Trabajo práctico 11: Comparar grupos

AL TERMINAR ESTA GUÍA

- Comparar dos medias con la prueba t para muestras independientes, y decidir si corresponde la versión con varianza combinada o la de Welch.
- Aplicar la prueba chi-cuadrado de independencia a una tabla de contingencia y medir la fuerza de la asociación.
- Distinguir la significación estadística de la relevancia práctica, con el intervalo de confianza y el tamaño del efecto.
- Leer la salida de un software y elegir la prueba adecuada para cada pregunta.

PARA EL DOCENTE

Los ejercicios 1 y 2 son los centrales del encuentro práctico, con las tablas de la t y de chi-cuadrado. El 3 y el 4 son más cortos y se pueden repartir entre grupos; el 4 prepara para leer salidas en el parcial.

El 5 necesita computadora. El 6 conviene hacerlo al final, en voz alta: es el repaso de toda la unidad y anticipa el ANOVA y la regresión de la semana que viene.

EJERCICIO 1 El peso al nacer y el tabaco

A mano

En una maternidad se registró el peso al nacer (en gramos) de los hijos de 12 madres que no fumaron durante el embarazo y de 10 que fumaron, todos nacidos a término y elegidos al azar de los registros del último año.

No fumadoras 1 a 6	3.210	3.630	2.500	3.630	3.570	3.780
No fumadoras 7 a 12	3.550	3.630	2.760	2.800	2.670	3.170
Fumadoras 1 a 5	3.220	2.720	2.560	2.810	2.940	
Fumadoras 6 a 10	2.170	3.000	3.170	2.600	3.330	

Se puede suponer que el peso al nacer tiene distribución aproximadamente normal en los dos grupos. Se usa $\alpha = 0{,}05$.

a) Planteá las hipótesis para responder si el peso medio al nacer difiere entre los dos grupos.

RESOLUCIÓN

$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$

El subíndice 1 es para las no fumadoras y el 2 para las fumadoras. La pregunta es si difieren, así que la prueba es bilateral, aunque se sospeche para qué lado va la diferencia.

b) Calculá la media y el desvío de cada grupo. ¿Es razonable suponer que las varianzas son iguales?

RESOLUCIÓN

$\bar{x}_1 = 3.241{,}7 \quad s_1 = 452{,}8 \qquad \bar{x}_2 = 2.852{,}0 \quad s_2 = 354{,}0$

Sí: el desvío mayor es 1,28 veces el menor. Una regla práctica habitual acepta varianzas iguales mientras el desvío mayor no llegue al doble del menor. Si no fuera así, se usa la prueba de Welch, que no supone varianzas iguales.

c) Calculá la varianza combinada, el error estándar de la diferencia y el estadístico t.

RESOLUCIÓN

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} = \frac{11 \cdot 452{,}8^2 + 9 \cdot 354{,}0^2}{20} = 169\,146 \qquad s_p = 411{,}3$$

$$EE = s_p \sqrt{\frac{1}{12} + \frac{1}{10}} = 176{,}10 \qquad t = \frac{\bar{x}_1 - \bar{x}_2}{EE} = \frac{389{,}7}{176{,}10} = 2{,}213$$

La varianza combinada es un promedio de las dos varianzas, ponderado por los grados de libertad de cada grupo: si las varianzas poblacionales son iguales, es la mejor estimación de esa varianza común.

d) Tomá la decisión con la región crítica y acotá el p-valor con la tabla de la t.

RESOLUCIÓN

$gl = 12 + 10 - 2 = 20 \qquad |t| = 2{,}213 > t_{20;\,0{,}975} = 2{,}086$

Se rechaza H_0. En la fila de 20 grados de libertad, 2,213 queda entre 2,086 (cola de 0,025) y 2,845 (cola de 0,005); como la prueba es bilateral, se duplican: 0,01 < p < 0,05. El valor exacto es p = 0,0387.

e) Construí el intervalo de confianza del 95 % para la diferencia de medias. ¿Es coherente con la decisión?

