REVIEWER A

1. COMENTARIOS SOBRE ASPECTOS DE CONTENIDO

Líneas 111-112. En los estudios de simulación es habitual utilizar más de un escenario relacionado con el tamaño del efecto simulado. ¿Por qué una diferencia de 15 puntos (1,5 desviaciones típicas), es decir, un efecto de tamaño muy grande? ¿Por qué no ver qué pasa con efectos de tamaño leve (0,2 desv. típ.), moderado (0,5 desv. típ.) y grande (0,8 desv. típ.)?

Líneas 115-116. En los estudios de simulación se suelen considerar diferentes grados de asimetría (por ejemplo, leve, moderada y severa). ¿Por qué aquí se simula únicamente un grado de asimetría? Por otro lado, cuando las medias de X e Y son iguales (primer escenario), no importa que la asimetría simulada sea positiva o negativa; pero cuando la media de Y es mayor que la de X (segundo escenario), ¿por qué no se simulan ambas formas de asimetría? Por cierto, cuando se asume que las medias de X e Y son iguales, es redundante simular As_x=.0_As_y=.8 y As_x=.8_As_y=.0.

Líneas 136-140. ¡¡Ojo!! No se están calculando intervalos de confianza para la diferencia entre dos medias o medianas (que es la estrategia convencional y, en mi opinión, apropiada), sino para cada media o mediana por separado. ¿Por qué se hace así?

Línea 162-165. Esto no es cierto. La tasa de falsos positivos se dispara cuando la asimetría afecta solo a una de las dos distribuciones, no a las dos.

Línea 199-200.l Esta afirmación no es correcta. El estadístico U cumple con esta afirmación más de lo que lo hacen otros estadísticos.

Líneas 202-206. Este comentario no se corresponde con los datos de la Tabla 2.

Líneas 217-219. La primera afirmación de la discusión no es cierta. Los resultados de la tabla 1 indican que, con "k Md CI", la tasa de falsos positivos es demasiado alta cuando n vale 50 (además, es demasiado baja cuando n vale 10). Y Con "MJ Md CI" la tasa es demasiado alta cuando n vale 30 y 50, etc.

Línea 227. Me parece exagerado afirmar que "their performance is almost perfect".

Línea 236-237. ¿Por qué se recomienda esto? La tasa de errores Tipo I asociada a este procedimiento es inaceptable en varias condiciones. Y la tasa de errores Tipo II también es inaceptable con n = 10 (hay que tener en cuenta que se está simulando un efecto de tamaño muy grande).

Línea 237-238. ¿Por qué se recomienda esto? La tasa de errores Tipo I asociada a este procedimiento es inaceptable, excepto con n = 10. Y la tasa de errores Tipo II es inaceptable con n = 10.

Líneas 240-242. Esto sería cierto si no fuera por lo que ocurre con la tasa de errores Tipo II con n = 10.

Línea 246. En lo relativo a los errores Tipo I, los resultados indican que el estadístico t de Yuan-Welch se comporta mejor que el estadístico U de Mann-Whitney. Pero en lo relativo a los errores Tipo II, el estadístico U de Mann-Whitney se comporta mejor que el de Yuan-Welch. ¿Por qué entonces se dice que el estadístico de

Yuan-Welch ofrece un buen comportamiento general (línea 241) y que el estadístico de Mann-Whitney se comporta de forma incorrecta?

Línea 257-258. Como ya he señalado, los autores no construyen intervalos de confianza para la diferencia entre dos medias o medianas, sino para cada media o mediana individualmente considerada. Esto no es "similar" (y desde luego no es equivalente) a un contraste de hipótesis.

Tabla 1. Resulta bastante sorprendente que el comportamiento del estadístico t empeore cuando aumenta el tamaño muestral. Ayudaría a confiar en esto resultados una tabla con información sobre las características de las distribuciones simuladas con cada tamaño muestral. Se echa en falta saber cuánto vale la media de las distribuciones simuladas, la desviación típica promedio en cada condición simulada, el grado de asimetría promedio en cada condición simulada, etc. Sin esta información es difícil formarse una idea acerca de la calidad de la simulación.

