote que ha llegado) tiene una buena o mala calidad. Los elementos individuales pueden mostrar sólo esta propiedad, que está codificada por 0 (bueno) y 1 (defectuoso). Al inspeccionar una muestra de n ítems, se debe tomar una decisión sobre la calidad. El número de defectos en la muestra está distribuido hipergeométricamente; descuidando que el muestreo es sin reemplazo, podemos usar la distribución binomial en su lugar. Se comparan dos escenarios, que representan diferentes partes interesadas: un lote ha entrado en la fábrica del consumidor, enviado por el productor; la buena calidad está representada por $p = 0{,}04$, la mala calidad por $p = 0{,}10$ (p representa la proporción de defectos).

En lugar de utilizar la distribución binomial, hemos simulado 5000 muestras de tamaño $n = 100$ y determinado la frecuencia relativa a partir del escenario de simulación para estimar las probabilidades. En la Figura 11, mostramos la implicación de un número de rechazo, digamos, rechazar el lote como malo (rechazar la hipótesis nula de buena calidad) a favor de la hipótesis alternativa. Se puede cambiar el número de rechazo (desplazar la barra de puntos en el diagrama) y se hace visible cómo cambian los dos tipos de error y reconocer que son antagónicos, es decir, mientras uno se hace más pequeño, el otro se hace más grande.

Figura 11. Un número de rechazo (línea de puntos vertical) se asocia con dos tipos de errores: la decisión se basa en una muestra con n = 100 (izquierda) y n = 400 (derecha)



Fuente. Los autores

En el lado derecho de la Figura 11, vemos las consecuencias de una muestra más grande para la decisión. Con 400 datos, ambos errores se vuelven pequeños. La elección del número de rechazo equilibra los intereses divergentes del proveedor y del comprador. El diagrama (Figura 11) con el umbral por encima del cual se rechaza la hipótesis nula a favor de la alternativa nos recuerda a los diagramas de la Sección 3 del diagnóstico médico. Básicamente, es el mismo tipo de situación de decisión. Estar por encima del número de rechazos equivale a estar por encima del punto de corte; la decisión consiguiente es que el lote tiene una mala calidad, que corresponde al diagnóstico positivo (el paciente está clasificado para tener la enfermedad bajo escrutinio).

Tenemos muchas otras tareas, en las que no tenemos que separar dos o más grupos, sino estudiar las consecuencias de los escenarios para los diferentes grupos.

- Por ejemplo, el caso de los exámenes de opción única. Mientras que las suposiciones para una distribución binomial para el número de ítems resueltos correctamente con sólo adivinar son indiscutibles, las suposiciones no son realmente apropiadas para describir a alguien que ha aprendido. Sin embargo, los escenarios pueden ser jugados con ideas valiosas sobre la forma en que tales exámenes tienen que ser diseñados para mantener el riesgo pequeño de que alguien pase el examen que está simplemente adivinando y – al mismo tiempo – para asegurar que el riesgo se vuelve pequeño para

fallar en el examen para alguien que ha aprendido y tiene una capacidad de resolución de 0,55 o incluso 0,80.

- Otro prototipo de tareas se origina en el control estadístico del proceso, donde la inspección horaria de la calidad de las mediciones debe orientarse hacia la calibración actual de la máquina de producción con una alarma incorporada en caso de que la calibración se haya desplazado. Para más detalles, véase Borovcnik (s.f.).
- También se pueden investigar escenarios de muestreo repetido con respecto a los intervalos de confianza para mostrar que la interpretación usual puede ser violada severamente. Si determinamos los intervalos de confianza para la media en caso de varianza desconocida, entonces no sabemos si la varianza en la muestra está por debajo o por encima de la varianza real de la población. Por lo tanto, no sabemos si los intervalos son más cortos que la media o más largos. Es fácil ver en un escenario de simulación que los intervalos de confianza cortos tienen una tasa de cobertura mucho menor que los intervalos largos. Esto significa que definitivamente no podemos transferir la cobertura a largo plazo del 95% para todos los intervalos de confianza repetidos al intervalo actual que acabamos de calcular. Para más detalles, véase Borovcnik (s.f.).
- Además, es necesario calibrar las expectativas conceptuales sobre la precisión de la simulación y el remuestreo. Un escenario de simulación puede aclarar rápidamente las expectativas engañosas y revelar la variabilidad de los resultados de la simulación. Si simulamos una distribución normal, por ejemplo, podríamos esperar un histograma en forma de campana. Con 50 datos simulados, la forma del histograma parece extremadamente errática; con 1000 datos simulados, el histograma comienza a aparecer como una distribución normal. Sin embargo, todavía hay desviaciones imprevistas y muy cambiantes de esa forma. Repetir la simulación de los datos y mostrar el impacto en los histogramas relacionados, como en un vídeo, podría convencer de que la simulación sólo tiene valor si se generan muchos datos; 1000 no son suficientes para lograr resultados estables. Eso refleja el viejo proverbio “en la lotería – todo es posible”. Una forma válida y generalizable sólo aparece en muestras simuladas muy grandes. Para muestras más pequeñas en las que todavía hacemos estadísticas, tenemos que medir la variabilidad del azar (el error de azar) por la amplitud de los intervalos de confianza.