RESOLUCIÓN

$389{,}7 \pm 2{,}086 \cdot 176{,}10 = 389{,}7 \pm 367{,}3 \;\Rightarrow\; (22{,}3;\ 757{,}0)$

El intervalo no incluye al cero, igual que la prueba rechaza que la diferencia sea cero: son la misma información vista de dos maneras. Pero el intervalo dice más: la diferencia podría ser de apenas 22 g o de 757 g. La muestra es chica y la estimación, imprecisa.

f) Redactá la conclusión. ¿Se puede afirmar que fumar durante el embarazo causa un menor peso al nacer?

RESOLUCIÓN

«Los hijos de madres fumadoras pesaron al nacer, en promedio, 390 g menos que los de madres no fumadoras; la diferencia es estadísticamente significativa (t = 2,21, gl = 20, p = 0,039).»

Este estudio solo, no: es observacional. Las madres que fuman pueden diferir de las que no fuman en otras cosas que también afectan el peso (edad, nivel socioeconómico, alimentación, controles prenatales). La prueba dice que la diferencia no parece obra del azar; no dice de dónde viene. La relación causal entre tabaco y bajo peso está establecida por muchos estudios distintos, no por una prueba t.

EJERCICIO 2 El control prenatal y el nivel educativo

A mano

Una encuesta a 220 mujeres que tuvieron un hijo en el último año, elegidas al azar en una provincia, registró su nivel educativo y si tuvieron un control prenatal adecuado (cinco controles o más, el primero antes de la semana 13).

Nivel educativo	Control adecuado	Control no adecuado	Total
Hasta primario	34	36	70
Secundario	62	38	100
Superior	37	13	50
Total	133	87	220

a) Calculá el porcentaje de mujeres con control adecuado en cada nivel educativo. ¿Qué sugieren los datos?

RESOLUCIÓN

Nivel educativo	Con control adecuado
Hasta primario	34 de 70: 48,6 %
Secundario	62 de 100: 62,0 %
Superior	37 de 50: 74,0 %

El porcentaje crece con el nivel educativo. Los porcentajes se calculan dentro de cada nivel (por fila), porque la pregunta es cómo cambia el control según la educación. La prueba responde si una diferencia así puede deberse al azar del muestreo.

b) Planteá las hipótesis de la prueba chi-cuadrado de independencia.

RESOLUCIÓN

H_0: el nivel educativo y el control prenatal adecuado son independientes (la proporción con control adecuado es la misma en los tres niveles). H_1: no son independientes (al menos un nivel tiene una proporción distinta). La alternativa no tiene dirección: el chi-cuadrado detecta cualquier forma de asociación.

c) Calculá las frecuencias esperadas bajo H_0. ¿Se cumple la condición para usar la prueba?

RESOLUCIÓN

$$E_{ij} = \frac{\text{total de la fila } i \cdot \text{total de la columna } j}{n} \qquad E_{11} = \frac{70 \cdot 133}{220} = 42{,}32$$

Nivel educativo	Adecuado (esperado)	No adecuado (esperado)
Hasta primario	42,32	27,68
Secundario	60,45	39,55
Superior	30,23	19,77

Sí: la menor frecuencia esperada es 19,77, mayor que 5. La condición se mira sobre las esperadas, no sobre las observadas. Las esperadas mantienen los totales de la tabla y reparten cada fila en la misma proporción que el total: 60,5 % con control adecuado.

d) Calculá el estadístico chi-cuadrado y sus grados de libertad.

RESOLUCIÓN

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Nivel educativo	Aporte (adecuado)	Aporte (no adecuado)
Hasta primario	1,635	2,500
Secundario	0,040	0,060
Superior	1,517	2,320
Suma	**3,192**	**4,880**

$$\chi^2 = 3{,}192 + 4{,}880 = 8{,}072 \qquad gl = (3-1)(2-1) = 2$$

Los aportes más grandes dicen dónde está la diferencia: en las celdas donde lo observado se aleja más de lo esperado bajo independencia.

e) Tomá la decisión con α = 0,05 y acotá el p-valor con la tabla de chi-cuadrado.

