

LA PRUEBA T DE STUDENT EN INVESTIGACIÓN: GUÍA DE USO CON UNA MIRADA PRÁCTICA.

AUTOR: GUILLERMO RAMÓN FRANCO DEL RÍO

DIVISION DE CIENCIAS DE LA SALUD, UNIVERSIDAD ANAHUAC QUERETARO.

RESUMEN

La prueba t de Student es una herramienta sencilla y poderosa para comparar medias en variables continuas. Bien empleada, ayuda a traducir datos en decisiones clínicas; mal aplicada, genera resultados frágiles y conclusiones engañosas. Este artículo, repasa su origen, cuándo usarla, qué condiciones deben cumplirse, cómo reportar resultados con sentido clínico y cuáles son los errores de uso más habituales. El mensaje central es claro: la t de Student es valiosa cuando se aplica correctamente, pero su uso indiscriminado y el desconocimiento de sus supuestos pueden llevar a hallazgos poco fiables y de escasa utilidad.

Palabras clave: prueba t de Student, Welch, intervalos de confianza, tamaño de efecto, bioestadística clínica, interpretación del valor p.

SUMMARY

The Student t-test is a simple yet powerful tool for comparing means in continuous variables. When properly used, it helps translate data into clinical decisions; when misapplied, it produces fragile results and misleading conclusions. This article reviews its origin, when to use it, the conditions that must be met, how to report results with clinical relevance, and the most common misuse errors. The central message is clear: the Student t-test is valuable when correctly applied, but indiscriminate use and ignorance of its assumptions can lead to unreliable and clinically unhelpful findings.

Keywords: Student's t-test, Welch, confidence intervals, effect size, clinical biostatistics, p-value interpretation.

Una historia que comienza en una cervecería.

A comienzos del siglo XX, William S. Gosset trabajaba controlando la calidad de la cerveza en la fábrica de Guinness en Irlanda. Necesitaba comparar promedios con pocos datos y sin conocer la variabilidad "verdadera" de toda la producción de cerveza. Para resolverlo, desarrolló la distribución y la prueba que hoy conocemos como t de Student (Student fue su seudónimo, Guinness prohibía a sus empleados divulgar sus secretos). Su aporte permitió hacer inferencias sobre medias en situaciones realistas: muestras pequeñas y varianza desconocida. Ese espíritu práctico es el que mantiene a la t vigente en la investigación. (Murtagh & Bryner, 2025)

¿Qué pregunta responde la t?

En términos sencillos, la t compara la diferencia observada en promedios tomando en cuenta la variabilidad de los datos. Pregunta: "¿Esta diferencia podría deberse al azar, dado lo variable que son las mediciones, o es suficientemente grande para considerarla evidencia de un efecto real?" La respuesta se expresa con un valor p y, más importante, con un intervalo de confianza, que describe un rango plausible para el tamaño del efecto. (Chicco et al., 2025)

Tres variantes, un mismo objetivo

La t adopta formatos según el diseño del estudio: (White, 2021)

- **Dos grupos independientes.** Se usa para comparar medias entre grupos distintos (por ejemplo, tratamiento vs. control).

Cuando la variabilidad de ambos grupos no es similar o los tamaños son desbalanceados sin normalidad, la versión *t* de Welch es preferible porque no exige igualdad de varianzas y mantiene mejor el control del error.

- **Pareada (antes–después o mediciones apareadas).** Compara a cada paciente consigo mismo (pre y post intervención) o pares naturales (por ejemplo, dos mediciones en la misma persona). Al trabajar con diferencias dentro de sujeto, reduce el “ruido” interindividual.

- **Una muestra frente a un valor de referencia.** Útil cuando se quiere contrastar la media de una muestra con un estándar clínicos (por ejemplo, el peso promedio de una muestra de recién nacidos contra el peso promedio de la población general).

Requisitos que conviene verificar (Daniel & Chad, 2019)

La *t* funciona bien bajo condiciones razonables y, en la práctica, es robusta. Aun así, conviene revisar:

1. **Tipo de variable.** Debe ser continua y tener sentido resumirla con una media (mmHg, mg/dL, minutos, puntuaciones continuas). Para variables binarias (sí/no) o conteos, hay otros métodos más adecuados.

2. **Independencia.** Las observaciones deben ser independientes dentro de cada grupo. No es lo mismo 100 pacientes que 50 pacientes medidos dos veces cada uno. Tampoco se deben analizar “como si fueran independientes” datos que están agrupados (por hospital, por familia) o ambos ojos de la misma persona.