5. “Inferencia informal”

La “inferencia informal” centra todas las consideraciones sobre la generalización de la información contenida en un conjunto de datos determinado únicamente en estos datos. Los primeros pasos del desarrollo del enfoque informal son los siguientes: El remuestreo como técnica didáctica (Borovcnik, 1996), como etapa transitoria de la inferencia estadística (Borovcnik, 2006a, b); como método para reemplazar la inferencia estadística (Cobb, 2007); pruebas de realeatorización como reemplazo de la prueba de significación (Rossman, 2008); intervalos de bootstrap para reemplazar los intervalos de confianza (Engel, 2010). Stohl Lee, Angotti y Tarr (2010) presentan un enfoque panorámico con ejemplos. Los métodos se explican a continuación; se puede consultar más sobre la metodología en Lunneborg (2000); para una amplia crítica del enfoque, véase Howell (s.f.).

5.1. Introducción al enfoque de “inferencia informal”

La estimación Bootstrap se utiliza para estimar el error estándar. En lugar de tomar muestras de la función F de distribución acumulativa (que describe la población), se toma una muestra (con reemplazo) de la estimación de F basado en la muestra inicial. El bootstrap produce intervalos aproximados para el parámetro desconocido.

Prueba de hipótesis Esto se reduce a pruebas de aleatorización. La reatribución aleatoria de la asignación a grupos para ser comparados proporciona datos artificiales que se utilizan para la prueba. Se investigan todas las permutaciones de los datos, o el muestreo se realiza a partir de los datos sin reemplazo, lo que equivale al muestreo de todas las permutaciones. Este enfoque proporciona pruebas exactas no paramétricas en casos específicos.

El caso de la hipótesis nula natural La intención de la “inferencia informal” es incrustar la situación compleja de la inferencia estadística en un entorno natural y material (es decir, los datos), dejando fuera cualquier consideración sobre las hipótesis, excepto la hipótesis nula natural (véase la Sección 5.3) de los efectos aleatorios puros sobre las unidades estadísticas.

Inferencia sobre un “grupo” Si se ha de juzgar un conjunto de datos, por ejemplo, para un parámetro de localización, se obtiene un intervalo de bootstrap muestreando repetidamente a partir de los datos dados (siempre calculando esta medida estadística). Este escenario de remuestreo proporciona una base empírica (una distribución empírica),

que está abierta para un análisis estadístico. Si un valor de parámetro (hipotético) cae fuera del intervalo de bootstrap, entonces es "rechazado".

Inferencia sobre dos grupos Si se van a comparar dos conjuntos de datos dados para un parámetro de localización (o cualquier otro parámetro), entonces hay dos opciones: Primero, podemos remuestrear a partir de los datos dados en cada grupo para derivar el intervalo de bootstrap para este parámetro. En segundo lugar (y mucho más intuitivo), podemos volver a aleatorizar la atribución de datos individuales a uno de los grupos mediante una decisión aleatoria. Si se aplica la hipótesis nula de que no hay diferencia entre los dos grupos, entonces los datos pueden agruparse y, a partir de este grupo, los datos del grupo 1 (y 2) pueden seleccionarse aleatoriamente, de modo que una vez más, una base empírica de la estadística de interés puede ser generada únicamente por los datos dados. La atribución aleatoria inicial se rehace aleatoriamente sobre los datos existentes, lo que refleja la hipótesis del efecto nulo natural.