RESOLUCIÓN

El valor crítico con 2 grados de libertad y α = 0,05 es 5,991. Como χ^2 = 8,072 > 5,991, se rechaza H_0. Con α = 0,01 el crítico es 9,210 y no se rechazaría: 0,01 < p < 0,05.

Con 2 grados de libertad, el p-valor sale exacto: $p = e^{-\chi^2/2} = e^{-8{,}072/2} = 0{,}0177$

La región crítica está solo a la derecha: un χ^2 chico significa que lo observado se parece a lo esperado, y eso no es evidencia contra la independencia.

f) Redactá la conclusión. Calculá también la V de Cramér. ¿Se puede decir que estudiar más hace que las mujeres se controlen mejor?

RESOLUCIÓN

$$V = \sqrt{\frac{\chi^2}{n \cdot (\min(f, c) - 1)}} = \sqrt{\frac{8{,}072}{220 \cdot 1}} = 0{,}192$$

«Los datos dan evidencia de que el control prenatal adecuado está asociado al nivel educativo (χ² = 8,07, gl = 2, p = 0,018): va del 49 % entre las mujeres con hasta primario al 74 % entre las de nivel superior.» La V de Cramér indica una asociación débil a moderada.

La asociación no prueba causalidad: el nivel educativo va junto con el ingreso, la cobertura de salud, la distancia a un centro de salud y otras cosas que también influyen en el control. Para orientar una política, importa saber cuál de esos factores es el que pesa.

EJERCICIO 3 Significativo no quiere decir importante

A mano

Dos estudios evaluaron un programa de apoyo escolar comparando la nota en una evaluación provincial (de 0 a 100) entre estudiantes que participaron y que no participaron. En los dos, el desvío de la nota fue de 12 puntos en cada grupo.

Estudio	Estudiantes por grupo	Media con el programa	Media sin el programa
A	2.500	62,2	61,0
B	8	68,0	61,0

a) Para el estudio A, calculá el error estándar de la diferencia, el estadístico y el p-valor bilateral. Con muestras tan grandes, la t es prácticamente una normal.

RESOLUCIÓN

$$EE_A = 12\sqrt{\frac{1}{2500} + \frac{1}{2500}} = 0{,}3394 \qquad z = \frac{1{,}2}{0{,}3394} = 3{,}536$$

z queda fuera de la tabla de la normal: el p-valor es menor que 0,001. Con software, p = 0,0004. La diferencia es muy significativa.

b) Para el estudio B, calculá el estadístico t y tomá la decisión con α = 0,05.

RESOLUCIÓN

$$EE_B = 12\sqrt{\frac{1}{8} + \frac{1}{8}} = 6{,}000 \qquad t = \frac{7}{6{,}000} = 1{,}167 \qquad t_{14;\,0{,}975} = 2{,}145$$

No se rechaza H_0: $|t| < 2{,}145$ (p = 0,2628). La diferencia observada es 5,8 veces la del estudio A, pero con ocho estudiantes por grupo no alcanza para descartar el azar.

c) Calculá los intervalos de confianza del 95 % para la diferencia en los dos estudios.

RESOLUCIÓN

$$A: 1{,}2 \pm 1{,}96 \cdot 0{,}3394 \Rightarrow (0{,}53;\ 1{,}87) \qquad B: 7 \pm 2{,}145 \cdot 6{,}000 \Rightarrow (-5{,}87;\ 19{,}87)$$

El intervalo de A es angosto y está lejos de cero, pero todos sus valores son chicos. El de B incluye al cero y también diferencias enormes: es compatible con que el programa no sirva para nada y con que sea muy bueno.

d) Calculá el tamaño del efecto (la diferencia dividida por el desvío) en cada estudio.

