Método general de la razón de verosimilitud

- La prueba anterior fue un ejemplo específico de aplicación del Método de Razón de Verosimilitud:
- En general sea θ un vector que consiste en todos los parámetros desconocidos de la población, y $L(\theta)$ la función de verosimilitud obtenida evaluando la densidad conjunta de $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ en sus valores observados $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$.
- θ toma sus valores en el conjunto de parámetros $\boldsymbol{\Theta}$. Por ejemplo, en el caso normal p-dimensional $\theta = [\mu_1, \ldots, \mu_p; s_{11}, \ldots, s_{1p}; s_{21}, \ldots, s_{2p}; \ldots; s_{p1}, \ldots, s_{pp}]$.
- Entonces $\boldsymbol{\Theta}$ consiste del espacio p-dimensional de las μ_i, $-\infty \leq \mu_i \leq \infty$, $i = 1, \ldots, p$, combinado con el espacio $p(p+1)/2$-dimensional de las varianzas y covarianzas tal que $\boldsymbol{\Sigma}$ es positivo definido. Po lo tanto, $\boldsymbol{\Theta}$ tiene una dimensión $v = p + p(p+1)/2$.

Método general de la razón de verosimilitud

- Bajo la hipótesis nula $H_0 : \theta = \theta_{\mathbf{0}}$, θ está restringido a estar en algun subconjunto $\mathbf{\Theta_0} \subseteq \mathbf{\Theta}$.
- Por ejemplo, en el caso normal p-dimensional con $\mu = \mu_{\mathbf{0}}$ y $\mathbf{\Sigma}$ no especificada, tenemos que

$$\mathbf{\Theta_0} = [\mu_1 = \mu_{1_0}, .., \mu_p = \mu_{p_0}; s_{11}, .., s_{1p}; s_{21}, .., s_{2p}; ...; s_{p1}, .., s_{pp}, \text{ con } \mathbf{\Sigma} \text{ p}$$

- Así que $\mathbf{\Theta_0}$ tiene dimensión $v_0 = 0 + p(p+1)/2 = p(p+1)/2$
- Entonces una prueba de razón de verosimilitud de $H_0 : \theta \in \mathbf{\Theta_0}$ rechaza H_0 en favor de $H_1 : \theta \ni \theta_{\mathbf{0}}$ si

$$\Lambda = \frac{\max_{\theta \in \mathbf{\Theta_0}} L(\mathbf{\Theta})}{\max_{\theta \in \mathbf{\Theta}} L(\Theta)} < c$$

- Para una constante c escogida adecuadamente (algún percentil de la distribución de Λ).

Método general de la razón de verosimilitud

- En cada aplicación del método de la razón de verosimilitud, se debe obtener la distribucion muestral del estadístico de la razon de verosimilitud.
- Entonce c puede ser seleccionado para producir una prueba con un específico α
- Sin embargo, cuando n es suficientemente grande, la distribución

$$-2ln(\Lambda) = -2ln\left(\frac{\max_{\theta \in \mathbf{\Theta_0}} L(\mathbf{\Theta})}{\max_{\theta \in \mathbf{\Theta}} L(\Theta)}\right)$$

  bajo la hipótesis nula $H_0 : \theta = \mathbf{\Theta_0}$ sigue aproximadamente una $\chi^2_{v-v_0}$.
- donde v es el número de parámetros que se estiman sin considerar restricciones y v_0 el número de parámetros que se estiman cuando algunos de los parámetros son fijados

Comentarios sobre método de la razón de verosimilitud

- La aproximación de la razón de verosimilitud a una χ^2 cuando n es grande nos permite obtener de manera directa las regiones de rechazo y aceptación en prueba de hipótesis, de acuerdo a los percentiles de la χ^2
- Evitando así el cálculo de sus valores críticos partir de los valores criticos de una T^2 que a su vez obtiene los valores criticos a partir de una F
- Entre todas las pruebas estadísticas para hacer inferencia sobre los parámetros de la población, las pruebas de la razón de verosimilitud son la que tiene mayor potencia, cuando la muestra es grande.

