Contraste de dos muestras

Josep Gibergans Bàguena

P08/75057/02309

Universitat Oberta de Catalunya
www.uoc.edu

Índice

Contrastes sobre la diferencia de medias

1. Introducción

En esta sesión veremos cómo debemos hacer un contraste de hipótesis sobre diferencias de medias poblacionales. Consideraremos dos muestras de observaciones y compararemos sus medias contrastando hipótesis sobre su diferencia y construyendo intervalos de confianza para esta diferencia.

¡Atención!

Es muy importante no confundir este tipo de problemas con los de "datos aparejados", en los que tenemos una muestra de observaciones de dos variables.

Por ejemplo, podemos estar interesados en conocer si hay una diferencia significativa entre los aspectos siguientes:

a) Los sueldos de los hombres y las mujeres que trabajan en una misma empresa.

b) El consumo de combustible de dos vehículos de marcas diferentes.

c) El tiempo de vida de bombillas de dos marcas diferentes, etc.

2. Contrastes sobre la diferencia de medias

Supongamos que tenemos una muestra aleatoria de tamaño n_1 obtenida de una población 1 de media μ_1 y una muestra aleatoria independiente de la anterior de tamaño n_2 obtenida de una población 2 de media μ_2. Queremos contrastar la hipótesis nula (H_0) que afirma que los valores de las medias de las dos poblaciones son iguales:

1) Hipótesis nula: H_0: $\mu_1 = \mu_2$

Notación

A veces en lugar de:

H_0: $\mu_1 = \mu_2$

escribiremos:

H_0: $\mu_1 - \mu_2 = 0$

2) Hipótesis alternativa. La hipótesis alternativa (H_1) puede ser bilateral o unilateral:

a) Bilateral. La media de una población es superior o inferior a la de la otra población.

$$H_1: \quad \mu_1 \neq \mu_2$$

b) Unilateral, en la que se consideran dos situaciones:

- La media de la población 1 es superior a la media de la población 2:

$$H_1: \quad \mu_1 > \mu_2$$

- La media de la población 1 es inferior a la media de la población 2:

$$H_1: \quad \mu_1 < \mu_2$$

De la misma manera que en el contraste de la media, supondremos poblaciones normales y consideraremos el caso de varianzas poblacionales conocidas y el de varianzas poblacionales desconocidas.

2.1. Caso de varianzas poblacionales conocidas

A partir de dos poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$ se toman dos muestras de tamaños n_1 *y* n_2. Las medias muestrales $\overline{X}_1$ y $\overline{X}_2$ están distribuidas según las distribuciones normales:

$$\overline{X_1} \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \quad \text{y} \quad \overline{X_2} \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$$

A partir de las propiedades de la distribución normal, la variable aleatoria $\overline{X_1} - \overline{X_2}$ también se distribuye normalmente:

$$(\overline{X}_1 - \overline{X}_2) \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Suma de distribuciones

Recordemos que dadas dos distribuciones normales:
$Z_1 \sim N(\mu_1, \sigma_1)$ y $Z_2 \sim N(\mu_2, \sigma_2)$
la variable $Z_1 + Z_2$ se distribuye según una normal:

$$N\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

Tipificando esta variable aleatoria, obtenemos una nueva variable que se distribuye según una $N(0,1)$:

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Si suponemos que la hipótesis nula es cierta, entonces $\mu_1 - \mu_2 = 0$; por tanto:

$$\frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

En resumen, bajo el supuesto de la hipótesis nula cierta (H_0: $\mu_1 = \mu_2$) tenemos que el estadístico de contraste:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma_{\overline{X}_1 - \overline{X}_2}}$$, donde $\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ es el error **estándar,**

corresponde a una observación de una ley $N(0,1)$.

Una vez que hemos calculado el valor del estadístico de contraste, debemos determinar el *p*-valor. El *p*-valor depende de la hipótesis alternativa planteada:

- Si H_1: $\mu_1 - \mu_2 \neq 0$, entonces $p = 2P(Z > |z|)$
- Si H_1: $\mu_1 - \mu_2 < 0$, entonces $p = P(Z < z)$
- Si H_1: $\mu_1 - \mu_2 > 0$, entonces $p = P(Z > z)$

El *p*-valor

El *p*-valor es la probabilidad de que un resultado sea "al menos tan extremo" como el estadístico de contraste obtenido.

Diremos que el *p*-valor es significativo y rechazaremos la hipótesis nula si es menor que el nivel de significación α fijado.

Ejemplo de contraste sobre la diferencia de medias en el caso de varianzas poblacionales conocidas

Un fabricante de vidrios quiere comparar la resistencia media de los vidrios que fabrica con la resistencia de los que fabrica la competencia. Se toma una muestra de cristales de cada fabricante y se miden las respectivas resistencias. Se obtienen los resultados siguientes:

Fabricante:	$n_1 = 150$,	$\bar{x}_1 = 111{,}2$
Competencia:	$n_2 = 125$,	$\bar{x}_2 = 109{,}6$

Suponiendo que las muestras son independientes y se han obtenido de dos poblaciones normales con desviaciones típicas conocidas $\sigma_1 = 10{,}4$ y $\sigma_2 = 12{,}5$, ¿con qué conclusión podemos llegar al 5% de significación?