5.2. Intervalo de bootstrap e intervalo de confianza para la media

Dada: Una muestra del tamaño *n* con la media y la deviación típica de una variable específica (los datos se muestran en la Figura 12). ¿Cuán precisa es la media de la muestra como medida para la población? La variable Tiempo se refiere al tiempo trabajado en un seminario. En lugar de volver a tomar muestras de la población, lo cual es imposible, tomamos muestras de la primera muestra (con reemplazo). El primer bootstrap produce una nueva medición de la media de la población, que no difiere demasiado de la media de la muestra original. Repetimos el bootstrap y obtenemos 1000 (o más) mediciones artificiales.

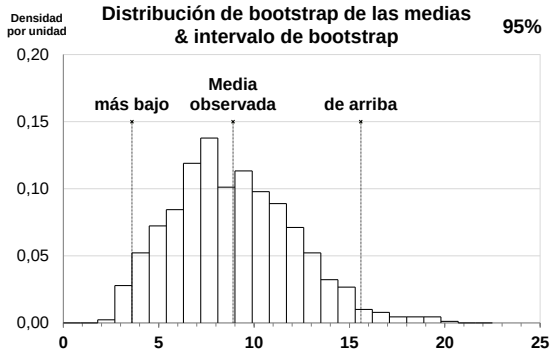
Figura 12: Muestra del tamaño *n* con la media y la deviación típica de una variable específica

Datos primarios	
Nr	Tiempo
1	12
2	2
3	6
4	2
5	19
6	5
7	34
8	4
9	1
10	4

Primero bootstrap	
Nr	Tiempo
7	34
8	4
4	2
9	1
9	1
10	4
9	1
7	34
8	4
5	19

n	Media
10	8,90

Media
10,40



Fuente. Lo autor.

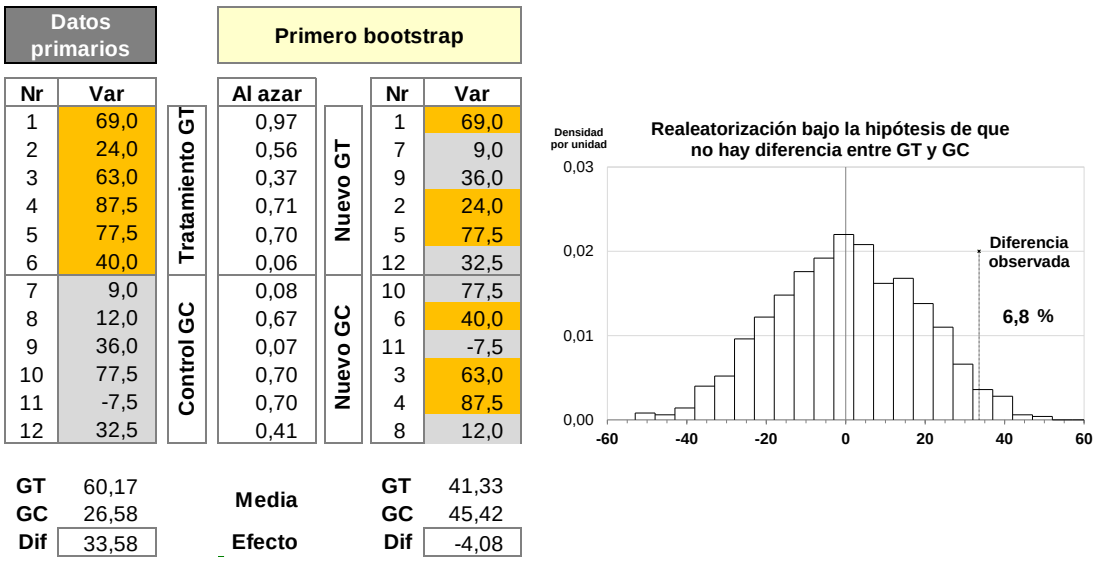
Los datos artificiales obtenidos por este método reflejan la variabilidad de las mediciones repetidas de la media desconocida de la población. A partir de la distribución de bootstrap para la media, podemos cortar fácilmente el 2,5% más bajo y el 2,5% más alto de la media de bootstrap para obtener el intervalo de bootstrap del 95%, que rinde (3,72, 15,39) en nuestro escenario de simulación. Esto puede compararse con el intervalo de confianza clásico de (2,46; 15,34). Vemos un buen acuerdo en los resultados de ambos métodos. Sin embargo, la interpretación difiere. El intervalo de bootstrap refleja la precisión de las mediciones repetidas de la media de la población, mientras que el intervalo de confianza contiene la media de la población en el 95% de las muestras "repetidas".