En esos casos, puede usarse la versión pareada o modelos que tengan en cuenta el agrupamiento.

3. **Forma de la distribución.** La *t* no exige perfección “gaussiana” de los datos crudos. Lo importante es que el promedio no esté dominado por asimetrías extremas o por unos pocos valores atípicos. Hay que revisar los histogramas, gráficos de caja y gráficos Q-Q; si observa asimetría marcada u datos muy extremos influyentes, hay que transformar la variable con logaritmos o alternativas no paramétricas.

4. **Homogeneidad de varianzas (solo para la versión clásica).** Si sospecha varianzas distintas entre grupos, la *t* de Welch suele ser la elección prudente.

Cómo reportar para que el resultado sirva clínicamente

Un buen reporte no se limita a “*p* menor de 0,05”. Debe incluir:

- **La estimación del efecto:** la diferencia de medias entre grupos o el cambio medio en el grupo pareado.

- **El intervalo de confianza al 95% (IC95%):** ofrece el rango de valores compatibles con los datos. Comunica precisión y ayuda a juzgar relevancia clínica.

- **El valor *p*:** evidencia estadística contra la hipótesis de “no diferencia”, pero no indica tamaño ni importancia clínica.

- **Un tamaño de efecto estandarizado:** por ejemplo, una diferencia expresada respecto a la variabilidad facilita comparar resultados entre estudios.

- **Contexto clínico:** contraste explícito con la diferencia mínimamente clínicamente importante (MCID).

importante (MCID). Un resultado estadísticamente “positivo” puede ser irrelevante si la magnitud es trivial; y un resultado “no significativo” puede ser compatible con beneficios relevantes si el IC es amplio por falta de precisión.

¿Cómo puedes reportarlo?:

“En comparación con el grupo control, el tratamiento redujo en promedio la presión arterial sistólica. La diferencia estimada fue de X unidades, con un IC del 95% de Y a Z. El valor p fue W. El tamaño de efecto sugiere [pequeño/moderado/grande]. La magnitud observada [sí/no] supera la MCID predefinida.”

Usos típicos en clínica (Pandey, 2015)

- **Ensayos clínicos:** comparación de medias de desenlaces continuos (glucosa sérica, presión arterial, escala de dolor). La prueba ha sido encontrada junto con ANOVA hasta en el 84.5% de las publicaciones científicas, sin embargo, debe considerarse que el porcentaje exacto depende del área, revista y criterios del análisis, pero los valores coinciden en que la *t* de Student figura entre los métodos estadísticos más comunes. (Weissgerber et al., 2018).

- **Estudios pre-pos:** evaluar el cambio medio tras una intervención, por ejemplo, ejercicio físico.

- **Observacionales:** comparar medias entre expuestos vs. no expuestos, idealmente con análisis ajustados cuando hay covariables relevantes como edad o valor basal.

Doce errores frecuentes (y cómo evitarlos)

- 1. Usarla con variables inadecuadas.** Proporciones y conteos requieren otras familias de modelos.

- 2. Ignorar la dependencia.** Datos por paciente y por centro, ojos en oftalmología, mediciones repetidas... todo eso introduce correlación. Solución: versión pareada, modelos mixtos o métodos para datos correlacionados.

- 3. Insistir en la versión clásica con varianzas desiguales.** Si hay duda razonable, usa *t* de Welch.

- 4. No tratar outliers ni asimetrías.** Explora, justifica transformaciones, haz análisis de sensibilidad y considera métodos robustos o no paramétricos.

- 5. Convertir una prueba bilateral en unilateral después de ver los datos.** Es mala práctica y aumenta falsos positivos. La dirección debe preespecificarse.

- 6. Confundir “no significativo” con “equivalente” o “no inferior”.** La ausencia de significación no demuestra equivalencia. Para ello existen procedimientos específicos con márgenes clínicos definidos de antemano.

- 7. Pruebas de normalidad mal interpretadas.** En muestras pequeñas carecen de poder; en grandes detectan trivialidades. Da más peso a los gráficos y al impacto clínico.

- 8. Multiplicidad sin control.** Comparar muchos desenlaces, tiempos o subgrupos sin ajustar eleva el riesgo de falsos positivos. Determina objetivos y usa ajustes cuando corresponda.