2. COMENTARIOS SOBRE ASPECTOS FORMALES

Línea 27. Dice "which is often taken" y debería decir "which is often assumed" o "which is often taken for granted". En general, el ingles de este manuscrito necesita una revisión importante.

Line 41. Dice "...such as the effect size test". Por lo general, las medidas del tamaño del efecto no van acompañadas de un contraste.

Línea 43. La cita "Willinas y Cumming, 2005" no se ha incluido en las referencias.

Línea 45. En "…Task Force for Statistical…" cambiar "for" por "on".

Línea 45. El año de la cita es 1999, no 1995.

Línea 58. La frase "cases in which the analysis is based on the mean" no está bien construida. Quizá sobra "is".

Líneas 59 y 60. Peró y Guàrdia están citados de forma diferente en estas dos referencias (tanto aquí como en las referencias finales). En la línea 220 ocurre algo parecido.

Línea 60. Creo que el significado de la siguiente frase no está claro: "… show a high performance to recognize the conditions around the true H0 for the median confidence intervals when comparing two independent groups, but a bad performance in the conditions around the false H0"

Línea 63. Tampoco ltengo claro el significado de la siguiente frase: "The main aim of this work is to study the goodness of confidence interval comparison for two independent groups when the distributions are asymmetrical." Quizá habría que utilizar "benefit" o "superiority" en lugar de "goodness".

Líneas 69 y 70. En la ecuación aparece "s minúscula" y en la línea 70 aparece "S mayúscula".

Línea 70. "S_w" debe colocarse en cursiva.

Línea 71. Yo creo que la expresión "trimmed quantiles"no es correcta. Lo que se recorta no son los cuantiles, sino los casos que no alcanzan o que superan ciertos cuantiles. Creo que es preferible referirse a gamma como "la proporción de casos recortados".

Línea 71. Utilizar "tail" en lugar de "cue".

Línea 74. Falta un espacio antes de “Bland” y antes de “and”.

Línea 80. La ecuación de “SE(median)” hay que separarla bastante más de la anterior.

Líneas 80 y 82. Evitar utilizar el punto para representar el producto. Particularmente cuando es innecesario.

Línea 88. Las expresiones “inferior limit” y “superior limit” no son correctas. Utilizar “lower limit” y “upper limit”.

Línea 90. Poner k en cursiva.

Líneas 91 y 101. Quitar el cero a los valores decimales menores que 1.

Línea 102. Se habla de cinco intervalos de confianza para la media y este es el cuarto. ¿Dónde está el quinto?

Línea 114. Poner “n” en cursiva (al final de la línea).

Línea 120. Falta un paréntesis para cerrar la sentencia.

Línea 130. ¿Quizá habría que decir ”kurtosis” en lugar de “pointedness”?

Línea 145. “…five used procedures.” ¿No son siete?

Línea 167. “…the error rate is below”. ¿Cambiar “below” por “around”?

Líneas 208. Quitar las comillas que van después de group.

Línea 224. ¿Qué significa una “interpretación parsimoniosa”?

Referencias

- Incluir todas las citas del texto.

- Revisar el formato, el espaciado entre iniciales, el tipo de guión utilizado en las páginas, etc. En el manuscrito he marcado varias erratas.

REVIEWER B

El documento enviado muestra una investigación relevante e interesante en el que se analizan varios métodos robustos para comparar dos medias poblacionales independientes. Es interesante porque los autores analizan métodos robustos modernos en contraposición a las técnicas al uso que se siguen utilizando masivamente en investigación. Es relevante porque tiene un carácter muy aplicado para que los investigadores puedan escoger entre métodos alternativos. En el estudio se plantean varios escenarios de simulación en el que se comparan las técnicas convencionales de comparación de medias (T de Student, U de Mann-Whitney e Intervalos de confianza para la diferencia de medias) con métodos robustos más modernos (medias truncadas, la T de Yuen-Welch y varios métodos de estimación de intervalos de confianza para la mediana).

Considero que es una contribución interesante sobre la temática. Sin embargo, antes de recomendar el artículo para su publicación pediría a los autores tres cosas: (1) que se planteen hacer nuevos escenarios de simulación dado que en las conclusiones hay varios aspectos cuestionables, (2) revisar algunas de las conclusiones sobre los resultados obtenidos y (3) corregir algunas cuestiones menores.