El bootstrap puede aplicarse para estimar otros parámetros de la distribución. El procedimiento es análogo al de la media. Estimar el parámetro relevante a partir de la muestra original, bootstrap la primera muestra (mediante muestreo con sustitución) y calcular el parámetro de la muestra remuestreada para obtener datos artificiales sobre el parámetro de interés y analizar su distribución. Corte el extremo superior e inferior de esta distribución para obtener un intervalo de bootstrap. También se puede determinar un intervalo de bootstrap para la correlación de la misma manera.

5.3. Prueba de aleatorización para la diferencia de medias

¿Es efectivo el tratamiento (medical) con respecto a una variable objetivo? El grupo de tratamiento (GT) recibe Verum – el grupo de control (GC) recibe Placebo. La realeatorización ofrece una alternativa a la prueba t de dos muestras. El procedimiento es similar al de la prueba de significación de la Sección 2. En lugar de ocuparnos de los rangos de observaciones, analizamos aquí los valores de los datos. En lugar de determinar todas las posibilidades de varias sumas de rango, simulamos a partir de la distribución de todas las posibilidades. Bajo la hipótesis nula de NO DIFERENCIA, es intuitivo que cualquier reatribución de personas a tratamientos no debe tener NINGÚN EFECTO. Por lo tanto, permutamos a las personas y el siguiente grupo de tratamiento consiste en 1, 2, 5, 6, 3 y 4. La primera reatribución produce una nueva medición de la diferencia de medias (como la medición del efecto del tratamiento); la diferencia entre el grupo de tratamiento y el grupo de control en la muestra original es de 33,58, mientras que la primera reatribución produce una diferencia de –4,08 (ver Figura 13).

Figura 13. Izquierda: Muestra original de la variable objetivo y primera reatribución de personas al tratamiento – Derecha: Histograma de 1000 reatribuciones



Fuente. Lo autor

La distribución para la repetida realeatorización se muestra en la Figura 13 (derecha); produce los resultados artificiales basados en la hipótesis de NO DIFERENCIA, es decir, la hipótesis nula. Podemos encajar el resultado de la primera muestra en esta distribución y ver que el valor de *P* de la misma es del 6,8% (bilateral). Todo el escenario de simulación puede repetirse para mostrar que el resultado es estable. Una vez más, podemos comparar este resultado de la reatribución con la prueba t de dos muestras, que arroja 2,16 con un valor de *P* de 5,6% o 2,16 con 5,9% (dependiendo de la suposición adicional de varianzas iguales o desiguales).

Una vez más, la similitud de los resultados clásicos con la prueba de reatribución es sorprendente. El procedimiento puede aplicarse también a otras comparaciones; por ejemplo, la tarea de correlación puede reformularse como prueba de la hipótesis de que la correlación en la población es cero. Ver Borovcnik (s.f.) para más detalles.

5.4. “Inferencia informal” contrastada con nociones estadísticas clave

La “inferencia informal” no es una inferencia en un sentido estadístico. NO es un enfoque informal de lo que la disciplina de las estadísticas llama inferencia. Presenta un enfoque más bien restringido para hacer inferencias sin vínculos obvios sobre cómo proceder desde allí hasta la inferencia formal (a menos que se omita la inferencia estadística tradicional). En la Tabla 3 (Borovcnik, 2017), bootstrap y re-aleatorización se comparan

con ideas estadísticas clave para resaltar los déficits si se toman como el único enfoque hacia la inferencia estadística.