RESOLUCIÓN

$$d_A = \frac{1{,}2}{12} = 0{,}10 \qquad d_B = \frac{7}{12} = 0{,}58$$

A diferencia del p-valor, el tamaño del efecto no depende del tamaño de la muestra. Como guía muy general, 0,2 se considera un efecto chico, 0,5 mediano y 0,8 grande; pero si un efecto importa o no lo decide quien conoce el tema, no una tabla.

e) Un diario titula: «El estudio A prueba que el programa funciona; el B, que no sirve». ¿Qué le corregirías?

RESOLUCIÓN

Las dos mitades. El estudio A muestra que el efecto existe, pero es de alrededor de 1,2 puntos en 100: si eso justifica el costo del programa es una discusión de política educativa, no estadística. El estudio B no muestra que el programa no sirva: muestra que con ocho estudiantes por grupo no se puede saber. «No significativo» no es «sin efecto».

Con una muestra enorme, casi cualquier diferencia da significativa; con una muy chica, casi ninguna. Por eso hay que informar la estimación y su intervalo, no solo el p-valor.

EJERCICIO 4 Leer la salida de un software

A mano

Un equipo de bienestar estudiantil comparó las horas de sueño por noche de estudiantes universitarios que trabajan (22) y que no trabajan (31), elegidos al azar. El software dio esta salida:

Grupo	n	Media (horas)	Desvío
Trabajan	22	6,11	1,57
No trabajan	31	7,04	0,83

Prueba t de Welch, bilateral	t	gl	p-valor	IC 95 % de la diferencia
Trabajan − No trabajan	−2,538	29,35	0,0167	−1,679 a −0,181

a) ¿Qué hipótesis pone a prueba la salida? ¿Por qué tiene sentido que el software use la prueba de Welch y no la t con varianza combinada?

RESOLUCIÓN

$$H_0 : \mu_{\text{trabajan}} = \mu_{\text{no trabajan}} \qquad H_1 : \mu_{\text{trabajan}} \neq \mu_{\text{no trabajan}}$$

Los desvíos son muy distintos: 1,57 contra 0,83, casi el doble. La prueba de Welch no supone varianzas iguales, y muchos programas (R entre ellos) la usan por defecto. Cuando las varianzas sí son parecidas, da casi lo mismo que la otra.

b) Verificá a mano el estadístico t a partir de las medias y los desvíos.

RESOLUCIÓN

$$EE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{1{,}57^2}{22} + \frac{0{,}83^2}{31}} = 0{,}3664 \qquad t = \frac{6{,}11 - 7{,}04}{0{,}3664} = -2{,}538$$

Cada grupo aporta su propia varianza al error estándar: no se combinan. Es negativo porque el software restó en el orden «trabajan − no trabajan»; el signo depende de ese orden y no cambia la conclusión.

c) ¿Por qué los grados de libertad tienen decimales? ¿Cuántos serían con la prueba de varianza combinada?

RESOLUCIÓN

La prueba de Welch aproxima la distribución del estadístico con una t cuyos grados de libertad salen de una fórmula (la de Welch-Satterthwaite), que en general no da un entero. Están entre el tamaño del grupo más chico menos uno (21) y $n_1 + n_2 - 2 = 51$, que es lo que usaría la prueba de varianza combinada. Acá dan 29,35, más cerca del mínimo, porque el grupo más variable es el más chico.

d) ¿Qué se concluye con $\alpha = 0{,}05$? Interpretá el intervalo de confianza.

RESOLUCIÓN

Como $p = 0{,}0167 < 0{,}05$, se rechaza H_0: los estudiantes que trabajan duermen, en promedio, menos que los que no trabajan. El intervalo dice cuánto menos: entre 11 y 101 minutos por noche, con 95 % de confianza (la diferencia estimada es de 56 minutos).