Regiones de confianza para el vector de medias μ

- Cuando H_0 no se rechaza, concluimos que μ_0 es un posible valor para la media de la poblacion normal. Sin embargo, existen otros conjuntos de valores de μ que también pueden ser consistentes con los datos.
- Un intervalo de confianza del $100(1-\alpha)\%$ para la media μ tiene la forma

$$P(\overline{x} - t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}) = 1-\alpha$$

- Un intervalo de confianza univariado del $100(1-\alpha)\%$ para un parámetro θ de la población se define de manera general como

$$P(\widehat{\theta} - \varepsilon \leq \theta \leq \widehat{\theta} + \varepsilon) = P(|\widehat{\theta} - \theta| \leq \varepsilon) = 1-\alpha$$

donde $\widehat{\theta}$ es el estimador del parámetro y ε mide la desviación del parámetro respecto a su estimador

Regiones de confianza para el vector de medias μ

- El concepto de intervalo de confianza univariado se puede extender a un espacio multivariado p-dimensional, que son las **regiones de confianza multivariadas**.
- Sea θ un vector de parámetros desconocidos de una poblacion y sea $\mathbf{\Theta}$ el conjunto de todos los posibles valores de θ.
- Una **región de confianza** que llamaremos $R(\boldsymbol{X})$ es una región de probables valores de θ, y se determina a partir de la muestra de datos $\boldsymbol{X} = [\boldsymbol{x_1}, ..., \boldsymbol{x_1}]$.
- La región $R(\boldsymbol{X})$ es una región de confianza del $100(1-\alpha)\%$, si antes de que la muestra sea seleccionada

$$P(R(\boldsymbol{X}) \text{ cubra al verdadero } \theta) = 1-\alpha$$

esta probabilidad es calculada bajo el verdadero pero desconocido valor de θ

Regiones de confianza para el vector de medias μ

- Recordando que el estadístico para probar $H_0 : \mu = \mu_{\mathbf{0}}$ está dado por
$$T^2 = n(\overline{\mathbf{x}} - \mu_{\mathbf{0}})' \mathbf{S}^{-\mathbf{1}} (\overline{\mathbf{x}} - \mu_{\mathbf{0}})$$
- No se rechazaba H_0 si $T^2 \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$, es decir si
$$n(\overline{\mathbf{x}} - \mu_{\mathbf{0}})' \mathbf{S}^{-\mathbf{1}} (\overline{\mathbf{x}} - \mu_{\mathbf{0}}) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$$
- Por tanto la región de confianza para μ de una población normal p-variada está dado por
$$P\left(n(\overline{\mathbf{x}} - \mu)' \mathbf{S}^{-\mathbf{1}} (\overline{\mathbf{x}} - \mu) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha) \right) = 1 - \alpha$$

Regiones de confianza para el vector de medias μ

- Más formalmente, una región de confianza $R(\boldsymbol{X})$ del $100(1-\alpha)\%$ para el vector de medias μ de una distribución normal p-dimensional es el elipsoide determinado por todos los puntos posibles μ que satisfacen

$$n(\overline{\boldsymbol{x}}-\mu)^{'}\mathbf{S}^{-\mathbf{1}}(\overline{\boldsymbol{x}}-\mu) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$$

donde

$$\overline{\mathrm{x}} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{x}_j \text{ y } \boldsymbol{S} = \frac{1}{n-1}\sum_{j=1}^{n} (\mathrm{x}_j - \overline{\mathrm{x}})(\mathrm{x}_j - \overline{\mathrm{x}})^{'}$$

y $\boldsymbol{x_1}, \ldots \boldsymbol{x_n}$ son las muestras de observaciones

Ejemplo: Construcción de una region de confianza del $100(1-\alpha)\%$ para el vector de medias

Utilizar la muestra original de datos bivariados para construir una región bidimensional de confianza del 90% y determine si el punto $(4.0, -1.5)$ se encuentra en esta región

x_{j1}	x_{j2}
1.43	-0.69
1.62	-5.00
2.46	-1.13
2.48	-5.20
2.97	-6.39
4.03	2.87
4.47	-7.88
5.76	-3.97
6.61	2.32
6.68	-3.24
6.79	-3.56
7.46	1.61
7.88	-1.87
8.92	-6.60
9.42	-7.64

