Comparaciones por pares de medias en diseños de medidas repetidas

(ANOVA intra-sujetos)

Frías Navarro, D. (2025). *Comparaciones por pares de medias en diseños de medidas repetidas (ANOVA intra-sujetos)*. Universidad de Valencia.

Contexto y objetivos

En los diseños de **medidas repetidas**, las comparaciones por pares de medias se realizan habitualmente con **pruebas *t* de medidas repetidas o para muestras relacionadas** (o contrastes sobre medias marginales) con correcciones o ajuste de la significación estadística (ajusta el valor de **α)** para controlar la tasa de error Tipo I por experimento (FWER). Las correcciones más utilizadas en este contexto son **Bonferroni**, **Šidák** y **Holm-Bonferroni**, porque no requieren suponer independencia entre las comparaciones y mantienen bajo control la probabilidad de cometer al menos un error de Tipo I en el conjunto de pruebas.

A partir de simulaciones Monte Carlo, Maxwell (1980) mostró que, en diseños intra-sujeto, los procedimientos tipo **Tukey no controlan adecuadamente la tasa de error de Tipo I** (aumenta de forma considerable la tasa de error de Tipo I) cuando se viola la esfericidad, incluso de forma moderada, mientras que los contrastes *t* con corrección **Bonferroni** resultan muy robustos frente a estas violaciones (Maxwell, 1980, pp. 282-283). Revisiones posteriores llegan a conclusiones similares y desaconsejan el uso rutinario de Tukey HSD o métodos análogos como el de **Scheffé** en ANOVA de medidas repetidas por estar pensados para comparaciones entre grupos independientes y/o ser innecesariamente conservadores cuando solo se analizan pares de medias (Field, 2013; Park, Cho, & Ki, 2009; Maxwell, Delaney, & Kelley, 2017). Por lo tanto, **Tukey HSD no es una opción adecuada en diseños de medidas repetidas** porque su fundamentación teórica asume grupos independientes y, cuando se viola la esfericidad, deja de controlar correctamente la tasa de error de Tipo I.

Entre las correcciones basadas en *p*-valores, **Bonferroni** es la opción conceptualmente más sencilla (divide α entre el número de comparaciones) y garantiza que la tasa de error Tipo I por experimento no supere el nivel nominal de α, pero puede ser conservadora cuando se realizan muchas comparaciones. La corrección de **Šidák** es un refinamiento ligeramente más potente, (menos conservador) que Bonferroni aunque su derivación asume independencia entre pruebas; en contextos de diseños de medidas repetidas, donde las comparaciones están positivamente correlacionadas, tiende a ser conservador y mantiene el control del error Tipo I (Field, 2013). El procedimiento **Holm-Bonferroni** es un método secuencial que controla fuertemente la tasa de error de Tipo I sin requerir independencia estricta y es, teóricamente, **más potente que Bonferroni simple** en todos los casos (Holm, 1979).

Por ello, siguiendo las recomendaciones metodológicas de Maxwell (1980) y los manuales aplicados de análisis de datos (Field, 2013; Park et al., 2009), es apropiado utilizar **pruebas *t* de medidas repetidas** para las comparaciones por pares, combinadas con una corrección **Holm-Bonferroni** (o, en su defecto, **Bonferroni o Šidák**) para controlar adecuadamente la tasa de error de Tipo I. En cambio, **Tukey HSD** no se empleará como procedimiento estándar en diseños de medidas repetidas, y el método de **Scheffé** se reserva, en su caso, para contrastes complejos, siendo muy conservador cuando solo se comparan pares de medias.

Resumen:

En diseños de medidas repetidas, para comparaciones por pares utilizaremos pruebas *t* para muestras relacionadas con correcciones del tipo **Holm-Bonferroni, Šidák o Bonferroni**.

1) Tasa de error de Tipo I. Bonferroni, Šidák y Holm controlan adecuadamente la tasa de error de Tipo I.