El intervalo no incluye al cero, coherente con el rechazo. Pasar de horas a minutos ayuda a juzgar si la diferencia importa, que es lo que le interesa a quien lee el informe.

e) Si antes de recolectar los datos el equipo hubiera planteado que quienes trabajan duermen menos (prueba unilateral), ¿cuánto valdría el p-valor?

RESOLUCIÓN

$$p_{\text{unilateral}} = \frac{0{,}0167}{2} = 0{,}0084$$

La mitad, porque la diferencia observada va en la dirección planteada. Si hubiera ido para el otro lado, el p-valor unilateral sería 1 − p/2, cercano a uno. Y otra vez: la dirección se decide antes de ver los datos.

EJERCICIO 5 Varones y mujeres en la EPH

Con software

Se usa la base «Ocupación e ingresos en Gran Catamarca», extracto de la EPH del primer trimestre de 2026 (estadistica.ar/datos/eph-catamarca-2026t1). Se puede usar R, Python, Excel, InfoStat o SPSS. En R, t.test(P21 ~ CH04, data = ...) hace la prueba de Welch y

chisq.test(table(...)) la de independencia. CH04 vale 1 para varones y 2 para mujeres.

Como en las semanas anteriores, se trata a la base como una muestra aleatoria simple de personas, sin el ponderador ni el diseño muestral de la EPH.

a) Entre las personas ocupadas que declararon ingreso (ESTADO = 1 y P21 > 0), compará el ingreso medio de la ocupación principal de varones y mujeres con la prueba de Welch.

> RESOLUCIÓN
>
> | Sexo | n | Media | Mediana | Desvío |
> |---|---|---|---|---|
> | Varones | 323 | 894.560 | 750.000 | 816.306 |
> | Mujeres | 271 | 867.011 | 650.000 | 1.062.148 |
>
> t = 0,349, gl = 500,8, p = 0,7271. IC 95 % de la diferencia (varones − mujeres): de $−127.476 a $182.575. No se rechaza H_0: con estos datos, no hay evidencia de que el ingreso medio difiera entre varones y mujeres.

b) ¿Quiere decir eso que varones y mujeres ganan lo mismo? Mirá las medianas, el intervalo y repetí la prueba con el logaritmo del ingreso.

> RESOLUCIÓN
>
> No. El intervalo va de una diferencia a favor de las mujeres a una de unos $183.000 a favor de los varones: la muestra no alcanza para saberlo. Las medianas difieren más que las medias (750.000 contra 650.000): unos pocos ingresos muy altos de mujeres levantan su media.
>
> Con el logaritmo, que achica la influencia de esos valores extremos, t = 1,554 y p = 0,1208: tampoco es significativo, aunque el p-valor baja mucho. Además, esta comparación no tiene en cuenta las horas trabajadas ni el tipo de ocupación: una brecha de ingresos se estudia con esas variables controladas.

c) Con todas las personas ocupadas, armá la tabla de categoría ocupacional (CAT_OCUP) por sexo y aplicá la prueba chi-cuadrado de independencia.

> RESOLUCIÓN
>
> | Categoría ocupacional | Varones | Mujeres | % de varones | % de mujeres |
> |---|---|---|---|---|
> | Patrón o empleador | 33 | 14 | 9,5 % | 4,9 % |
> | Cuenta propia | 76 | 43 | 21,9 % | 15,1 % |
> | Obrero o empleado | 238 | 227 | 68,6 % | 79,9 % |
>
> χ^2 = 10,911, gl = 2, p = 0,0043. Se rechaza la independencia: la distribución por categoría ocupacional es distinta entre varones y mujeres. Las mujeres ocupadas son más a menudo asalariadas, y menos a menudo patronas o cuentapropistas. Acá los porcentajes van por columna, porque se compara cómo se reparte cada sexo.

d) Calculá la V de Cramér. ¿Es una asociación fuerte?