- 9. Probar “diferencias basales” en ensayos aleatorizados.** No aporta; la aleatorización equilibra en expectativa. Describe basales y, si hace falta, ajusta en el análisis principal.

10. Analizar “cambios” sin pensar en el basal.

Con frecuencia es mejor comparar el valor final ajustado por el basal que comparar deltas (diferencias)

11. Olvidar la unidad de análisis. Defínala con claridad (paciente, centro, ojo). No mezcle niveles.

12. Reducir el informe al valor p. Sin IC, tamaño de efecto y comparación con la MCID, el resultado estadístico carece de traducción clínica.

Recuadro práctico: verificación de 9 pasos

1. Variable continua y la media es un resumen clínicamente pertinente.

2. Diseño correcto: independiente, pareado o una muestra frente a referencia.

3. Unidad de análisis clara; evite mezclar niveles (paciente/centro/órgano).

4. Exploración gráfica: asimetría, valores atípicos, dispersión por grupo.

5. Versión adecuada: use Welch si hay dudas sobre varianzas iguales o tamaños desbalanceados.

6. Reporte completo: estimación, IC, p, tamaño de efecto, y relación con la MCID.

7. Multiplicidad controlada y objetivos primaria/secundarios preespecificados.

8. Sensibilidad: resultados con y sin transformaciones o exclusión razonada de outliers.

9. Interpretación clínica, no sólo estadística.

¿Como lo puedes redactar en tu publicación o cartel?

- **Métodos:** “La variable primaria fue X (continua). Se compararon medias entre grupos mediante prueba t de Welch, por heterogeneidad de varianzas.

- **Resultados:** “El grupo intervención mostró una media inferior de X respecto al control. La diferencia estimada fue de...; IC del 95%...; p...; el tamaño de efecto fue... y supera la diferencia clínicamente importante definida a priori.”

- **Discusión:** “Aunque el valor p fue... el IC sugiere que el efecto plausible se sitúa entre...; esta magnitud [sí/no] tiene relevancia clínica. Los análisis de sensibilidad (transformación log, método no paramétrico) apoyan la robustez de la conclusión.”

CONCLUSION

La t de Student sigue siendo una aliada cotidiana del clínico-investigador: simple, robusta y útil cuando se respeta su ámbito de aplicación. Su poder reside en poner en relación efecto observado y variabilidad para juzgar si una diferencia merece atención. Pero la misma facilidad con la que se ejecuta puede llevar a abusos: aplicarla a variables inadecuadas, ignorar dependencias y varianzas desiguales, atribuir todo al valor p y olvidar la MCID.

La buena práctica exige diseño claro, elección de variante apropiada (a menudo Welch), exploración gráfica, reporte completo con IC y tamaños de efecto, y una interpretación anclada en la relevancia clínica. Así, la prueba t de Student cumple su promesa: ayudar a tomar decisiones mejor informadas, con rigor estadístico y sentido clínico.

REFERENCIAS

- Chicco, D., Sichenze, A., & Jurman, G. (2025). A simple guide to the use of Student's t-test, Mann-Whitney U test, Chi-squared test, and Kruskal-Wallis test in biostatistics. *BioData Min*, 18(1), 56. <https://doi.org/10.1186/s13040-025-00465-6>
- Daniel, W. W., & Chad, L. C. (2019). *Biostatistics. A Foundation for analysis in the health sciences*. (11 ed.). John Wiley & Sons, Inc.
- Murtagh, J., & Bryner, J. (2025). How the Guinness Brewery Invented the Most Important Statistical Method in Science. Retrieved 06-11-2025, from <https://www.scientificamerican.com/article/how-the-guinness-brewery-invented-the-most-important-statistical-method-in/>
- Pandey, R. M. (2015). Commonly used t-tests in medical research. *Journal of the Practice of Cardiovascular Sciences*, 1(2). <https://doi.org/10.4103/2395-5414.166321>
- Weissgerber, T. L., Garcia-Valencia, O., Garovic, V. D., Milic, N. M., & Winham, S. J. (2018). Why we need to report more than 'Data were Analyzed by t-tests or ANOVA'. *Elife*, 7. <https://doi.org/10.7554/eLife.36163>
- White, S. E. (2021). *Bioestadística Básica y Clínica* (5a ed.). McGraw Hill. <https://accessmedicina.mhmedical.com/content.aspx?bookid=3043§ionid=257473577>