1. Expresamos las hipótesis:

- Hipótesis nula: H_0: $\mu_1 - \mu_2 = 0$
- Hipótesis alternativa: H_1: $\mu_1 - \mu_2 \neq 0$

2. Determinamos el nivel de significación: $\alpha = 0{,}05$

3. Calculamos el estadístico de contraste, que sigue una distribución $N(0,1)$:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{111{,}2 - 109{,}6}{1{,}40} = 1{,}14$$

Sin embargo, antes hemos tenido que calcular el error estándar:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{10{,}4^2}{150} + \frac{12{,}5^2}{125}} = 1{,}40$$

4. Ahora ya estamos en condiciones de calcular el *p*-valor:

$$p = 2P(Z > |z|) = 2\,P(Z > 1{,}14) = 2 \cdot 0{,}13 = 0{,}26$$

5. Puesto que $0{,}26 > 0{,}05$, no podemos rechazar la hipótesis nula, es decir, no podemos decir que las resistencias de los vidrios de los diferentes fabricantes sean diferentes.

2.2. Caso de varianzas poblacionales desconocidas pero iguales

En la vida real prácticamente nunca se conocen las varianzas poblacionales, por lo que es importante este caso, en el que consideramos que las varianzas poblacionales son desconocidas pero iguales a cierto valor σ^2, es decir,

$\sigma_1^2 = \sigma_2^2 = \sigma^2$, con σ desconocida. Esta desviación típica común σ se puede estimar por medio de la fórmula:

$$s = \sqrt{\frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{i2} - \bar{x}_2)^2\right)}$$

Fórmula

Esta fórmula para estimar σ es parecida a la de la desviación típica para una única muestra. La diferencia está en el hecho de que se suman los totales de los términos al cuadrado por separado y después se dividen por el tamaño muestral total menos dos.

donde x_{i1} es la *i*-ésima observación de la muestra 1 y x_{i2} es la *i*-ésima observación de la muestra 2. También podemos escribir esta expresión de la manera siguiente:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

donde s_1^2 y s_2^2 son las varianzas muestrales obtenidas a partir de las muestras.

Si la hipótesis nula es cierta ($\mu_1 = \mu_2$), obtenemos el **estadístico de contraste** siguiente:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{(\bar{X}_1 - \bar{X}_2)}} \text{ con un } \textbf{error estándar } s_{(\bar{x}_1 - \bar{x}_2)} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

En este caso el estadístico de contraste corresponde a una observación de una distribución *t* de Student con $n_1 + n_2 - 2$ grados de libertad.

Como siempre, a continuación calcularemos el *p*-valor correspondiente al estadístico de contraste calculado. Dependiendo de la hipótesis alternativa, tenemos:

- Si H_1: $\mu_1 - \mu_2 \neq 0$, entonces $p = 2P(t_{n1 + n2 - 2} > |t|)$
- Si H_1: $\mu_1 - \mu_2 < 0$, entonces $p = P(t_{n1 + n2 - 2} < t)$
- Si H_1: $\mu_1 - \mu_2 > 0$, entonces $p = P(t_{n1 + n2 - 2} > t)$

Según el *p*–valor obtenido en comparación con el nivel de significación α escogido, rechazaremos o no la hipótesis nula de igualdad de medias.

Ejemplo de contraste sobre la diferencia de medias en el caso de varianzas poblacionales desconocidas pero iguales

Un fabricante de bombillas asegura que sus bombillas tienen una mayor duración que las de una nueva marca coreana. A partir de la duración (en horas) de $n_1 = 25$ bombillas del fabricante y de $n_2 = 15$ bombillas de la nueva marca, elegidas de forma aleatoria, hemos obtenido:

Para el fabricante: $\bar{x}_1 = 827$, $s_1^2 = 9.005$

Para la nueva marca: $\bar{x}_2 = 812$, $s_2^2 = 7.984$

Supondremos que las dos poblaciones se distribuyen normalmente con varianzas iguales y desconocidas. Haremos un contraste de hipótesis a un nivel del 0,05 para determinar si, tal como parece, el fabricante tiene razón.

Debemos contrastar la diferencia de medias para saber si hay una diferencia significativa o podemos considerar que éstas son iguales. Nos dan las medias y las varianzas muestrales y desconocemos las varianzas poblacionales, que supondremos iguales. Con todos estos supuestos, hacemos lo siguiente:

1. Expresamos las hipótesis:

- Hipótesis nula: H_0: $\mu_1 - \mu_2 = 0$
- Hipótesis alternativa: H_1: $\mu_1 - \mu_2 > 0$

2. Determinamos el nivel de significación: $\alpha = 0{,}05$

3. Calculamos el estadístico de contraste:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{827 - 812}{0{,}959} = 15{,}64$$

que sigue una distribución t de Student con $n_1 + n_2 - 2 = 38$ grados de libertad y en la que:

$$s_{\bar{x}_1 - \bar{x}_2} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 2.937\sqrt{\frac{1}{25} + \frac{1}{15}} = 0{,}959$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(25-1)9{,}005 + (15-1)7{,}984}{25 + 15 - 2}} = \sqrt{\frac{24 \cdot 9{,}005 + 14 \cdot 7{,}984}{38}} =$$

$$= \sqrt{\frac{216{,}120 + 111{,}776}{38}} = \sqrt{\frac{327{,}896}{38}} = \sqrt{8{,}628842} = 2{,}937$$

4. Calculamos el p-valor: $p = P(t_{38} > t) \cong 0$

5. Puesto que $0 < 0{,}05$, entonces rechazamos la hipótesis nula en favor de la hipótesis alternativa. Efectivamente, el fabricante tiene razón, es decir, la media de vida de sus bombillas es superior a la de la nueva marca.

