Facultad de Ciencias Naturales y Exactas
Universidad del Valle

Comparación de la eficiencia de las pruebas de hipótesis e intervalos de confianza en el proceso de inferencia. Estudio sobre medias

Pablo Javier Flores Muñoz
Escuela Superior Politécnica de Chimborazo

Recibido: 23 de agosto de 2018 Aceptado: 11 de diciembre 2018

Pag.65-85

Resumen

Estudios pertinentes aseguran que cuando se requiere inferir sobre algún parámetro, los intervalos de confianza son una alternativa más adecuada que las pruebas de hipótesis. El presente estudio mide la eficiencia de inferencia de estas dos metodologías, aplicadas sobre contrastes de medias, comparando la probabilidad de cometer un error y un acierto en cada una de ellas ─Estas probabilidades son estimadas mediante un proceso de simulación estocástica de Montecarlo. Tomando en cuenta este criterio, se encontró que ambos procesos inferenciales son igualmente efectivos, esto es, presentan igual probabilidad de equivocarse e igual probabilidad de acertar, razón por la cual no se puede asegurar que un procedimiento es mejor que otro. Si bien es cierto, los intervalos de confianza proporcionan mayor información sobre la medición de diferencias significativas entre los parámetros comparados, esto no se debe confundir con mayor eficiencia. Parece ser que problema no son los modelos de pruebas de hipótesis, sino más bien el planteamiento que tradicionalmente estos presentan. Para superar este problema y conseguir que la metodología basada en el valor P proporcione la misma información inferencial que los intervalos de confianza se recomienda utilizar un enfoque de pruebas de hipótesis basado en equivalencia.

Palabras clave: pruebas de hipótesis, intervalos de confianza, equivalencia, simulación.

doi: 10.25100/rc.v22i2.7921

Comparison of the Efficiency of Hypotheses Tests and Confidence Intervals in the Inference Process. Study on Means

Abstract

Relevant studies ensure that when you need to infer about a parameter, confidence intervals are more appropriate alternatives than hypothesis testing. This study measures the efficiency of inference of these two methodologies, applied on contrasts of means, comparing the probability of an error and a success in each one of them ─These probabilities are estimated through a stochastic Monte Carlo simulation process. Taking into account this criterium, it was found that both inferential processes are equally effective, for this reason, it cannot be guaranteed that one procedure is better than another. Confidence intervals provide more information about the measurement of significant differences between the parameters compared, but this should not be confused with greater efficiency.

The problem is not hypothesis testing models, but rather the approach they traditionally present. To overcome this problem and get than methodology based on P value provides the same inferential information as the confidence intervals, it is recommended to use an approach of hypothesis testing based on equivalence.

Keywords: hypothesis test, confidence intervals, equivalence, simulation.

1 Introducción

Los procesos más utilizados por investigadores para determinar diferencias significativas entre un parámetro poblacional y un valor hipotético se basan en procesos clásicos de inferencia estadística. A pesar de que los modelos de pruebas de hipótesis son muy usados para este fin, existen múltiples investigaciones que desaconsejan su uso, argumentando la presencia de graves deficiencias y dudosa utilidad [1,2], algo que parece no ocurrir cuando se usan modelos basados en intervalos de confianza, los cuales se asegura, son procedimientos que aventajan el uso del valor P como instrumento inferencial para muchos tipos de estudios observacionales y experimentales investigados principalmente en las ciencias médicas y sociales [3–5]. El argumento presentado por estos estudios ha tenido tal impacto que algunas revistas que publican artículos sobre estas ciencias alientan a sus autores a trabajar con intervalos de confianza en lugar de pruebas de hipótesis [6].

El principal argumento por el cual se estimula el uso de intervalos de confianza en lugar del valor P es que los investigadores deben estar interesados en determinar el tamaño de la diferencia entre los parámetros comparados en lugar de una simple respuesta dicotómica que indique la existencia o no de dicha diferencia. En este sentido, con el uso del valor P nunca podremos determinar conclusiones más allá de aquella que nos permita determinar (o no) la existencia de diferencias significativas, independientemente de cuán grande sea esta, mientras que los intervalos de confianza nos mostrarán la magnitud de estas diferencias encontradas [7, 8]. A partir de este pensamiento se han escrito varias críticas sobre el uso de las pruebas de hipótesis en la valoración de la importancia de los resultados investigativos, dos muy famosas rezan de la siguiente manera "*Adiós, p menor del 0.05", equivoco y traicionero compañero de viaje. Tus efectos colaterales y toxicidad intracerebral son demasiado grandes para compensar cualquier beneficio que pudieras aportar*" [9] y "*Las típicas aseveraciones ($p<0.05, p>0.05$ o p= Nivel de significancia) dan poca información sobre los resultados de un estudio y se basan en el consenso arbitrario de utilizar el nivel de significación estadístico de 5% para definir los posibles resultados: significativo o no significativo. Esto no sirve para nada y, además, favorece la vagancia intelectual. Incluso cuando se indica el valor p en concreto, no se proporciona información alguna sobre las diferencias en los grupos estudiados*" [10].

Con lo expuesto hasta ahora, concordamos en pensar que el valor P proporciona menor información de la estimación de un parámetro que los intervalos de confianza. Aún más, nos gustaría aportar a estas ideas, ciertos análisis muy interesantes sobre el mal planteamiento que parecen presentar las pruebas de hipótesis tradicionalmente usadas. Por ejemplo, hablando de test de hipótesis para probar perfecta normalidad, George Box en 1979 mencionó: "*en la vida real no existe una distribución perfectamente normal, sin embargo, con modelos, que se sabe que son falsos, a menudo se pueden derivar resultados*

que coinciden, con una aproximación útil a los que se encuentran en el mundo real" [11]. Este pensamiento, fácilmente se puede trasladar a cualquier hipótesis nula que pretenda probar una perfecta igualdad de parámetros (medias, varianzas, proporciones, …) lo cual nos lleva a pensar que probar una hipótesis de perfecta igualdad carece de sentido, en su lugar parece ser que lo verdaderamente importante es saber si la aproximación de la prueba es lo suficientemente buena como para ser considerada útil. El criterio de Cochran establece que un test de hipótesis es considerado un buen modelo cuando la Probabilidad de cometer un error tipo I se aleja una distancia máxima del 20% de α por encima o por debajo de este nivel de significancia [12]. Exactamente el mismo criterio es utilizado para definir la robustez de una prueba de hipótesis [13].