1. Plantear nuevos escenarios de simulación. En las conclusiones, párrafo tres, se proponen algunas estrategias generales basadas en los resultados obtenidos. Dado que los escenarios de simulación planteados contienen siempre varianzas poblacionales iguales y tamaños muestrales iguales creo relevante plantear nuevas condiciones de simulación en las que se den varianzas distintas (puede ser 1:3 y 3:1) y tamaños muestrales distintos (las condiciones 10-30 y 30-10. Creo que esto añadiría tres tablas más al experimento. En las mismas 2 tablas ya presentadas (tabla 1: medias iguales; tabla 2: medias distintas) se pueden añadir las 8 filas correspondientes a las condiciones 10-30 y 30-10. Se generaría una tabla para medias poblacionales iguales con varianzas 1:3 (la tabla para varianzas 3:1 y medias iguales sería redundante y no haría falta incluirla), y otras dos tablas para medias distintas con varianzas 1:3 y 3:1. ¿Por qué esta petición? Porque al sugerir estrategias robustas para comparar dos medias, el artículo gana mucho en relevancia si se plantean situaciones corrientes dentro de la investigación (varianzas distintas y tamaños muestrales distintos). Es cierto que con estos nuevos escenarios no se sondea exhaustivamente la cuestión de varianzas distintas ni tamaños muestrales distintos, pero sí que añade conocimiento sobre tales asuntos y habrá indicios fuertes de cómo se comportan los métodos propuestos por los autores en situaciones más reales. Convendría ver, además, cómo se comporta el estadístico T de Yuen-Welch bajo estas circunstancias, dado que hay trabajos en los que se cuestionan su robustez precisamente con tamaños muestrales distintos y varianzas distintas. Podría ser que la verdadera ventaja de los métodos de intervalos de confianza para las medianas se den en condiciones de no igualdad de varianzas o tamaños distintos.

2. En relación a algunas conclusiones cuestionables sobre los resultados señalar que:

a. En el tercer párrafo como estrategia general se proponen los métodos Adaptive-kernel o Trimmed mean. Yo creo que el procedimiento Adaptive-Kernel y Trimmed mean tienen errores tipo I inaceptables prácticamente en todas las condiciones (especialmente con muestras grandes). Por ejemplo, funcionan peor que la U de Mann-Whitney. Además, Trimmed Means tiene otros inconvenientes cuando las distribuciones son asimétricas – pueden quedar sesgados los valores de los intervalos de confianza (Bonett and Price, 2002 citado por ustedes). Incluso, yo hago otra lectura de los datos y creo que el mejor método de intervalos de confianza para la mediana es Binomial Md CI, ya que presenta un error tipo I adecuado y detecta bien diferencia de medias excepto cuando n = 10.

b. Asimismo, cabe preguntarse si la U de Mann-Whitney tiene tan malas prestaciones (cuarto párrafo de la discusión) como se deja ver en la discusión. De hecho, muestra la misma potencia que los mejores procedimientos modernos analizados. En los resultados mostrados únicamente funciona mal porque tiene un error tipo I alto cuando los dos grupos tienen simetrías diferentes.

3. En relación a algunas cuestiones menores:

a. Preguntar si tiene sentido hacer las condiciones AS_X=.0-AS_Y=.8 y AS_X=.8-AS_Y=.0 con medias iguales, ya que es un escenario duplicado. Sin esta redundancia tal vez puedan quitarse filas y simplificarse la tabla de igualdad de medias poblacionales. De todas formas, si los autores consideran que por coherencia deben permanecer estas dos condiciones me parecería bien. Simplemente el comentario es para que valoren esto.

b. Sobre las referencias. Hay una inconsistencia en el año de publicación en el artículo cuando se cita Wilkinson and the Task Force for Statistical Inference, donde pone 1995 y en la referencia donde pone 1999. Tukey 1977 aparece en las referencias pero no se cita en el artículo y lo mismo sucede con Wilcox 2006.

Ricardo Olmos
Universidad Autónoma de Madrid