RESOLUCIÓN

$$V = \sqrt{\frac{10{,}911}{631 \cdot 1}} = 0{,}131$$

Es débil: muy significativa, pero de poca intensidad. Con más de 600 casos, diferencias moderadas de porcentajes ya dan p-valores chicos. Significación y fuerza de la asociación son cosas distintas, y hay que informar las dos.

e) Si se compararan los ingresos de varones y mujeres por separado en cada uno de los seis niveles educativos, y en uno solo diera significativo, ¿qué conclusión sacarías?

RESOLUCIÓN

Muy poca. Con seis pruebas al 5 % y H_0 cierta en todas, la probabilidad de que al menos una dé significativa es $1 - 0{,}95^6 = 0{,}265$: casi una de cada cuatro veces. Además, varios niveles tienen pocos casos. Antes de hacer muchas comparaciones hay que decidir cuáles importan, o corregir el nivel (por ejemplo, con Bonferroni: 0,05 / 6 para cada prueba).

EJERCICIO 6 ¿Qué prueba uso?

Para discutir

Para cada pregunta, decidí qué tipo de variables están en juego, si las muestras son independientes o apareadas, y qué prueba usarías. No hace falta resolver nada.

a) ¿El peso medio de los recién nacidos de una maternidad difiere de 3.300 g, el valor de referencia?

RESOLUCIÓN

Una variable cuantitativa, un solo grupo y un valor de referencia: prueba t para una media (semana 10).

b) ¿Los estudiantes de una comisión rinden distinto en el primer y en el segundo parcial?

RESOLUCIÓN

Las mismas personas medidas dos veces: las muestras son apareadas. Corresponde la prueba t pareada, que trabaja con la diferencia de cada estudiante. Usar la prueba para muestras independientes desperdicia que cada estudiante es su propio control, y pierde potencia.

c) ¿La proporción de personas que fuman es distinta entre varones y mujeres?

RESOLUCIÓN

Dos variables categóricas (sexo y fumar o no): chi-cuadrado de independencia en una tabla de 2 × 2. Es equivalente a la prueba de diferencia de dos proporciones: el χ^2 es el cuadrado del z.

d) ¿El ingreso medio difiere entre los tres aglomerados de una región?

RESOLUCIÓN

Una cuantitativa comparada entre tres grupos: análisis de la varianza (ANOVA), que se ve la semana que viene. Hacer tres pruebas t de a pares infla la probabilidad de un falso positivo. Además, con ingresos muy asimétricos conviene usar el logaritmo o una prueba no paramétrica (Kruskal-Wallis).

e) ¿Hay relación entre las horas semanales de estudio y la nota final?

> RESOLUCIÓN
>
> Dos variables cuantitativas medidas en las mismas personas: correlación y regresión lineal (semanas 4 y 12). No se compara grupos sino la forma en que una variable cambia con la otra.

f) Con un sistema nuevo de turnos, ¿bajó el tiempo de espera en una guardia? Se tienen 12 esperas antes y 12 después, de pacientes distintos, y unas pocas esperas son larguísimas.

> RESOLUCIÓN
>
> Dos grupos independientes y una variable cuantitativa, pero muy asimétrica y con muestras chicas: el supuesto de normalidad de la prueba t es dudoso. Conviene la prueba de Mann-Whitney (no paramétrica), o la t sobre el logaritmo del tiempo. La prueba es unilateral si la pregunta, planteada de antemano, es si bajó.

g) ¿La opinión sobre una política (a favor, en contra, indiferente) está asociada al nivel educativo (primario, secundario, superior)?

> RESOLUCIÓN
>
> Dos variables categóricas: chi-cuadrado de independencia en una tabla de 3 × 3, con 4 grados de libertad. Hay que verificar que las frecuencias esperadas no sean chicas; si lo son, se pueden agrupar categorías.

---

estadistica.ar/docentes/trabajos-practicos/semana-11 · Las lecturas de cada tema y la hoja de fórmulas están en el sitio.