Además, cabe recalcar la dificultad lógica al concluir estos tipos de test tradicionales, en el sentido de que cuando no rechazamos una hipótesis nula no estamos concluyendo igualdad entre los parámetros comparados, en su lugar lo único que se concluye es ausencia de evidencia para determinar diferencias significativas entre los parámetros contrastados, lo cual no debe ser confundido con una significativa igualdad, o como en este mismo sentido se menciona "*Ausencia de evidencia no es evidencia de ausencia*" [14] o "*Una diferencia no significativa no debe ser confundida con una significativa homogeneidad*" [15]. Si todas estas dificultades lógicas se producen al no rechazar una hipótesis nula tradicional de perfecta igualdad, por otra parte, el rechazarla solo podría estar probando una simple diferencia irrelevante entre los parámetros comparados, insuficiente como para ser considerada altamente significativa [16], lo cual nos lleva nuevamente a pensar la importancia de medir el tamaño de la diferencia entre los parámetros comparados.

A pesar de toda esta evidencia teórica en contra del uso de los test de hipótesis tradicionales, no estamos de acuerdo en afirmar que estadísticamente hablando se deba considerar que los intervalos de confianza son modelos inferenciales mejores que los test de hipótesis. A nuestro criterio, lo único que se ha demostrado con los estudios previos es que, a diferencia de las pruebas de hipótesis, el proceso de inferencia por intervalos de confianza proporciona mayor información de la magnitud de la diferencia entre el parámetro comparado y su valor hipotético, lo cual no debería ser confundido con un análisis objetivo de la comparación de la eficiencia de estos procedimientos. Si el problema es determinar la magnitud de la diferencia entre parámetros comparados, un enfoque de pruebas de hipótesis diferente al tradicional sería la solución. Este enfoque debe contener una hipótesis a probar en la que se establezca la existencia de identidad entre parámetros dentro de un intervalo de interés alrededor de la perfecta igualdad. Este tipo de pruebas de hipótesis existen, se denominan "test de hipótesis de equivalencia" y fueron planteadas por Steffan Wellek en el año 2010. El autor usa el término de equivalencia como una forma dilatada de una relación de semejanza entre los parámetros analizados y considera que esta dilatación en la hipótesis de equivalencia se induce al añadir en la hipótesis tradicional una zona de irrelevancia (que podría estar dada por la magnitud de la diferencia entre tratamientos comparados que se quieran probar) alrededor de la correspondiente región o punto en el espacio paramétrico que denota la igualdad perfecta. Esta zona de irrelevancia, cuyos límites son constantes positivas deben ser asignadas a priori y sin mayor conocimiento de la muestra. Inverso al test de hipótesis tradicional donde en la hipótesis nula, se especifica la igualdad de los

parámetros comparados. Este tipo de pruebas se plantean de tal forma que la hipótesis nula establece la no equivalencia, mientras que la alternativa establece la equivalencia. Este cambio de interés en la investigación conduce a diseñar un estudio que pretende demostrar ausencia de una diferencia relevante entre los efectos de dos o más tratamientos, es decir, equivalencia [17]. Además, es muy importante notar que las pruebas de equivalencia no buscan probar una perfecta igualdad de parámetros a compararse, en lugar de esto, la intención es declarar el cumplimiento de la relación de identidad incluso para desviaciones que pueden ser consideradas como irrelevantes o despreciables. Parece ser que, con estas precisiones, este enfoque supera las dificultades lógicas de los test tradicionales analizadas en los párrafos anteriores.

Aunque el problema de incluir la magnitud de la diferencia entre parámetros en un modelo de prueba de hipótesis parece estar superado con el uso de los test de equivalencia, aún nos resta comprobar (desde un punto de vista más objetivo que el planteado en investigaciones previas) si los modelos de intervalos de confianza pueden ser considerados mejores que los test estadístico tradicionales. Para ello, una medición y comparación de las probabilidades de error y/o acierto de los dos modelos nos parece un criterio estadístico no subjetivo. Por una parte, una estimación de la probabilidad de cometer un error al realizar un intervalo de confianza se puede obtener a partir de la proporción de veces que un parámetro cae afuera del intervalo cuando realmente debería estar adentro, el correspondiente estimador en una prueba de hipótesis es la proporción de veces que se rechaza la hipótesis nula de perfecta igualdad cuando esta es verdadera, lo cual se conoce como la Probabilidad de cometer un Error de tipo I. Por otra parte, una estimación de la probabilidad de cumplir un acierto en intervalos de confianza se puede ver como la proporción de veces que cierto parámetro cae fuera del intervalo cuando realmente debería hacerlo, el correspondiente estimador para pruebas de hipótesis es la proporción de veces que se rechaza la hipótesis nula de perfecta igualdad cuando esta es falsa, lo cual se conoce como la potencia de la prueba.

El presente estudio estima mediante un proceso de simulación las probabilidades de éxito y error descritas. Delimitamos estas mediciones a inferencias sobre una y dos medias cuando se asegura normalidad (Z test, t – Student test y Welch test), esto debido a que en la práctica estas pruebas suelen ser las más utilizadas.

2 Materiales y métodos

Delimitando el parámetro a estudiar y basados en la estimación de las probabilidades de éxito y error que presentan las técnicas de Intervalos de Confianza y Pruebas de Hipótesis, pretendemos determinar si existe evidencia para considerar una técnica mejor que otra, lo cual nos parece un criterio menos subjetivo que aquellos presentados en la sección de Introducción. La Tabla 1 muestra estas técnicas con sus correspondientes estadísticos [18], usados para estimar las probabilidades que deseamos comparar: las probabilidades deseadas son obtenidas a partir de funciones de simulación propias desarrolladas en el software estadístico R [19]. Con el fin de que estas funciones, computacionalmente hablando sean lo menos pesadas posibles, se crearon funciones auxiliares propias (Anexo A), las cuales se usaron en lugar de las existentes en las librerías del software. Estas funciones auxiliares calculan los intervalos de confianza y los estadísticos que se encuentran en La Tabla 1 y se las puede visualizar de acuerdo al detalle especificado en la misma tabla.

Tabla 1. Modelos *de inferencia estadística usados para estimar probabilidades de acierto y Probabilidades de error.*

Parámetro	Premisa	Intervalo de Confianza (IC) (ANEXO A)	Prueba de Hipótesis (PH)	Estadístico correspondiente (ANEXO A)
$\boldsymbol{\mu}$	σ conocida	$\bar{x} \pm Z_{\propto/2} \frac{\sigma}{\sqrt{n}}$ **(A.1)**	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ **(A.6)**
	σ desconocida	$\bar{x} \pm t_{\propto/2} \frac{S}{\sqrt{n}}$ **(A.2)**	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$t = \frac{\bar{x}-\mu}{S/\sqrt{n}} \sim (n-1)\ g.l$ **(A.7)**
$\boldsymbol{\mu_1 - \mu_2}$	σ_1^2, σ_2^2 conocidas	$(\bar{x}_1 - \bar{x}_2) \pm Z_{\propto/2} \sqrt{\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n2}}$ **(A.3)**	$H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 \neq 0$	$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n2}}}$ **(A.8)**
	σ_1^2, σ_2^2 desconocidas e iguales	$(\bar{x}_1 - \bar{x}_2) \pm t_{\propto/2}\, Sp \sqrt{\frac{1}{n1} + \frac{1}{n2}}$ **(A.4)**	$H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 \neq 0$	$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{Sp \sqrt{\frac{1}{n1} + \frac{1}{n2}}} \sim$ $(n_1 + n_2 - 2) g.l$ **(A.9)**
	σ_1^2, σ_2^2 desconocidas y diferentes	$(\bar{x}_1 - \bar{x}_2) \pm t'_{\propto/2} \sqrt{\frac{S_1^2}{n1} + \frac{S_2^2}{n2}}$ **(A.5)**	$H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 \neq 0$	$t' = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n1} + \frac{S_2^2}{n2}}} \sim v\ g.l$ **(A.10)**

Las probabilidades de éxito y error para intervalos de confianza y pruebas de hipótesis se obtuvieron a través de la función propia de simulación "*onemean*" (Anexo B) en el caso de una muestra y "*twomean*" (Anexo C) en el caso de dos muestras.