Table 3. Inferencia "informal" e ideas estadísticas clave

Conceptos	**Re-aleatorización**	**Bootstrap**
Hipótesis – escenarios	Sólo hipótesis de efecto NULO	No es posible conceptualizar
Error de tipo I (error α)	Sí	No
Error de tipo II (error β)	No	No
Hipótesis alternativas	No es posible conceptualizar	No es posible conceptualizar
Métodos	Sólo prueba de significación del efecto NULO	Sin relación con las pruebas de significación

Fuente: datos de la investigación

La inferencia estadística implica un enfoque hipotético Ha habido varios esfuerzos para comparar las diversas escuelas de inferencia a partir de Barnett (1982). Difieren según las hipótesis que se incluyan y cómo se traten.

- Hipótesis alternativas. No hay manera de introducir hipótesis alternativas excepto mediante supuestos probabilísticos y por simulación (o cálculos de probabilidad). Cualquier alternativa no puede ser objeto de un nuevo muestreo, ya que no ha sido objeto de muestreo. Los datos no son adecuados para investigar hipótesis alternativas mediante remuestreo.
- Comparación hipotética de modelos. La modelización implica comparar escenarios (descritos por distribuciones de probabilidad) y no sólo juzgar una hipótesis aislada. Las pruebas de hipótesis son *comparaciones de modelos* (mientras que la familia de modelos es restringida).

La "inferencia informal" marca un cambio de los modelos de probabilidad a los datos Causa un cambio en la connotación de las hipótesis a los hechos (los datos como hechos) y los modelos se absorben en los datos (remuestreados). Sin embargo, la forma en que se juzgan las hipótesis se encuentra en el centro de la inferencia estadística. Ya sea que se haga por métodos clásicos o bayesianos, no existe ningún vínculo entre el remuestreo y estos otros enfoques.

La "inferencia informal" reduce la probabilidad a la concepción del frecuentista Como la inferencia estadística se reduce al remuestreo de los datos, no existe un vínculo obvio para otras connotaciones de probabilidad, aunque son relevantes para las diversas

escuelas de inferencia (véase Barnett, 1982). Esto reduce también la conexión con la teoría de la decisión. Muchos problemas de inferencia estadística se mejoran si se consideran desde una perspectiva teórica de decisión.

El caso de las pequeñas probabilidades En bootstrap, se introduce un nuevo error. Si se trata de las colas de una distribución (pequeñas probabilidades), no se obtendrían datos sobre ellas para que las colas no sean remuestreadas. Si la primera muestra no es lo suficientemente grande, no se pueden muestrear bien más regiones de la distribución. Si la primera muestra es grande, entonces el teorema del límite central da mejores resultados. Si se aplica el método de remuestrar, entonces la variación adicional es grande a menos que se generen más de 10.000 remuestras. Eso lo hace intratable para la enseñanza. La simulación es inapropiada para el caso de las pequeñas probabilidades; un problema que se subestima ampliamente no sólo en la educación estadística (véase Batanero y Borovcnik, 2016).

6. Preocupaciones educativas sobre el remuestreo y las conclusiones

La "inferencia informal" ha sido sugerida como una forma de revolucionar la enseñanza de la inferencia estadística (Cobb, 2007; delMas, 2017; Ben-Zvi, Makar, & Garfield, 2018). Sin embargo, hay cuestiones que deben reconsiderarse no sólo desde el punto de vista educativo.

La realeatorización es una prueba de significación de hipótesis nula Se trata de probar una hipótesis de efecto nulo. La diferencia en la media, por ejemplo, entre dos poblaciones, se compara mediante un reordenamiento aleatorio de las unidades a uno de los grupos. Se genera una distribución para la diferencia de medias, que corresponde a la hipótesis nula de no diferencia entre grupos. No se puede realizar una reatribución de los datos dados para reflejar diferencias específicas entre los dos grupos, de modo que el método no incorpora cuestiones de errores de tipo II. Estamos atascados con una prueba de significación pura con toda la confusión que surge del valor de *P* (ver, por ejemplo, Hubbard & Bayarri, 2003).

Confusión entre los intervalos de bootstrap y de confianza Los intervalos de bootstrap no son fáciles de implementar cuando se aplican correcciones para adaptar su probabilidad de cobertura (para las correcciones sofisticadas, ver Efron & Tibshirani, 1993). Los intervalos de bootstrap no son fáciles de conectar con los temas tradicionales

más adelante. A menudo se omite la especificación "Bootstrap". Los mismos términos para conceptos muy diferentes pueden confundir a los alumnos.