2.3. Caso con muestras grandes no normales

En el caso de que no se pueda asegurar que las muestras provienen de poblaciones normales, sólo podremos contrastar la diferencia de medias si los tamaños de las muestras son superiores a treinta.

El teorema del límite central nos dice que si los tamaños de las muestras son superiores a 30, el **estadístico de contraste**:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

es una observación de una variable aleatoria que se distribuye aproximadamente como una $N(0,1)$.

En este módulo no consideraremos el caso de muestras pequeñas ($n < 30$) de distribuciones no normales.

Ejemplo de contraste sobre la diferencia de medias en el caso de muestras grandes no normales

Queremos saber si existe una diferencia significativa entre el consumo de gasolina de dos motores de coche diferentes. Recogemos los datos sobre el consumo a partir de una muestra aleatoria e independiente de cada tipo de motor y obtenemos los resultados siguientes:

Motor 1:	$n_1 = 80$,	$\bar{x}_1 = 11{,}2$,	$s_1 = 2{,}2$
Motor 2:	$n_2 = 75$,	$\bar{x}_2 = 11{,}8$,	$s_2 = 3{,}7$

No tenemos información sobre el tipo de distribución que tienen los consumos de estos motores. Con un nivel de significación del 1%, ¿podemos asegurar que el consumo es el mismo?

Siguiendo el mismo planteamiento de siempre, haremos un contraste de hipótesis de la diferencia de medias.

1. Expresamos las hipótesis:

- Hipótesis nula: H_0: $\mu_1 - \mu_2 = 0$
- Hipótesis alternativa: H_1: $\mu_1 - \mu_2 \neq 0$

2. Determinamos el nivel de significación: $\alpha = 0{,}01$

3. Calculamos el estadístico de contraste: en este caso no conocemos la distribución del consumo, pero dado que las muestras tienen un tamaño superior a 30, el estadístico de contraste viene dado por:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{11{,}2 - 11{,}8}{0{,}493} = -1{,}217 \approx -1{,}22$$

que sigue una distribución $N(0,1)$ y donde:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{2{,}2^2}{80} + \frac{3{,}7^2}{75}} = 0{,}493$$

4. Calculamos el p–valor: $p = 2P(Z > |z|) = 2P(Z > |-1{,}22|) = 2 \cdot 0{,}1112 = 0{,}2224$

5. Dado que $0{,}2224 > 0{,}01$, entonces no rechazamos la hipótesis nula en favor de la hipótesis alternativa, es decir, el consumo no es significativamente diferente para cada tipo de motor.

3. Intervalos de confianza para la diferencia de medias

El procedimiento que hay que seguir para construir un intervalo de confianza en torno a la diferencia de medias es el siguiente:

1) Fijamos el nivel de confianza, que escribimos como $(1 - \alpha)$.

2) Calculamos el error estándar de la diferencia de medias.

3) Calculamos el valor crítico teniendo en cuenta el tipo de distribución de nuestro estadístico de contraste.

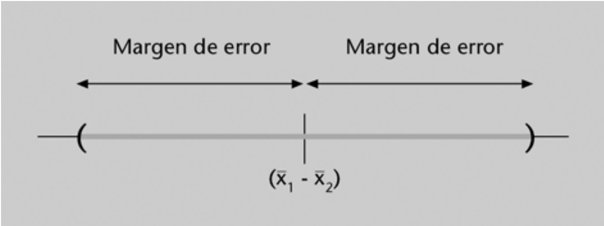
4) Calculamos el margen de error, que es el producto entre el valor crítico y el error estándar.

El **intervalo de confianza para la diferencia de medias** viene dado por:

$$(\bar{x}_1 - \bar{x}_2) \pm \text{(valor crítico) (error estándar)}$$

Este intervalo de confianza contiene la diferencia de medias con un nivel de certeza igual a $(1 - \alpha)$%.

Gráficamente, podemos representarlo así:



A continuación buscaremos para cada caso el intervalo de confianza correspondiente.

3.1. Caso de varianzas poblacionales conocidas

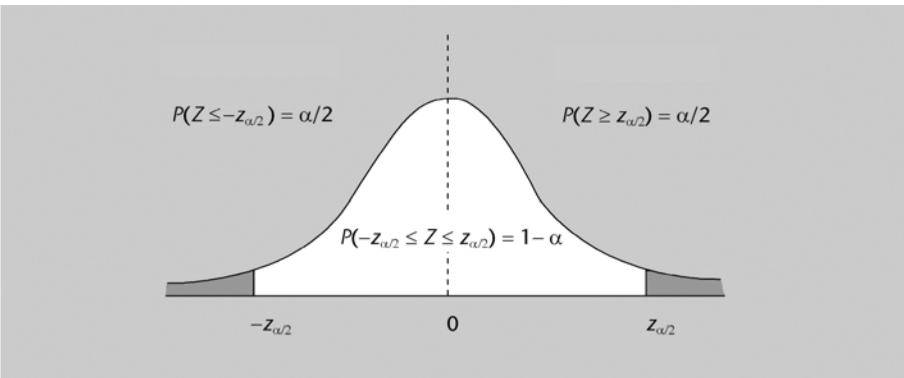
Ya hemos dicho antes que, dadas dos muestras de tamaños n_1 y n_2 obtenidas de dos poblaciones normales $N(\mu_1, \sigma_1)$ *y* $N(\mu_2, \sigma_2)$, la diferencia de sus medias se distribuye según una ley normal:

$$(\overline{X}_1 - \overline{X}_2) \sim N\left(\mu_1 - \mu_2,\ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Supongamos que queremos un nivel de confianza del $(1 - \alpha)\%$. En primer lugar, consideraremos la variable tipificada:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