En el caso de una muestra, el proceso de simulación implementado en la función "*onemean*" y sus argumentos se detallan a continuación: A través del comando rnorm del paquete base de R, se crea una variable aleatoria normal de tamaño *n*, con media *mu* y desviación estándar *sigma*. A partir de esta variable, se calculan los intervalos de confianza y los estadísticos correspondientes según sea el caso que indique el argumento *var.known*, esto es, inferencias cuando conocida (Z test) en el caso de que se indique un valor TRUE al argumento o inferencias cuando σ desconocida (t - Student) en el caso de tomar el valor FALSE.

Cuando los argumentos *mu* y *mu.hip* (media hipotética o de comparación) son iguales, es decir la media hipotética coincide con la media poblacional que genera la muestra, las estimaciones obtenidas a partir de *m = 1000000* (un millón) de réplicas de simulación, corresponden a las probabilidades de cometer un error, las cuales son calculadas en el caso de intervalos de confianza como la proporción de veces que la media hipotética está afuera de los intervalos generados (cuando en realidad debería estar adentro) y para pruebas de hipótesis como la proporción de veces que se rechaza la hipótesis nula de igualdad $H_0: \mu = 0$ (cuando en realidad, teóricamente la muestra sí se formó con esta media poblacional).

Cuando los argumentos *mu* y *mu.hip* son diferentes, es decir la media hipotética es distinta a la media poblacional que genera la muestra, las estimaciones obtenidas a partir de *m = 1000000* (un millón) de réplicas de simulación, corresponden a las probabilidades de acierto, las cuales son calculadas en el caso de intervalos de confianza como la proporción de veces que la media hipotética está afuera de los intervalos generados (siendo esto lo correcto) y para pruebas de hipótesis como la proporción de veces que se rechaza la hipótesis nula de igualdad $H_0: \mu = 1$ (cuando en realidad, teóricamente la muestra no se formó con esta media poblacional).

En el caso de dos muestras, el proceso de simulación implementado en la función *"twomean*" y sus argumentos se detallan a continuación: A través del comando *rnorm* del paquete base de R, se crean dos variables aleatorias normales de tamaño *n1* y *n2* respectivamente, estas variables tienen una diferencia teórica de medias dado por el argumento *diffmean* (este valor siempre será cero para asegurar igualdad poblacional de las medias comparadas, caso contrario denotaría diferencia) y un aseguramiento de la homocedasticidad dado por el argumento *ratioSigma* (un valor de 1 denota perfecta homocedasticidad y mientras más nos alejemos de 1 se representa una heterocedasticidad más fuerte). A partir de estas variables, se calculan los intervalos de confianza y los estadísticos correspondientes según sea el caso que indique el argumento *var.known*, esto es inferencias cuando σ_1 σ_2 son conocidas (Z - test) en el caso de que se indique un valor TRUE al argumento, o inferencias cuando σ_1 σ_2 son desconocidas en el caso de tomar el valor FALSE. Es importante recordar, que cuando estas desviaciones teóricas son desconocidas el uso del t – Student o del test de Welch dependerá del cumplimiento del supuesto de homocedasticidad.

Cuando los argumentos *diffmean* y *dm.hip* (diferencia de medias hipotética o de comparación) son iguales, es decir la diferencia hipotética de medias coincide con la diferencia de medias poblacionales que generan las muestras, las estimaciones obtenidas a partir de *m = 1000000* (un millón) de réplicas de simulación, corresponden a las probabilidades de cometer un error, las cuales son calculadas en el caso de intervalos de confianza como la proporción de veces que la diferencia de medias hipotética está fuera de los intervalos generados o no contiene el cero (cuando en realidad debería estar dentro o contener el cero) y para pruebas de hipótesis como la proporción de veces que se rechaza la hipótesis nula $H_0: \mu_1 - \mu_2 = 0$ (cuando en realidad, teóricamente las muestras sí se formaron con estas dos medias poblacionales iguales).

Cuando los argumentos *diffmean* y *dm.hip* son diferentes, es decir la diferencia hipotética de medias es diferente a la diferencia de medias poblacionales que generan las muestra, las estimaciones obtenidas a partir de *m = 1000000* (un millón) de réplicas de simulación, corresponden a las probabilidades de acierto, las cuales son calculadas en el caso de intervalos de confianza como la proporción de veces que la diferencia de medias hipotética está fuera de los intervalos generados (siendo esto lo correcto) y para pruebas de hipótesis como la proporción de veces que se rechaza la hipótesis nula $H_0: \mu_1 - \mu_2 = 1$ (cuando en realidad, teóricamente las muestras no se formaron con estas dos medias poblacionales distintas).

Finalmente, para ambas funciones principales, los argumentos *alpha* y *seed* indican respectivamente el nivel de confianza de las pruebas y la semilla utilizada con el fin de que todas las simulaciones sean replicables con los mismos resultados. Para las probabilidades de error, los valores de la media teórica y la media hipotética fueron ambas cero, mientras que para las probabilidades de acierto fueron de cero y uno respectivamente. Se obtuvieron los resultados para un nivel de significancia $\alpha = 0.05$, usando tamaños muestrales de $n = 5;10;15;20$ para los casos de estimación de una sola media y $n = (5,5);(10,10);(5,10);(10,5)$ cuando la estimación es sobre dos medias. Cabe recalcar que los estadísticos correspondientes en cada modelo de inferencia utilizado fueron calculados de acuerdo a cada premisa diferente sobre el conocimiento de la varianza teórica y sobre los criterios correspondientes de homocedasticidad o heterocedasticidad.

3 Resultados

En el Anexo D, se muestran los resultados completos de las estimaciones de las probabilidades de error y aciertos para todos los casos descritos. Además, se muestra en paralelo la precisión que tuvo el proceso de simulación al estimar estos parámetros, esto es el error estándar de la estimación de probabilidades a partir de proporciones. Para visualizarlo de mejor manera, estos resultados se resumen en los gráficos que se detallan a continuación.