Simplificando el enfoque de una disciplina en el ámbito de la didáctica Brousseau (1984) advierte de las implicaciones de la elementarización en las consideraciones didácticas. Simplificando, se puede llegar a enseñar un nuevo objeto que ni siquiera existe en matemáticas ("glissement methadidactique"). Biehler (2014) afirma que: "[...] el razonamiento inferencial formal como tal es controversial en sí mismo [...] Esto plantea preguntas con respecto a qué punto de vista de la inferencia formal [...] diseñamos [...] actividades de inferencia informal." La "inferencia informal" va más allá de la exploración informal de modelos probabilísticos; su objetivo es reemplazar la inferencia estadística tradicional. Las ventajas de un enfoque intuitivo hacia la inferencia se pierden si no se ve como una etapa transitoria en la enseñanza y el aprendizaje.

Temas a reconsiderar para un enfoque de "inferencia informal"

- La "inferencia informal" es muy convincente pero conduce a una metodología restringida que es un subconjunto estricto de la inferencia estadística.
- Los intervalos de bootstrap difieren de los intervalos de confianza clásicos; para adaptarlos, se requieren métodos sofisticados para que su ventaja intuitiva se pierda (ver Efron & Tibshirani, 1993, Lunneborg, 2000, o Howell, s.f.). La realeatorización no permite tener en cuenta los errores de tipo II, de modo que la prueba se reduce a la prueba de la significación, que es muy controvertida.
- El bootstrap falla con probabilidades pequeñas (de cola), que no están cubiertas en los datos a menos que la muestra sea muy grande, de modo que no se puedan remuestrear. Se trata de *modelizar pequeñas probabilidades* y riesgos en lugar de tratar los datos de forma inteligente.
- El enfoque de "inferencia informal" reduce de manera inviable la concepción de probabilidad al aspecto de frecuencia a medida que la inferencia se reduce al remuestreo y a las frecuencias relativas de los datos artificiales. Aquí nos enfrentamos al dilema formulado por Carranza y Kuzniak (2008) con un enfoque puramente frecuentista de la estocástica, donde los problemas aplicados requieren un enfoque teórico de decisión y una connotación cualitativa (subjetivista) de probabilidad.
- ¿Cómo adaptar el currículo de probabilidad? ¿Debemos dejar atrás la distribución normal? El modelo probabilístico utiliza muchas otras distribuciones (para el análisis

de riesgos, fiabilidad, etc.). Cómo tratar con otros enfoques e interpretaciones (por ejemplo, Bayes).

- ¿Cómo continuar el plan de estudios dentro de un entorno de este tipo? No hay camino para pasar del remuestreo a la teoría de la decisión, que está mucho más cerca de problemas de interés cotidiano, pero también de muchas aplicaciones, como en la medicina o la economía. No hay conexión entre el remuestreo y los métodos Bayes (aunque los bayesianos usan mucho la simulación), que constituyen un enfoque relevante para los problemas reales.
- La comprensión conceptual difiere de un acceso más fácil y de la resolución de tareas. Además, la modelización se absorbe en la simulación. Esto puede resultar en datos como hechos, mientras que los modelos representan una forma hipotética de pensar. Sería mejor enseñar la inferencia clásica y bayesiana en paralelo para resaltar las diferencias en los conceptos y así permitir una adquisición de conceptos sostenible (ver Vancsó, 2009).

La "inferencia informal" reduce la visión de la modelización probabilística más adelante. Las cuestiones educativas generales que surgen con el enfoque son: Los estadísticos utilizan modelos cada vez más sofisticados, pero ni siquiera hemos conseguido enseñar lo más sencillo. ¿Cómo se podrá desafiar a los expertos si se los educa sólo en esta vía lateral? ¿Debería ser la estadística de secundaria un campo que no tiene casi nada en común con la estadística universitaria y las abundantes aplicaciones que se dan en todos los sectores de la ciencia, la economía, y de la vida pública y privada? ¿Vamos a distraer a la gente de evaluar críticamente y desafiar esas aplicaciones de las estadísticas? Se sugiere utilizar el remuestreo (bootstrap y realeatorización) como una etapa transitoria para la inferencia estadística y centrarse en las formas de elementarizar toda la complejidad de la inferencia estadística.