y a continuación construiremos un intervalo centrado en $Z = 0$, de manera que la probabilidad de que la variable aleatoria Z tome un valor en este intervalo sea de $1 - \alpha$:



$$P\left(-z_{\alpha/2} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

donde $z_{\alpha/2}$ y $-z_{\alpha/2}$ son los valores críticos. Son aquellos que hacen que $P(Z \geq z_{\alpha/2}) =$ $= \alpha/2$ y $P(Z \leq -z_{\alpha/2} = \alpha/2$. Trabajando un poco con esta expresión, tenemos:

$$((\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sigma_{\bar{X}_1 - \bar{X}_2} \leq (u_1 - u_2) \leq (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sigma_{\bar{X}_1 - \bar{X}_2})$$

Error estándar

Recordad que el error estándar de la diferencia de medias es:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Finalmente, sustituyendo los valores muestrales, obtendremos el correspondiente **intervalo de confianza** para la diferencia de medias en el caso de varianzas conocidas:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sigma_{\bar{x}_1 - \bar{x}_2}$$

Ejemplo de intervalos de confianza para la diferencia de medias en el caso de varianzas poblacionales conocidas

Si consideramos el ejemplo del fabricante de vidrios, podemos calcular un intervalo de confianza del 95% para la diferencia de las medias de las resistencias de la manera que explicamos a continuación.

Ya tenemos la diferencia de medias y el error estándar de los cálculos anteriores:

$$(\bar{x}_1 - \bar{x}_2) = 111{,}2 - 109{,}6 = 1{,}6 \quad \text{y} \quad \sigma_{(\bar{x}_1 - \bar{x}_2)} = 1{,}40$$

Los valores críticos para un $\alpha/2 = 0{,}025$ son $\pm z_{\alpha/2} = \pm 1{,}96$. El intervalo de confianza es:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sigma_{\bar{x}_1 - \bar{x}_2}$$

$$1{,}6 \pm 1{,}96 \cdot 1{,}40, \text{ es decir, } 1{,}6 \pm 2{,}74$$

Por tanto, el intervalo de confianza es (–1,14; 4,34). Es importante observar que el intervalo contiene el valor cero, de manera que tenemos un 95% de confianza al pensar que la resistencia de los vidrios es la misma. Este resultado se ha obtenido antes haciendo el contraste de hipótesis.

El intervalo de confianza también nos sirve para hacer contrastes de hipótesis en caso de que la hipótesis alternativa sea bilateral.

3.2. Caso de varianzas poblacionales desconocidas e iguales

En el contraste de hipótesis para diferencias de medias en el caso de varianzas desconocidas e iguales, hemos visto que la variable:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

sigue una distribución t de Student con $n_1 + n_2 - 2$ grados de libertad. De manera que si fijamos un nivel de confianza $1 - \alpha$, podemos determinar los valores críticos $-t_{\alpha/2,n_1+n_2-2}$ y $t_{\alpha/2,n_1+n_2-2}$ que aseguran que:

$$P\left(-t_{\alpha/2,n_1+n_2-2} \leq \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1-\bar{x}_2}} \leq t_{\alpha/2,n_1+n_2-2}\right) = 1 - \alpha$$

Por tanto:

$$(\overline{X}_1 - \overline{X}_2) - t_{\alpha/2,n_1+n_2-2} \cdot s_{\bar{x}_1-\bar{x}_2} \leq \mu_1 - \mu_2 \leq (\overline{X}_1 - \overline{X}_2) + t_{\alpha/2,n_1+n_2-2} \cdot s_{\bar{x}_1-\bar{x}_2}$$

Y sustituyendo valores muestrales, tendremos el intervalo **de confianza** para la diferencia de medias en el caso de varianzas desconocidas e iguales:

$$(\bar{x}_1 - \bar{x}_2) \mp t_{\alpha/2,n_1+n_2-2} \cdot s_{\bar{x}_1-\bar{x}_2}$$

Ejemplo de intervalos de confianza para la diferencia de medias en el caso de varianzas poblacionales desconocidas e iguales

Si consideramos ahora los datos del ejemplo del fabricante de bombillas, calcularemos un intervalo de confianza para la diferencia de duraciones. Ya habíamos encontrado la diferencia de medias y el error estándar:

$$(\bar{x}_1 - \bar{x}_2) = 827 - 812 = 15 \ \text{ y } \ s_{\bar{x}_1-\bar{x}_2} = 0{,}959$$

Los valores críticos para $\alpha/2 = 0{,}025$ son $\pm t_{\alpha/2,n_1+n_2-2} = \pm t_{0{,}025;38} = \pm 2{,}0244$

El intervalo de confianza es: $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,n_1+n_2-2} \cdot s_{\bar{x}_1-\bar{x}_2}$

$$15 \pm 2{,}0244 \cdot 0{,}959, \text{ es decir, } 15 \pm 1{,}9414$$

Por tanto, el intervalo de confianza para la diferencia de duraciones es: (13,06; 16,94). En este caso podemos ver que el intervalo no contiene el cero y que es positivo, por lo que podemos concluir que hay evidencias de que las bombillas del fabricante duran más que las nuevas bombillas coreanas.

3.3. Caso con muestras no normales

Si las muestras no son normales pero sus tamaños son superiores a treinta, entonces, por el teorema del límite central, tenemos que la variable:

$$\frac{(\overline{X}_1 - \overline{X}_2)}{s_{\bar{x}_1-\bar{x}_2}}$$

está distribuida $N(0,1)$. De manera que hacemos las mismas consideraciones que en el caso del apartado 3.1.