La Figura 1 muestra que la probabilidad de error es exactamente la misma cuando se usan cualquiera de los dos métodos (pruebas de hipótesis o intervalos de confianza), independientemente del tamaño muestral y el estadístico utilizado (Z, t o Welch). La probabilidad de acierto también es la misma con cualquiera de los dos métodos, pero

incrementa conforme el tamaño muestral es mayor, en este sentido aunque estudiar casos para tamaños muestrales más grandes puede resultar interesante, no tendría sentido hacerlo, puesto que ya se observa claramente lo que ocurriría, esto es, que las probabilidades de cometer un error se mantendrían relativamente controladas muy cerca del nivel de significancia planteado y las probabilidades de aciertos irían aumentando de manera asintótica hacia 1 (probabilidad límite).

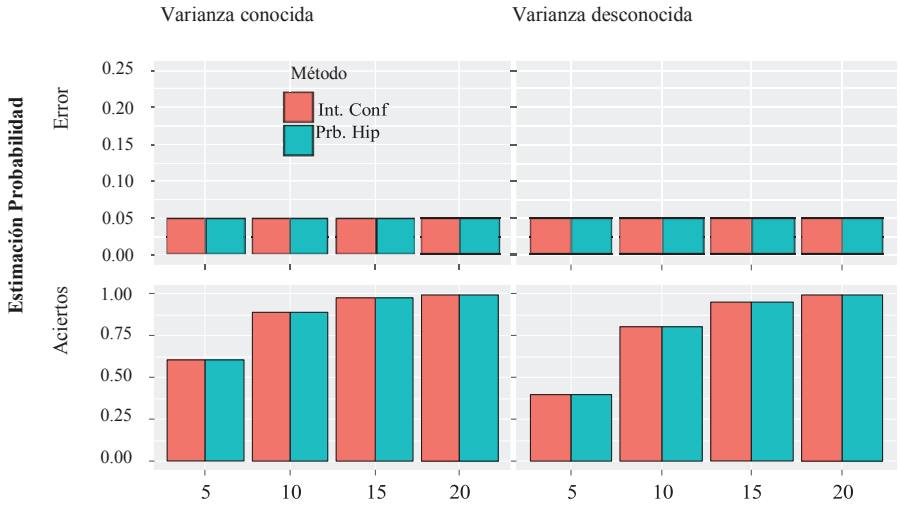

Figura 1. *Estimación de probabilidad de acierto y error para inferencias sobre una media usando $\alpha=0.05$.*

La Figura 2 y la Figura 3 muestran respectivamente las probabilidades de cometer un error y un acierto cuando se estiman dos medias con la premisa de varianzas conocidas. Podemos notar que el comportamiento es similar al presentado en el caso de una media en cuanto que no se observan diferencias entre las probabilidades estimadas cuando se usan intervalos de confianza y pruebas de hipótesis. El error es siempre el mismo independientemente del tamaño muestral y el nivel de heterocedasticidad, pero la probabilidad de acertar es mayor conforme el tamaño muestral incrementa (especialmente para casos balanceados) y el nivel de heterocedasticidad es menor.

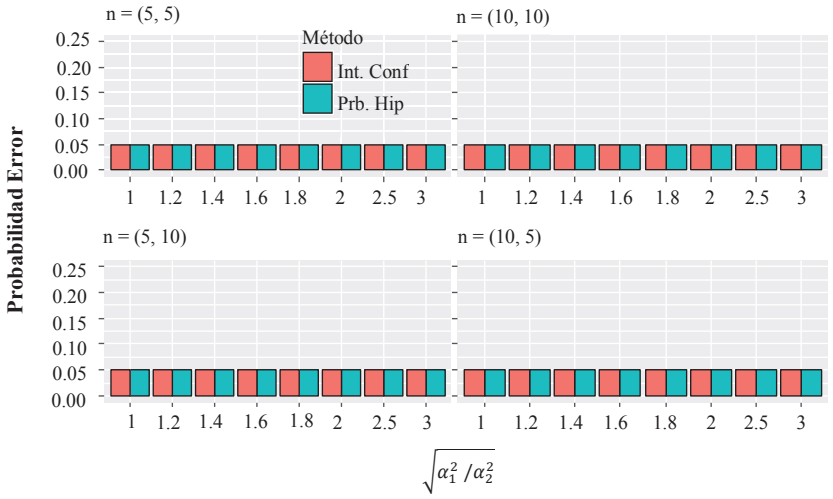

Figura 2. *Estimación de probabilidad de error para inferencias sobre dos medias cuando la varianza es conocida, usando α=0.05*

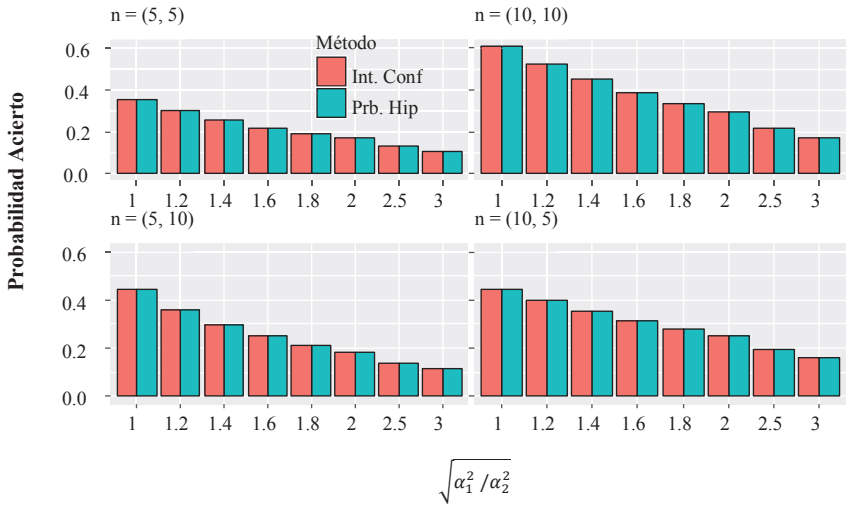

Figura 3. *Estimación de probabilidad de acierto para inferencias sobre dos medias cuando la varianza es conocida, usando α=0.05.*

Conforme lo muestran la Figura 4 y la Figura 5, exactamente este mismo comportamiento se observa cuando la varianza es desconocida, aunque con una leve disminución de la probabilidad de cometer un acierto, pero con la misma dinámica que con varianzas conocidas.