2) Potencia estadística (manteniendo el control de la tasa de error de Tipo I y máxima potencia): las diferencias entre las pruebas de **Holm-Bonferroni, Šidák y Bonferroni**, pero relevantes desde el punto de vista teórico:

- **Holm-Bonferroni** (conocido como procedimiento de Holm o Holm-Bonferroni) es la opción **más potente** entre los métodos que controlan bien la tasa de error de tipo I por experimento. **Holm** (1979) ha demostrado que es más potente que Bonferroni simple: nunca peor, a veces claramente mejor (es decir, rechaza con más facilidad hipótesis nulas falsas manteniendo la tasa de error de Tipo I por experimento $\leq \alpha$), sin asumir independencia estricta. Por lo tanto, es la opción preferente por su mejor equilibrio entre control del error de Tipo I y potencia estadística.
- A continuación, se sitúa la corrección de **Šidák, que** es un poco menos conservador que Bonferroni (por lo tanto, más potente), aunque la diferencia suele ser pequeña), después **Bonferroni** simple (es el más simple de aplicar y el más conservador de los tres), es muy robusto, no requiere independencia y es muy útil y fácil de entender por tener un cálculo sencillo: α ajustado = α / número de comparaciones. Por lo tanto, **Šidák** y **Bonferroni** son alternativas válidas, siendo Bonferroni la más conservadora y la más sencilla de aplicar.

3) Existen dos pruebas que no se recomiendan cuando se trata de diseños de medidas repetidas si el objetivo es realizar comparaciones por pares de medias: **Scheffé y Tukey**. Respecto a la prueba de **Scheffé**, siendo una prueba correcta en el diseño de medidas repetidas, Maxwell (1980) ya señalaba que el procedimiento tipo Scheffé/Roy-Bose en medidas repetidas es muy conservador ("too conservative for practical use") cuando solo nos interesan comparaciones por pares de medias y el tamaño muestral no es grande. También Field (2013) lo valora igual ya que señala que es útil para contrastes complejos, pero no lo recomienda para "post hoc de pares de medias" en diseños de medidas repetidas porque se pierde mucha potencia estadística respecto al resto de pruebas que tienen mayor potencia (menor error de Tipo II) como Holm, Šidák y Bonferroni. Por lo tanto, Scheffé no se usará en este contexto de diseño de medidas repetidas porque, aunque es teóricamente correcto, resulta excesivamente conservador cuando solo se analizan pares de medias.

4) Recordar que la prueba de **Tukey** en diseños de medidas repetidas no es adecuada. La prueba de Tukey está diseñada para comparaciones múltiples en ANOVA entre-sujetos con grupos independientes. En diseños de medidas repetidas no garantiza un control adecuado del error Tipo I, especialmente si la esfericidad se viola. Por lo tanto, Tukey HSD no es una prueba adecuada como opción estándar en diseños de medidas repetidas, ya que se desarrolló para comparaciones entre grupos independientes (diseños entre-grupos) y puede no controlar correctamente la tasa de error de Tipo I cuando se violan los supuestos en diseños intra-sujeto como la esfericidad.

Elementos clave:

En diseños de medidas repetidas las comparaciones por pares deben hacerse con pruebas *t* para muestras relacionadas más procedimientos de ajuste del nivel de significación. Se recomienda Holm-Bonferroni, seguido de Šidák y Bonferroni como alternativas válidas y robustas. Scheffé solo tiene sentido cuando se quieren proteger contrastes complejos, pero es demasiado conservador si solo se comparan pares de medias. Nunca aplicar la corrección de Tukey HSD.

En la práctica, esto significa que, si en un ANOVA de medidas repetidas encontramos un efecto principal estadísticamente significativo y queremos saber qué niveles difieren entre sí, utilizaremos pruebas *t* de medidas repetidas con corrección de Holm-Bonferroni, Šidák o, si se prefiere algo más sencillo, Bonferroni y evitaremos usar Tukey HSD como si estuviéramos en un diseño entre-sujetos.

Referencias

Field, A. P. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). SAGE.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70.

Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. *Journal of Educational Statistics, 5*(3), 269-287. https://doi.org/10.3102/10769986005003269

Park, E., Cho, M., & Ki, C.-S. (2009). Correct use of repeated measures analysis of variance. *Korean Journal of Laboratory Medicine, 29*(1), 1-9. https://doi.org/10.3343/kjlm.2009.29.1.1