El **intervalo de confianza** para la diferencia de medias para el caso de muestras no normales grandes (tamaño superior a treinta) viene dado por:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} s_{\bar{x}_1 - \bar{x}_2}$$

Ejemplo de intervalos de confianza para la diferencia de medias en el caso de muestras no normales

Consideremos ahora el ejemplo del consumo de gasolina de dos tipos de motores para ilustrar este último caso. Teníamos los datos siguientes, la diferencia de medias y el error estándar:

$$(\bar{x}_1 - \bar{x}_2) = 11{,}2 - 11{,}8 = -0{,}6 \text{ y } s_{\bar{x}_1 - \bar{x}_2} = 0{,}493$$

Los valores críticos para $\alpha/2 = 0{,}025$ son $\pm z_{\alpha/2} = \pm 1{,}96$.

El intervalo de confianza es:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \cdot s_{\bar{x}_1 - \bar{x}_2}$$

$$-0{,}6 \pm 1{,}96 \cdot 0{,}493, \text{ es decir, } -0{,}6 \pm 0{,}96628$$

Por tanto, el intervalo de confianza es: (–1,57; 0,366). Es importante observar que el intervalo contiene el valor cero. Este resultado es totalmente coherente con el obtenido con el contraste de hipótesis según el cual el consumo no es significativamente diferente para cada tipo de motor.

4. Resumen

En esta sesión hemos visto cómo hacer contrastes de hipótesis para la diferencia de medias de dos muestras aleatorias e independientes. Hemos distinguido tres casos:

1) Muestras normales con varianzas poblacionales conocidas
2) Muestras normales con varianzas poblacionales desconocidas e iguales
3) Muestras grandes no normales

También hemos aprendido a construir intervalos de confianza para la diferencia de medias considerando estos mismos tres casos.

Ejercicios

1. Una tienda de ordenadores portátiles equipa sus modelos con baterías de la marca Duramás. Estas baterías son de buena calidad, pero pasado cierto tiempo, comienzan a dar problemas. Para intentar dar una mejor calidad a las ventas, el responsable del negocio se plantea la posibilidad de cambiar la marca de baterías por la de Enerplus. Dado que el precio de estas nuevas baterías es superior al de las de la marca Duramás, antes de tomar una decisión quiere tener la seguridad de que con este cambio gana calidad en el producto final. Se prueban cincuenta baterías Duramás y cincuenta y cinco Enerplus, y se obtienen unas duraciones medias de treinta y siete, y cuarenta y tres meses, respectivamente. Suponiendo que las desviaciones típicas de las dos marcas son conocidas e iguales a cinco meses, ¿creéis que se ganará calidad con el cambio de marca?

2. Se quiere probar que, a la hora de cargarse al poner en funcionamiento un ordenador, el sistema operativo A es más rápido que el sistema operativo B. Se han medido los tiempo de arranque en seis ordenadores equipados con el sistema A y en otros seis con el sistema B. Los tiempos (en segundos) han sido los siguientes:

Sistema A	10,7	14,8	12,3	16,5	10,2	11,9
Sistema B	13,4	11,5	11,2	15,1	13,3	12,9

a) ¿Se puede aceptar la afirmación con un nivel de significación del 5%?
b) Calculad un intervalo de confianza del 90% para la diferencia de medias.

Supondremos que los tiempos están normalmente distribuidos y que las varianzas del tiempo son las mismas para las dos marcas.

Solucionario

1.
Datos del problema:

Baterías Duramás: $n_1 = 50$; $\bar{x}_1 = 37$; $\sigma_1 = 5$
Baterías Enerplus: $n_2 = 55$; $\bar{x}_2 = 43$; $\sigma_2 = 5$

En este problema deberemos hacer un contraste de hipótesis sobre la diferencia de medias de las duraciones de las baterías. En este caso conocemos las desviaciones típicas de las poblaciones. Puesto que lo que nos piden es saber si "se ganará en calidad", plantearemos un contraste de hipótesis con las hipótesis siguientes:

Tipos de contraste

Contraste de diferencia de medias en el caso de varianzas conocidas.

1) Expresamos las hipótesis:

- Hipótesis nula. Las medias son iguales:

$$H_0: \quad \mu_1 = \mu_2 \qquad (\mu_1 - \mu_2 = 0)$$

- Hipótesis alternativa. Las medias no son iguales, ya que la media de duración de la batería Enerplus es mayor que la media de la duración de la batería Duramás, de manera que “se gana calidad con el cambio”:

$$H_1: \quad \mu_1 < \mu_2 \qquad (\mu_1 - \mu_2 < 0)$$

2) Fijamos el nivel de significación $\alpha = 0{,}01$.

3) Calculamos el estadístico de contraste, pero antes tendremos que calcular el error estándar:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{25}{50} + \frac{25}{55}} = 0{,}977$$

El estadístico de contraste: $z = \dfrac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \dfrac{37 - 43}{0{,}977} = -6{,}14$

4) Mediante las tablas de la ley normal (0,1) calculamos el *p*-valor correspondiente a este estadístico de contraste. Tenemos que: $p = P(Z < z) = P(Z < -6{,}14\ ) = 0{,}0$.

5) Puesto que $0{,}0 < 0{,}05$, entonces rechazamos la hipótesis nula y llegamos a la conclusión de que con el cambio de baterías ganaremos calidad.