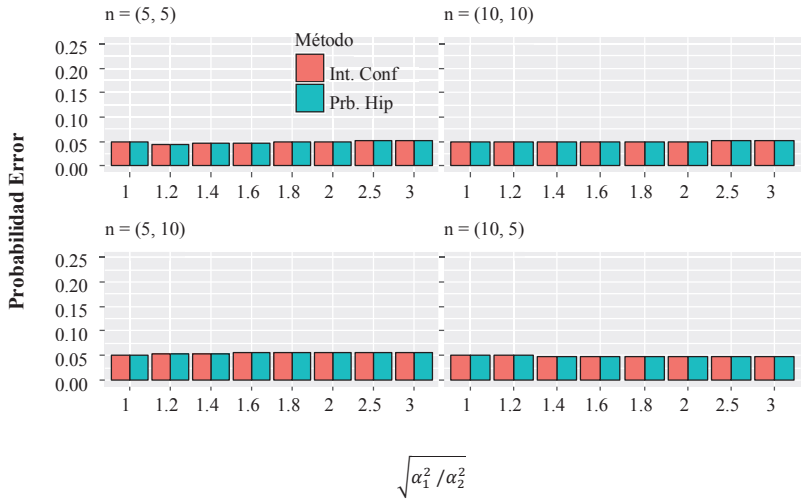

Figura 4. *Estimación de probabilidad de error para inferencias sobre dos medias cuando la varianza es desconocida, usando α = 0.05.*

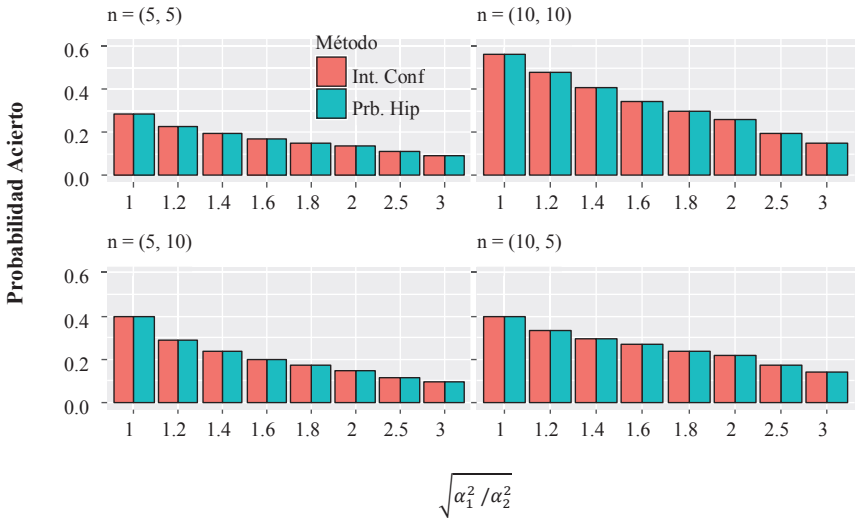

Figura 5. *Estimación de probabilidad de acierto para inferencias sobre dos medias cuando la varianza es desconocida, usando α = 0.05*

Del mismo modo que para una sola media, es un tanto predecible el comportamiento que tendrían las probabilidades estimadas cuando se usan tamaños muestrales más grandes, esto es, que en todos los casos balanceados y desbalanceados la probabilidad de error tanto para intervalos de confianza como para pruebas de hipótesis quedaría bastante controlada y cerca del nivel de significancia planteado, mientras que la probabilidad de acierto iría aumentando conforme el tamaño muestral crezca, aunque con una mayor rapidez para los casos balanceados que para los desbalanceados.

4 Conclusiones

Los resultados observados muestran claramente que sin importar el número de parámetros a compararse (uno o dos), el estadístico (Z, t o Welch), el tamaño muestral o el nivel de heterocedasticidad, no existe evidencia para concluir estadísticamente que la metodología basada en intervalos de confianza tenga una mayor probabilidad de aciertos o una menor probabilidad de cometer un error que la presentada por la metodología basada en pruebas de hipótesis. En este sentido, no podemos emitir un criterio estadístico formal que confirme que los intervalos de confianza (estadísticamente hablando) son procedimientos que superan la eficiencia de las pruebas de hipótesis en los procesos de inferencia estadística.

Si bien es cierto, como ya lo dijimos, confirmamos que los intervalos de confianza proporcionan mayor información sobre la magnitud de la diferencia estadística hallada entre tratamientos, sin embargo, los resultados de la medición de la probabilidad de acierto o error que tienen estos procedimientos muestran que ambos tienen la misma efectividad de inferencia. Concluimos que los modelos de pruebas de hipótesis no son erróneos ni menos potentes que los intervalos de confianza, lo único que ocurre es que estos tienen un incorrecto planteamiento que no permite observar la magnitud de las diferencias significativas halladas, sin embargo, creemos que un planteamiento de hipótesis basado en equivalencia proporcionará esta información requerida, la cual se puede establecer en los límites de irrelevancia que el investigador proporcione al test de hipótesis.

Anexos:

Anexos A: Funciones Auxiliares

A.1: Intervalo de Confianza para una media cuando varianza conocida

```
icomscon <- function(x, sigma, n = length(x), alpha){
  z <- qnorm(1 - (alpha/2))
  c(mean(x) - z*(sigma/sqrt(n)), mean(x) + z*(sigma/sqrt(n)))
}
```

A.2: Intervalo de Confianza para una media cuando varianza desconocida

```
icomsds <- function(x, n = length(x), alpha){
  t <- qt(1-(alpha/2), n - 1)
  c(mean(x) - t*(sd(x)/sqrt(n)), mean(x) + t*(sd(x)/sqrt(n)))
}
```

A.3: Intervalo de Confianza para dos medias cuando las varianzas son conocidas

```
ic2mscon <- function(x, y, n1 = length(x), n2 =  length(y), alpha,
                     sigma1, sigma2){
  z <- qnorm(1 - (alpha/2))
  c((mean(x) - mean(y)) - z * sqrt((sigma1^2/n1) + (sigma2^2/n2)),
    (mean(x) - mean(y)) + z * sqrt((sigma1^2/n1) + (sigma2^2/n2)))
}
```

A.4: Intervalo de Confianza para dos medias cuando las varianzas son desconocidas e iguales

```
ic2tstud <- function(x, y, n1 = length(x), n2 =  length(y), alpha){
```

```
  t <- qt(1-(alpha/2), n1 + n2 - 2)
  sp2 <- (var(x) * (n1 - 1) + var(y) * (n2 - 1)) / (n1 + n2 - 2)
  c((mean(x) - mean(y)) - (t*sqrt(sp2)*sqrt((1/n1) + (1/n2))),
    (mean(x) - mean(y)) + (t*sqrt(sp2)*sqrt((1/n1) + (1/n2))))
}
```

A.5: Intervalo de Confianza para dos medias cuando las varianzas son desconocidas y diferentes

```
ic2welch <- function(x, y, n1 = length(x), n2 =  length(y), alpha){
  v <- (((var(x)/n1)+(var(y)/n2))^2)/((((var(x)/n1)^2)/(n1-
1))+(((var(y)/n2)^2)/(n2-1)))
  t <- qt(1-(alpha/2), v)
  c((mean(x) - mean(y)) - t*sqrt((var(x)/n1) + (var(y)/n2)),
    (mean(x) - mean(y)) + t*sqrt((var(x)/n1) + (var(y)/n2)))
}
```