Referencias

BARNETT, V. (1982). Comparative statistical inference (2nd ed.). New York: Wiley.

BATANERO, C. & BOROVCNIK, M. (2016). *Statistics and probability in high school*. Rotterdam: Sense Publishers.

BEN-ZVI, D., MAKAR, K., & GARFIELD, J. (2018). *International handbook of research in statistics education*. Cham, Switzerland: Springer International.

BIEHLER, R. (2014). On the delicate relation between informal statistical inference and formal statistical inference. In K. Makar (Ed.), *Proceedings of the Ninth International Conference on Teaching Statistics*. The Hague: ISI.

BOROVCNIK, M. (1996). Trends und Perspektiven in der Stochastik-Didaktik [Trends and perspectives in the didactics of stochastics]. In: G. Kadunz, H. Kautschitsch, G. Ossimitz, & E. Schneider (Eds.): Trends und Perspektiven (pp. 39-60). Wien: HPT.

BOROVCNIK, M. (2006a). Daten – Zufall – Resampling [Data–randomness–resampling]. In J. Meyer (Ed.), *Anregungen zum Stochastikunterricht Band 3. Tagungsband 2004/5 des Arbeitskreises "Stochastik in der Schule"* (p. 143-158). Berlin: Franzbecker.

BOROVCNIK, M. (2006b). On outliers, statistical risks, and a resampling approach towards statistical inference. *Paper presented at* CERME V). Larnaka.

BOROVCNIK, M. (2017). Informal inference – Some thoughts to reconsider. In *Proceedings of the 61st World Statistics Congress*. The Hague: ISI.

BOROVCNIK, M. (n.d.). *Spreadsheets in Statistics Education*. Online: wwwg.uni-klu.ac.at/stochastik.schule/Boro/index_inhalt.

BROUSSEAU, G. (1984). Le rôle central du contrat didactique dans l'analyse et la construction des situations. Non-published paper.

CARRANZA, P. & KUZNIAK, A. (2008). Duality of probability and statistics teaching in French education. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education.* Monterrey: ICMI and IASE.

COBB, G.W. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education* 1(1).

DELMAS, R. (2017). A 21st century approach towards statistical inference – Evaluating the effects of teaching randomization methods on students' conceptual understanding. In *Proceedings of the 61st World Statistics Congress*. The Hague: ISI.

EFRON, B., & TIBSHIRANI, R. J. (1993). *An introduction to the bootstrap*. New York – London: Chapman & Hall.

ENGEL, J. (2010). On teaching bootstrap confidence intervals. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society.* Voorburg: International Statistical Institute.

GIGERENZER, G. (2002). Calculated risks: How to know when numbers deceive you. New York: Simon & Schuster.

HOWELL, D. (n.d.). Resampling statistics: Randomization & Bootstrap. *Statistical page of D. Howell*. Online: www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html.

HUBBARD, R. & BAYARRI, M. J. (2003). Confusion over measures of evidence (p) versus errors (α) in classical statistical testing. *The American Statistician 57*(3), 171-182.

LUNNEBORG, C. E. (2000). *Data analysis by resampling: concepts and applications*. Pacific Grove, CA: Duxbury Press.

NEYMAN J. & PEARSON E. (1933). On the problem of most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A, 231,* 289-337.

NOETHER, G. (1967). *Elements of noparametric statistics*. New York: Wiley.

ROSSMAN, A. J. (2008). Reasoning about informal statistical inference: one statistician's view. *Statistics Education Research Journal 7*(2), 5-19.

STOHL LEE, H., ANGOTTI, R. L., & TARR, J. E. (2010). Making comparisons between observed data and expected outcomes: students' informal hypothesis testing with probability simulation tools. *Statistics Education Research Journal 9*(1), 68-96.

VANCSÓ, Ö. (2009). Parallel discussion of classical and Bayesian ways as an introduction to statistical inference. *International Electronic Journal of Mathematics Education 4*(3), 291-322.

Texto recebido: 14/04/2019
Texto aprovado: 14/04/2019