2. Es un problema de diferencia de medias.

a) Debemos contrastar la diferencia de medias para saber si hay una diferencia significativa o podemos considerar que éstas son iguales. A partir de los datos calculamos:

Tipos de contraste

Contraste de diferencias de medias en el caso de varianzas desconocidas.

Media de la muestra A: $\bar{x}_A = \dfrac{1}{n_A} \sum_{i=1}^{n} x_{Ai} = 12{,}73$

Desviación típica de la muestra A: $s_A = \sqrt{\dfrac{1}{n_A - 1} \sum_{i=1}^{n} (\bar{x}_A - x_{Ai})^2} = 2{,}44$

Media de la muestra B: $\bar{x}_B = \dfrac{1}{n_B} \sum_{i=1}^{n} x_{Bi} = 12{,}9$

Desviación típica de la muestra B: $s_B = \sqrt{\dfrac{1}{n_B - 1} \sum_{i=1}^{n} (\bar{x}_B - x_{Bi})^2} = 1{,}42$

A continuación seguiremos el esquema general para hacer un contraste de hipótesis:

1) Expresamos las hipótesis:

- Hipótesis nula: H_0: $\mu_A - \mu_B = 0$
- Hipótesis alternativa: H_1: $\mu_A - \mu_B < 0$

2) Fijamos un nivel de significación: $\alpha = 0{,}05$

3) Estadístico de contraste: $t = \dfrac{(\bar{x}_A - \bar{x}_B)}{s_{\bar{x}_A - \bar{x}_B}} = -0{,}144$

Este estadístico sigue una distribución t de Student, con $n_A + n_B - 2 = 10$ grados de libertad y en la que el error estándar $s_{\bar{x}_A - \bar{x}_B}$ viene dado por la expresión:

$$s_{\bar{x}_A - \bar{x}_B} = s\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 1{,}155, \text{ donde } s = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}} = 1{,}999$$

4) Finalmente calculamos el p-valor: $p = P(t_{n_A + n_{B-2}} < t) = P(t_{10} < -0{,}144) = 0{,}444$.

5) Puesto que $0{,}444 > 0{,}05$, no rechazamos la hipótesis nula en favor de la hipótesis alternativa, de manera que estos datos no confirman que el sistema A sea más rápido que el sistema B.

b) Calcularemos un intervalo de confianza del 90% para la diferencia de medias. Ahora tenemos que $\alpha = 0{,}1$, de manera que $\alpha/2 = 0{,}05$ y el valor crítico correspondiente es $t_{0,05;10} = 1{,}8125$. Por tanto, el intervalo vendrá dado por:

$$(\bar{x}_A - \bar{x}_B) \pm t_{\alpha/2,\, n_A + n_B - 2} \cdot s_{\bar{x}_A - \bar{x}_B}$$

con: $(\bar{x}_A - \bar{x}_B) = -0{,}17$; $\pm t_{0,05;10} = \pm 1{,}8125$; $s_{\bar{x}_A - \bar{x}_B} = 1{,}11$

De manera que obtenemos $-0{,}17 \pm 1{,}8125 \cdot 1{,}155$, y haciendo las operaciones tenemos el intervalo de confianza siguiente para la diferencia de medias:

$$(-2{,}259;\ 1{,}926)$$

En este caso podemos observar que el intervalo de confianza contiene el cero, por lo que podemos concluir con una confianza del 95% que las medias de los tiempos de carga son iguales. Este resultado está totalmente de acuerdo con el obtenido en el contraste de hipótesis del apartado anterior.

Contrastes sobre la diferencia de proporciones

1. Introducción

En esta sesión veremos cómo tenemos que hacer un contraste de hipótesis sobre la diferencia entre dos proporciones y cómo tenemos que determinar un intervalo de confianza con un nivel de significación determinado.

Este hecho puede ser de interés en algunos casos; veamos algunos ejemplos:

- Para saber si hay diferencia entre la proporción de alumnos de la UOC que se conectan por la mañana o los que lo hacen por la noche.

- Para saber si hay diferencia entre la proporción de personas que están a favor de una propuesta y de las que están en contra.

- Para saber si existe diferencia entre la proporción de consumidores que prefieren un producto de un fabricante determinado y los que lo prefieren de la competencia, etc.

Estudiaremos la diferencia de proporciones para saber cómo se distribuyen, determinaremos el error estándar y el estadístico de contraste. Con esto podremos hacer el contraste de hipótesis, así como encontrar intervalos de confianza para la diferencia de proporciones.

2. Contrastes sobre la diferencia de proporciones

Supongamos que tenemos una muestra de tamaño n_1 que proviene de una distribución de Bernoulli de parámetro p_1. La proporción muestral de éxitos es $\hat{p}_1$ y una muestra independiente de la anterior de tamaño n_2 y que proviene de una distribución de Bernoulli de parámetro p_2. La proporción muestral de éxitos es $\hat{p}_2$. Queremos comparar los parámetros poblacionales p_1 y p_2 a partir de las muestras para poder decir si éstos son iguales. Esto lo haremos mediante el contraste de hipótesis:

- Hipótesis nula: H_0: $p_1 = p_2$

- Hipótesis alternativa:

  - Bilateral: H_1: $p_1 \neq p_2$
  - Unilateral: H_1: $p_1 > p_2$
  - Unilateral: H_1: $p_1 < p_2$