A.6: Estadístico de Prueba para un Z test de una muestra

```
zval1m <- function(x, n = length(x), mu.hip, sigma){
  abs((mean(x) - mu.hip)/(sigma/sqrt(n)))
}
```

A.7: Estadístico de Prueba para un t - Student test de una muestra

```
tval1m <- function(x, n = length(x), mu.hip){
  abs((mean(x) - mu.hip)/(sd(x)/sqrt(n)))
}
```

A.8: Estadístico de Prueba para un Z test de dos muestras

```
z.val2m <- function(x, y, n1 = length(x), n2 = length(y), mu.hip = 0,
                    sigma1, sigma2){
  abs(((mean(x) - mean(y)) - mu.hip)/sqrt((sigma1^2/n1) + (sigma2^2/
n2)))
}
```

A.9: Estadístico de Prueba para un t - Student test de dos muestras

```
t.val2m <- function(x, y, n1 = length(x), n2 = length(y), mu.hip = 0){
  sp2 <- (var(x)*(n1 - 1) + var(y)*(n2 - 1))/(n1 + n2 - 2)
  abs(((mean(x) - mean(y)) - mu.hip)/(sqrt(sp2) * sqrt((1/n1) + (1/
n2))))
}
```

A.10: Estadístico de Prueba para un Welch test de dos muestras

```
wel.val2m <- function(x, y, n1 = length(x), n2 = length(y), mu.hip = 0)
{
  abs(((mean(x) - mean(y)) - mu.hip)/sqrt((var(x)/n1) + (var(y)/n2)))
}
```

A.11: Función devuelve TRUE si el valor esta fuera del intervalo int y FALSE en caso contrario

```
val.out <- function(q, int){
  (q < min(int) | q > max(int))
}
```

Anexo B: Función principal para estimar las probabilidades de error y acierto en el caso de una muestra

```
onemean <- function(m = 1000000, n, mu = 0, mu.hip = 0, alpha = 0.05,
                    sigma, var.known = F, seed = 1234){
  old.seed <- .Random.seed
  set.seed(seed)
  vcz <- qnorm(1 - alpha/2)
  vct <- qt(1 - alpha/2, n - 1)
  z <- qnorm(1- alpha/2)
  if(var.known){
   sim <- replicate(m,
   {
    xcontrol <- rnorm(n)
    x <- mu + (xcontrol * sigma)
    ic <- icomscon(x, sigma, n, alpha)
    est <- zval1m(x, n, mu.hip, sigma)
    c("Err_IC" = val.out(mu.hip, ic), "Err_PH" = est > vcz)
   })
  r <- rowSums(sim)/m
  attr(r, "precis") <- z*sqrt((r*(1-r))/m)
 } else{
  sim <- replicate(m,
   {
    xcontrol <- rnorm(n)
    x <- mu + (xcontrol * sigma)
    ic <- icomsds(x, n, alpha)
    est <- tval1m(x, n, mu.hip)
    c("Err_IC" = val.out(mu.hip, ic), "Err_PH" = est > vct)
     })
  r <- rowSums(sim)/m
  attr(r, "precis") <- z*sqrt((r*(1-r))/m)
 }
  print(r)
}
```

Anexo C: Función principal para estimar las probabilidades de error y acierto en el caso de dos muestras

```
twomean <- function(m = 1000000, n1, n2, diffmean = 0, dm.hip = 0,
                    alpha = 0.05, ratioSigma = 1, var.known = F,
                    seed = 1234){
  old.seed <- .Random.seed
  set.seed(seed)
  vcz <- qnorm(1 - alpha/2)
  vct <- qt(1 - alpha/2, n1 + n2 - 2)

  if(var.known){
    sim <- replicate(m,
              {
                x1control <- rnorm(n1)
                x2control <- rnorm(n2)
```

```
                x <- x1control * ratioSigma
                y <- x2control + diffmean
                ic <- ic2mscon(x, y, n1, n2, alpha, ratioSigma, 1)
                est <- z.val2m(x, y, n1, n2, dm.hip, ratioSigma, 1)
                c("ZErr_IC" = val.out(dm.hip, ic), "Err_PH" = est >
vcz)
              })
    r <- rowSums(sim)/m
    attr(r, "precis") <- z*sqrt((r*(1-r))/m)
  }
  if(!var.known & ratioSigma == 1){
    sim <- replicate(m,
              {
                x1control <- rnorm(n1)
                x2control <- rnorm(n2)
                x <- x1control * ratioSigma
                y <- x2control + diffmean
                ic <- ic2tstud(x, y, n1, n2, alpha)
                est <- t.val2m(x, y, n1, n2, dm.hip)
                c("tErr_IC" = val.out(dm.hip, ic), "Err_PH" = est >
vct)
              })
    r <- rowSums(sim)/m
    attr(r, "precis") <- z*sqrt((r*(1-r))/m)
  }
  if(!var.known & ratioSigma > 1){
    sim <- replicate(m,
              {
                x1control <- rnorm(n1)
                x2control <- rnorm(n2)
                x <- x1control * ratioSigma
                y <- x2control + diffmean
                ic <- ic2welch(x, y, n1, n2, alpha)
                est <- wel.val2m(x, y, n1, n2, dm.hip)
                v <- (((var(x)/n1) + (var(y)/n2))^2) / ((((var(x)/
n1)^2)/                          (n1-1)) + (((var(y) / n2)^2) / (n2-1)))
                vcw <- qt(1 - alpha/2, v)
                c("wErr_IC" = val.out(dm.hip, ic), "Err_PH" = est >
vcw)
              })
    r <- rowSums(sim)/m
    attr(r, "precis") <- z*sqrt((r*(1-r))/m)
  }
  print (r)
}
```

Anexo D: Resultados de las estimaciones mediante el proceso de simulación

D.1: Para una muestra

n	Intervalos de Confianza				Pruebas de Hipótesis			
	Varianza Conocida		Varianza Desconocida		Varianza Conocida		Varianza Desconocida	
	Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)
5	0.050142 (0.0004277384)	0.609088 (0.0009563737)	0.050071 (0.0004274515)	0.402103 (0.0009610145)	0.050142 (0.0004277384)	0.609088 (0.0009563737)	0.050071 (0.0004274515)	0.402103 (0.0009610145)
10	0.050326 (0.000428481)	0.885599 (0.000623852)	0.049718 (0.0004260212)	0.803106 (0.0007793827)	0.050326 (0.000428481)	0.885599 (0.000623852)	0.049718 (0.0004260212)	0.803106 (0.0007793827)
15	0.050415 (0.050415)	0.971926 (0.0003237553)	0.05018 (0.0004278919)	0.94887 (0.0004317073)	0.050415 (0.050415)	0.971926 (0.0003237553)	0.05018 (0.0004278919)	0.94887 (0.0004317073)
20	0.050214 (0.0004280292)	0.993986 (0.0001515374)	0.050089 (0.0004275242)	0.988596 (0.0002081067)	0.050214 (0.0004280292)	0.993986 (0.0001515374)	0.050089 (0.0004275242)	0.988596 (0.0002081067)

e.e* denota el error estándar de la estimación, una forma de medir la precisión del proceso de simulación para cada estimador puntual

D.2: Para dos muestras

n	σ_1^2/σ_2^2	Intervalos de Confianza				Pruebas de Hipótesis			
		Varianza Conocida		Varianza Desconocida		Varianza Conocida		Varianza Desconocida	
		Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)
(5, 5)	1.0	0.050058 (0.0004273989)	0.353644 (0.0009370594)	0.050398 (0.0004287711)	0.287239 (0.0008868328)	0.050058 (0.0004273989)	0.353644 (0.0009370594)	0.050398 (0.0004287711)	0.287239 (0.0008868328)
	1.2	0.049941 (0.0004269254)	0.300207 (0.0008983453)	0.044908 (0.0004059126)	0.225538 (0.0008191398)	0.049941 (0.0004269254)	0.300207 (0.0008983453)	0.044908 (0.0004059126)	0.225538 (0.0008191398)
	1.4	0.050015 (0.0004272249)	0.256249 (0.0008556437)	0.045928 (0.0004102772)	0.193888 (0.0007748564)	0.050015 (0.0004272249)	0.256249 (0.0008556437)	0.045928 (0.0004102772)	0.193888 (0.0007748564)
	1.6	0.049834 (0.0004264918)	0.220905 (0.000813104)	0.047041 (0.0004149765)	0.169035 (0.0007345608)	0.049834 (0.0004264918)	0.220905 (0.000813104)	0.047041 (0.0004149765)	0.169035 (0.0007345608)
	1.8	0.049736 (0.0004260942)	0.193189 (0.0007737937)	0.048241 (0.0004199714)	0.150018 (0.0006998817)	0.049736 (0.0004260942)	0.193189 (0.0007737937)	0.048241 (0.0004199714)	0.150018 (0.0006998817)
	2.0	0.049606 (0.0004255661)	0.170989 (0.0007379252)	0.049323 (0.0004244136)	0.134763 (0.0006692694)	0.049606 (0.0004255661)	0.170989 (0.0007379252)	0.049323 (0.0004244136)	0.134763 (0.0006692694)
	2.5	0.049593 (0.0004255133)	0.132764 (0.000665054)	0.051537 (0.0004333291)	0.109071 (0.0006109762)	0.049593 (0.0004255133)	0.132764 (0.000665054)	0.051537 (0.0004333291)	0.109071 (0.0006109762)
	3.0	0.049723 (0.0004260415)	0.10974 (0.0006126169)	0.052603 (0.0004375416)	0.093574 (0.0005708104)	0.049723 (0.0004260415)	0.10974 (0.0006126169)	0.052603 (0.0004375416)	0.093574 (0.0005708104)

n	σ_1^2/σ_2^2	Intervalos de Confianza				Pruebas de Hipótesis			
		Varianza Conocida		Varianza Desconocida		Varianza Conocida		Varianza Desconocida	
		Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)
(10, 10)	1.0	0.050213 (0.0004280251)	0.608113 (0.0009567989)	0.050217 (0.0004280413)	0.561533 (0.0009725326)	0.050213 (0.0004280251)	0.608113 (0.0009567989)	0.050217 (0.0004280413)	0.561533 (0.0009725326)
	1.2	0.05029 (0.0004283358)	0.52535 (0.0009787217)	0.048868 (0.0004225526)	0.476056 (0.0009788577)	0.05029 (0.0004283358)	0.52535 (0.0009787217)	0.048868 (0.0004225526)	0.476056 (0.0009788577)
	1.4	0.050268 (0.0004282471)	0.451233 (0.0009753096)	0.049074 (0.0004233964)	0.405149 (0.0009621872)	0.050268 (0.0004282471)	0.451233 (0.0009753096)	0.049074 (0.0004233964)	0.405149 (0.0009621872)
	1.6	0.0501 (0.0004275687)	0.387838 (0.0009550068)	0.049497 (0.0004251227)	0.345741 (0.000932177)	0.0501 (0.0004275687)	0.387838 (0.0009550068)	0.049497 (0.0004251227)	0.345741 (0.000932177)
	1.8	0.049975 (0.0004270631)	0.335649 (0.000925528)	0.049789 (0.0004263093)	0.297628 (0.000896125)	0.049975 (0.0004270631)	0.335649 (0.000925528)	0.049789 (0.0004263093)	0.297628 (0.000896125)
	2.0	0.049992 (0.0004271319)	0.293128 (0.000892169)	0.05006 (0.000427407)	0.258813 (0.0008584303)	0.049992 (0.0004271319)	0.293128 (0.000892169)	0.05006 (0.000427407)	0.258813 (0.0008584303)
	2.5	0.049967 (0.0004270307)	0.21706 (0.000807983)	0.05058 (0.0004295035)	0.192335 (0.00077249)	0.049967 (0.0004270307)	0.21706 (0.000807983)	0.05058 (0.0004295035)	0.192335 (0.00077249)
	3.0	0.050091 (0.0004275323)	0.170503 (0.0007370917)	0.050892 (0.0004307553)	0.152046 (0.0007037554)	0.050091 (0.0004275323)	0.170503 (0.0007370917)	0.050892 (0.0004307553)	0.152046 (0.0007037554)

n	σ_1^2/σ_2^2	Intervalos de Confianza				Pruebas de Hipótesis			
		Varianza Conocida		Varianza Desconocida		Varianza Conocida		Varianza Desconocida	
		Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)
(5, 10)	1.0	0.050244 (0.0004281503)	0.446505 (0.000974357)	0.05024 (0.0004281341)	0.394432 (0.00095789)	0.050244 (0.0004281503)	0.446505 (0.000974357)	0.05024 (0.0004281341)	0.394432 (0.00095789)
	1.2	0.050182 (0.0004279)	0.361913 (0.0009418682)	0.052644 (0.0004377026)	0.288937 (0.0008883901)	0.050182 (0.0004279)	0.361913 (0.0009418682)	0.052644 (0.0004377026)	0.288937 (0.0008883901)
	1.4	0.050117 (0.0004276374)	0.296689 (0.0008953081)	0.053935 (0.0004427351)	0.236315 (0.0008326277)	0.050117 (0.0004276374)	0.296689 (0.0008953081)	0.053935 (0.0004427351)	0.236315 (0.0008326277)
	1.6	0.049962 (0.0004270104)	0.248722 (0.0008472382)	0.054834 (0.0004461975)	0.198085 (0.0007811565)	0.049962 (0.0004270104)	0.248722 (0.0008472382)	0.054834 (0.0004461975)	0.198085 (0.0007811565)
	1.8	0.049922 (0.0004268485)	0.212035 (0.0008011343)	0.055416 (0.000448421)	0.17016 (0.0007365021)	0.049922 (0.0004268485)	0.212035 (0.0008011343)	0.055416 (0.000448421)	0.17016 (0.0007365021)
	2.0	0.049913 (0.000426812)	0.184296 (0.0007599278)	0.055744 (0.0004496681)	0.149245 (0.0006983936)	0.049913 (0.000426812)	0.184296 (0.0007599278)	0.055744 (0.0004496681)	0.149245 (0.0006983936)
	2.5	0.050037 (0.0004273139)	0.138781 (0.0006775945)	0.055797 (0.0004498691)	0.115867 (0.0006273166)	0.050037 (0.0004273139)	0.138781 (0.0006775945)	0.055797 (0.0004498691)	0.115867 (0.0006273166)
	3.0	0.050075 (0.0004274676)	0.11286 (0.0006201749)	0.055391 (0.0004483258)	0.096655 (0.0005791447)	0.050075 (0.0004274676)	0.11286 (0.0006201749)	0.055391 (0.0004483258)	0.096655 (0.0005791447)