Ya sabemos que si el tamño de las muestras es grande (superior a 30), entonces $\hat{P}_1$ y $\hat{P}_2$ presentan distribuciones aproximadamente normales, es decir:

$$\hat{P}_1 \sim N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right) \text{ y } \hat{P}_2 \sim N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right)$$

la diferencia $\hat{P}_1 - \hat{P}_2$ sigue también una distribución normal:

$$\hat{P}_1 - \hat{P}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

Y estandarizándola, obtenemos una nueva variable que sigue una distribución normal tipificada:

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$$

Si se verifica la hipótesis nula, entonces $p_1 = p_2 = p$ y la anterior expresión nos queda así:

$$Z = \frac{(\hat{P}_1 - \hat{P}_2)}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}} = \frac{(\hat{P}_1 - \hat{P}_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

El valor p desconocido que aparece en la expresión se tiene que sustituir por una estimación $\hat{p}$, proporción poblacional común que podemos estimar a partir de la información proporcionada por las dos muestras:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Nota

Hay que tener presente que cuanto mayores sean las muestras, más precisa será la aproximación. Se obtienen resultados muy buenos con muestras de tamaño superior a 100.

Dos formas de calcular la $\hat{p}$

De una encuesta a 1.500 habitantes de Gerona se obtiene que 725 están a favor de que la velocidad sea ilimitada en las autopistas. En otra encuesta realizada a 2.000 habitantes de Barcelona, resulta que 1.050 están a favor de lo mismo. En estos casos tenemos, como resultado de cada muestra, las proporciones siguientes:

$$\hat{p}_1 = \frac{725}{1.500} = 0{,}483\ ;\ \hat{p}_2 = \frac{1.050}{2.000} = 0{,}525$$

Si suponemos que la hipótesis nula es cierta, entonces las proporciones poblacionales de Gerona y Barcelona son iguales, y son iguales a un valor p. Este valor se puede estimar a partir del cociente entre el número total de encuestados que están a favor y el número total de encuestados:

$$\hat{p} = \frac{725 + 1.050}{1.500 + 2.000} = \frac{1.775}{3.500} = 0{,}507$$

Este mismo resultado se puede obtener a partir de las proporciones muestrales, ya que:

- Número de encuestados de Gerona que están a favor:

$$\hat{p}_1 \cdot n_1 = 0{,}483 \cdot 1.500 = 724{,}5 \sim 725$$

- Número de encuestados de Barcelona que están a favor:

$$\hat{p}_2 \cdot n_2 = 0,507 \cdot 2.000 = 1.050$$

Por tanto: $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = 0{,}507$

En resumen, si la hipótesis nula (H_0: $p_1 = p_2$) es cierta, el **estadístico de contraste** que obtenemos y el error estándar son:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\hat{p}_1 - \hat{p}_2}{s_{\hat{p}}}; \quad s_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

y $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ es la **estimación de la proporción poblacional común.**

Este estadístico de contraste es una observación de una ley $N(0,1)$.

Como siempre, una vez calculado el valor del estadístico de contraste, determinaremos el p-valor. Este valor depende de la hipótesis alternativa planteada:

- Si H_1: $p_1 - p_2 \neq 0$, entonces $p = 2\,P(Z > |z|)$
- Si H_1: $p_1 - p_2 < 0$, entonces $p = P(Z < z)$
- Si H_1: $p_1 - p_2 > 0$, entonces $p = P(Z > z)$

¡Atención!

Es muy importante no confundir la p de la proporción con la p del p-valor.

Ejemplo de contraste sobre la diferencia de proporciones

Se quiere construir una central nuclear cerca de un pueblo. Por un lado, la central puede proporcionar puestos de trabajo tanto a gente del pueblo como del resto de la comarca; por el otro, algunas personas del pueblo creen que puede resultar peligrosa para la salud. Se hace una encuesta entre los habitantes del pueblo y los del resto de la comarca. Los resultados son los siguientes: 120 de 200 encuestados del pueblo y 240 de 500 encuestados del resto de la comarca están de acuerdo con su construcción. Haremos un contraste de hipótesis a un nivel del 0,05 para determinar si la proporción de encuestados del pueblo que están a favor de la propuesta es mayor que la proporción de encuestados del resto de la comarca.

Sea p_1 la proporción real de votantes del pueblo y p_2 la de la comarca que están a favor de la propuesta. Ahora debemos hacer una prueba de la diferencia entre dos proporciones:

1. Expresamos las hipótesis:

Hipótesis nula:	H_0:	$p_1 - p_2 = 0$, es decir, $p_1 = p_2$
Hipótesis alternativa:	H_1 :	$p_1 - p_2 > 0$, es decir, $p_1 > p_2$

2. Seleccionamos un nivel de significación: $\alpha = 0{,}05$

3. Calculamos el valor del estadístico de contraste:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{s_{\hat{p}}} = \frac{0{,}60 - 0{,}48}{0{,}0418} = 2{,}87$$

Este valor es una observación de una variable $N(0,1)$. Sin embargo, antes hemos tenido que calcular el error estándar:

$$s_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 0{,}0418 \quad \text{donde: } \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = 0{,}514$$

$$\text{con } \hat{p}_1 = \frac{x_1}{n_1} = \frac{120}{200} = \frac{3}{5} = 0{,}60 \qquad \text{y} \qquad \hat{p}_2 = \frac{x_2}{n_2} = \frac{240}{500} = \frac{12}{25} = 0{,}48$$

4. Calculamos el *p*-valor: $p = P(Z > 2{,}87) = 0{,}0021$.

5. Conclusión: puesto que $0{,}0021 < 0{,}05$, entonces rechazamos H_0 y estamos de acuerdo con el hecho de que la proporción de votantes del pueblo a favor de la propuesta es superior que la proporción de votantes de la comarca.