n	σ_1^2/σ_2^2	Intervalos de Confianza				Pruebas de Hipótesis			
		Varianza Conocida		Varianza Desconocida		Varianza Conocida		Varianza Desconocida	
		Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)	Prob. Error (e.e*)	Prob. Acierto (e.e*)
(10, 5)	1.0	0.050009 (0.0004272007)	0.446416 (0.0009743382)	0.049874 (0.000426654)	0.394061 (0.0009577326)	0.050009 (0.0004272007)	0.446416 (0.0009743382)	0.049874 (0.000426654)	0.394061 (0.0009577326)
	1.2	0.049975 (0.0004270631)	0.399289 (0.0009598968)	0.049096 (0.0004234864)	0.330334 (0.0009218364)	0.049975 (0.0004270631)	0.399289 (0.0009598968)	0.049096 (0.0004234864)	0.330334 (0.0009218364)
	1.4	0.049864 (0.0004266134)	0.355706 (0.0009382871)	0.048061 (0.0004192268)	0.297472 (0.0008959896)	0.049864 (0.0004266134)	0.355706 (0.0009382871)	0.048061 (0.0004192268)	0.297472 (0.0008959896)
	1.6	0.049873 (0.0004266499)	0.316487 (0.0009115897)	0.047517 (0.0004169665)	0.266672 (0.266672)	0.049873 (0.0004266499)	0.316487 (0.0009115897)	0.047517 (0.0004169665)	0.266672 (0.266672)
	1.8	0.049892 (0.0004267269)	0.281733 (0.0008816778)	0.047331 (0.0004161903)	0.239289 (0.0008362176)	0.049892 (0.0004267269)	0.281733 (0.0008816778)	0.047331 (0.0004161903)	0.239289 (0.0008362176)
	2.0	0.050092 (0.0004275364)	0.252047 (0.000850993)	0.0474 (0.0004164785)	0.215623 (0.0008060427)	0.050092 (0.0004275364)	0.252047 (0.000850993)	0.0474 (0.0004164785)	0.215623 (0.0008060427)
	2.5	0.050197 (0.0004279606)	0.196269 (0.0007784475)	0.047989 (0.0004189285)	0.170023 (0.0007362664)	0.050197 (0.0004279606)	0.196269 (0.0007784475)	0.047989 (0.0004189285)	0.170023 (0.0007362664)
	3.0	0.050142 (0.0004277384)	0.158685 (0.0007161357)	0.048633 (0.0004215874)	0.139582 (0.0006792311)	0.050142 (0.0004277384)	0.158685 (0.0007161357)	0.048633 (0.0004215874)	0.139582 (0.0006792311)

e.e* denota el error estándar de la estimación, una forma de medir la precisión del proceso de simulación para cada estimador puntual

Referencias bibliográficas

1. Berkson J. Some difficulties of interpretation encountered in the application of the chi-square test. J Am Stat Assoc. 1938;33(203):526–536. DOI: 10.1080/01621459.1938.10502329.

2. Berkson J. Tests of significance considered as evidence. Int J Epidemiol. 2003;32(5):687–691.

3. Clark ML. Los valores P y los intervalos de confianza: ¿En qué confiar? Editorial. Revista Panamericana de Salud Pública. 2004;15(5):293.

4. Cohen J. The earth is round (p <. 05). In: What if there were no significance tests? Routledge; 2016. p. 69–82.

5. Carver R. The case against statistical significance testing. Harv Educ Rev. 1978;48(3):378–399.

6. Sarria Castro M, Silva Aycaguer LC. Las pruebas de significación estadística en tres revistas biomédicas: una revisión crítica. Rev Panam Salud Pública. 2004;15:300--306.

7. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed). 1986;292(6522):746-750.

8. Escalante Angulo C. Prueba de hipótesis frente a intervalos de confianza. Cienc & Tecnol para la Salud Vis y Ocul. 2010;8(2):145–9.

9. Feinstein AR. Clinical Epidemiology: The Architecture of Clinical Research. Philadelphia, PA. Ed. WB Saunders Company: 2nd Edition. 1985; 812.

10. Haynes RB, Mulrow CD, Huth EJ, Altman DG, Gardner MJ. More Informative Abstracts Revisited. Cleft Palate-Craniofacial J. 1996;33(1–9).

11. Box GE. Robustness in the strategy of scientific model building. Army Res Off Work Robustness Stat. 1979;1:201–36.

12. Cochran WG. The X2 correction for continuity. Iowa State Coll J Sci. 1942;16:421–36.

13. Rasch D, Guiard V. The robustness of parametric statistical methods. Psychol Sci. 2004;46(2):175–208.

14. Altman, Douglas G, Bland JM. Statistics Notes: Absence of evidence is not evidence of absence. Bmj. 1995;311(7003):485.

15. Wellek S. Testing Statistical Hypotheses of Equivalence and Noninferiority. 2nd Edition, Chapman and Hall/CRC. 2010. 3 p.

16. Flores P, Ocana J. Heteroscedasticity irrelevance when testing means difference. SORT. 2018;42(1):59–72. DOI: 10.2436/20.8080.02.6910

17. Flores P. Un pre test de irrelevancia de la diferencia de varianzas en la comparación de medias. 2017.

18. Walpole R, Myers R, Myers S, Keying Y. Probabilidad y Estadística para ingeniería y Ciencias. 9th ed. México: Pearson Education; 2012.

19. Team RC. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2016. Available from: https://www.r-project.org/

Dirección del autor

Pablo Javier Flores Muñoz
Grupo de Investigación en Ciencias de Datos CIDED, Facultad de Ciencias. Escuela Superior Politécnica de Chimborazo, Riobamba, Ecuador
p_flores@espoch.edu.ec