3. Intervalos de confianza para la diferencia de proporciones

El procedimiento que hay que seguir para construir un intervalo de confianza para la diferencia de proporciones es análogo al que se sigue para la diferencia de medias y, en general, para cualquier tipo de intervalo de confianza:

1) Fijamos el nivel de confianza, que escribimos como $(1 - \alpha)$.

2) Calculamos el error estándar de la diferencia de proporciones.

3) Calculamos el valor crítico teniendo en cuenta el tipo de distribución de nuestro estadístico de contraste. En este caso, una normal (0,1).

4) Calculamos el margen de error a partir del valor crítico y el error estándar.

El **intervalo de confianza** para la diferencia de proporciones viene dado por:

$$(\hat{p}_1 - \hat{p}_2) \pm (\text{valor crítico})\ (\text{error estándar})$$

es decir, $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} s_{\hat{p}}$, donde:

$$s_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \text{ y } \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Ejemplo de cálculo de los intervalos de confianza para la diferencia de proporciones

Considerando de nuevo el ejemplo de la central nuclear, calcularemos el intervalo de confianza al 95%.

$$\hat{p}_1 - \hat{p}_2 = 0{,}60 - 0{,}48 = 0{,}12; \pm z_{\alpha/2} = \pm 1{,}96; s_{\hat{p}} = 0{,}0418$$

Por tanto, tenemos: $0{,}12 \pm 1{,}96 \cdot 0{,}0418$.

Así pues, el intervalo de confianza es (0,0380; 0,2019).

Dado que estamos tratando con proporciones que se pueden expresar en tanto por ciento, también es posible expresar el intervalo de confianza en tanto por ciento: el valor de la proporción está entre el 3,80% y el 20,19%.

Podemos observar que el cero no se incluye dentro de este intervalo. Este resultado está de acuerdo con el obtenido en el contraste de hipótesis efectuado anteriormente, en el que se ha rechazado la hipótesis nula, según la cual la diferencia de proporciones es igual a cero.

4. Resumen

En esta sesión hemos aprendido a hacer contrastes de hipótesis para la diferencia de dos proporciones en el caso de muestras grandes. Después hemos visto cuál es el procedimiento para construir intervalos de confianza para la diferencia de proporciones.

Ejercicios

1. Una firma manufacturera de cigarrillos distribuye dos marcas. De una muestra de 150 fumadores, 29 prefieren la marca A, y de otra muestra de 200 fumadores, 56 prefieren la marca B. A partir de estos datos, ¿podemos concluir que los fumadores prefieren más una marca que otra? Utilizad un nivel de significación $\alpha = 0{,}1$.

2. Se realiza un estudio para determinar la efectividad de una nueva vacuna contra la gripe. Se administra la vacuna a una muestra aleatoria de 2.000 personas y de este grupo, 23 sufren la enfermedad. Como grupo de control se seleccionan al azar 2.500 personas que no han sido vacunadas. De este grupo, 98 padecen la gripe. Construid un intervalo de confianza del 95% para la diferencia de proporciones. ¿Qué podéis decir de la efectividad de la nueva vacuna?

Solucionario

1.
Sean p_1 y p_2 las proporciones reales de consumidores de la marca A y B, respectivamente. Ahora tenemos que hacer una prueba de la diferencia entre dos proporciones:

1) Expresamos las hipótesis:

– Hipótesis nula: H_0: $p_1 - p_2 = 0$
– Hipótesis alternativa: H_1: $p_1 - p_2 \neq 0$

2) Determinamos un nivel de significación: $\alpha = 0{,}1$

3) Estadístico de contraste:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})[(1/n_1)+(1/n_2)]}} = -1{,}871 \text{ donde: } \hat{p} = \frac{29+56}{150+200} = 0{,}243$$

$$\hat{p}_1 = \frac{29}{150} = 0{,}193 \text{ y } \hat{p}_2 = \frac{56}{200} = \frac{7}{25} = 0{,}280$$

es una observación de una variable $N(0,1)$.

4) Calculamos el p-valor: $p = 2P(Z > |z|) = 2P(Z > 1{,}871) = 2 \cdot 0{,}031 = 0{,}062$.

5) Conclusión: puesto que $0{,}061 < 0{,}1$, rechazamos H_0 y, por tanto, sí que hay diferencia en las preferencias de los consumidores.

2.
Datos:

– Personas vacunadas: $n_1 = 2.000$, $x_1 = 23$
– Personas no vacunadas: $n_2 = 2.500$, $x_2 = 98$

$$\hat{p}_1 = \frac{23}{2.000} = 0{,}0115, \quad \hat{p}_2 = \frac{98}{2.500} = 0{,}0392$$

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{2.000 \quad 0{,}0115 + 2.500 \quad 0{,}0392}{2.000 + 2.500} = 0{,}0269$$

$$s_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0{,}0269(1-0{,}0296)\left(\frac{1}{2.000} + \frac{1}{2.500}\right)} = 0{,}0049$$

Ya podemos calcular el intervalo de confianza: $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}s_{\hat{p}}$

$$(0{,}0115 - 0{,}0392) \pm 1{,}96 \cdot 0{,}0049; \; -0{,}0277 \pm 0{,}0096$$

Por tanto, el intervalo de confianza es: (–0,0373; –0,0181).

Puesto que el valor cero no se encuentra dentro del intervalo, llegamos a la conclusión de que las proporciones son diferentes y, dado que el intervalo es negativo, la vacuna tiene realmente algún efecto beneficioso.