Resumen de los estadísticos

$$\bar{x} = \begin{bmatrix} 5.26 \\ -3.09 \end{bmatrix}, \mu_0 = \begin{bmatrix} 4.0 \\ -1.5 \end{bmatrix}$$

$$S = \begin{bmatrix} 7.12 & -.72 \\ -.72 & 12.43 \end{bmatrix}, S^{-1} = \begin{bmatrix} .1412 & .0082 \\ .0082 & .0809 \end{bmatrix}$$

y

$$\frac{(n-1)p}{n-p} F_{p,n-p}(\alpha) = 5.95$$

Entonces una región de confianza 100(1 - α)% para el vector de medias $\underset{\sim}{\mu}$ es el elipsoide determinado por todos los posibles puntos $\underset{\sim}{\mu}$ que satisfacen

$$n\left(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}\right)' \underset{\sim}{S}^{-1}\left(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}\right) \leq \frac{(n-1)\,p}{(n-p)} F_{p,n-p}(\alpha) \text{ i.e.,}$$

$$15\left(\begin{bmatrix} 5.26 \\ -3.09 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)' \begin{bmatrix} 0.1412 & 0.0082 \\ 0.0082 & 0.0809 \end{bmatrix} \left(\begin{bmatrix} 5.26 \\ -3.09 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right) \leq 5.95$$

Queremos determinar si

$$\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 4.0 \\ -1.5 \end{bmatrix}$$

cae dentro de esta región de confianza del 90%

Tenemos que

$$15\left(\begin{bmatrix} 5.26 \\ -3.09 \end{bmatrix} - \begin{bmatrix} 4.0 \\ -1.5 \end{bmatrix}\right)' \begin{pmatrix} 0.1412 & 0.0082 \\ 0.0082 & 0.0809 \end{pmatrix} \left(\begin{bmatrix} 5.26 \\ -3.09 \end{bmatrix} - \begin{bmatrix} 4.0 \\ -1.5 \end{bmatrix}\right)$$

$$= 5.97 > 5.95$$

No cae dentro de la región de confianza del 90%!!!

¿Está el punto (-1.5, 4.0) en esta región.

En este caso

$$15\left(\begin{bmatrix} 5.26 \\ -3.09 \end{bmatrix} - \begin{bmatrix} -1.5 \\ 4.0 \end{bmatrix}\right)' \begin{pmatrix} 0.1412 & 0.0082 \\ 0.0082 & 0.0809 \end{pmatrix} \left(\begin{bmatrix} 5.26 \\ -3.09 \end{bmatrix} - \begin{bmatrix} -1.5 \\ 4.0 \end{bmatrix}\right)$$

$$= 146.201 > 5.95$$

Este punto cae muy afuera de la región de confianza del 90%!

Podemos calcular los ejes del elipsoide de confianza, que representa la región de confianza del 100 (1 - α)% para el vector de medias $\boldsymbol{\mu}$. Estos son determinados por los valores y vectores propios de **S**. La longitud de los ejes del elipsoide de confianza

$$n\left(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}\right)' \underset{\sim}{S}^{-1}\left(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}\right) \leq c^2 = \frac{(n-1)\,p}{(n-p)} F_{p,n-p}(\alpha)$$

Son determinados por

$$\frac{\sqrt{\lambda_i}\,c}{\sqrt{n}} = \sqrt{\lambda_i}\sqrt{\frac{(n-1)\,p}{n\,(n-p)} F_{p,n-p}(\alpha)}$$

Y sus direcciones están definidas por los vectores propios correspondientes $\mathbf{e}_i$

Comenzando en el centroide $\bar{x}$, los ejes del elipsoide de confianza son

$$\pm\sqrt{\lambda_i}\sqrt{\frac{(n-1)\,p}{n\,(n-p)}\,F_{p,n-p}(\alpha)}\,\underset{\sim}{e}_i$$

Obsérvese que los cocientes entres los λ_is ayudan a identificar que tan grande es un eje respecto al otro

Entonces para obtener los ejes del elipsoide de confianza del 100(1 - α)% del vector de medias $\boldsymbol{\mu}$ – calculamos los pares de valores y vectores propios λ_i, e_i de la matriz de covarianza muestral **S**.

$$\lambda_1 = 12.522,\ \underset{\sim}{e}_1 = \begin{bmatrix} -0.13316 \\ 0.99109 \end{bmatrix}$$

$$\lambda_2 = 7.024,\ \underset{\sim}{e}_2 = \begin{bmatrix} 0.99109 \\ 0.13316 \end{bmatrix}$$

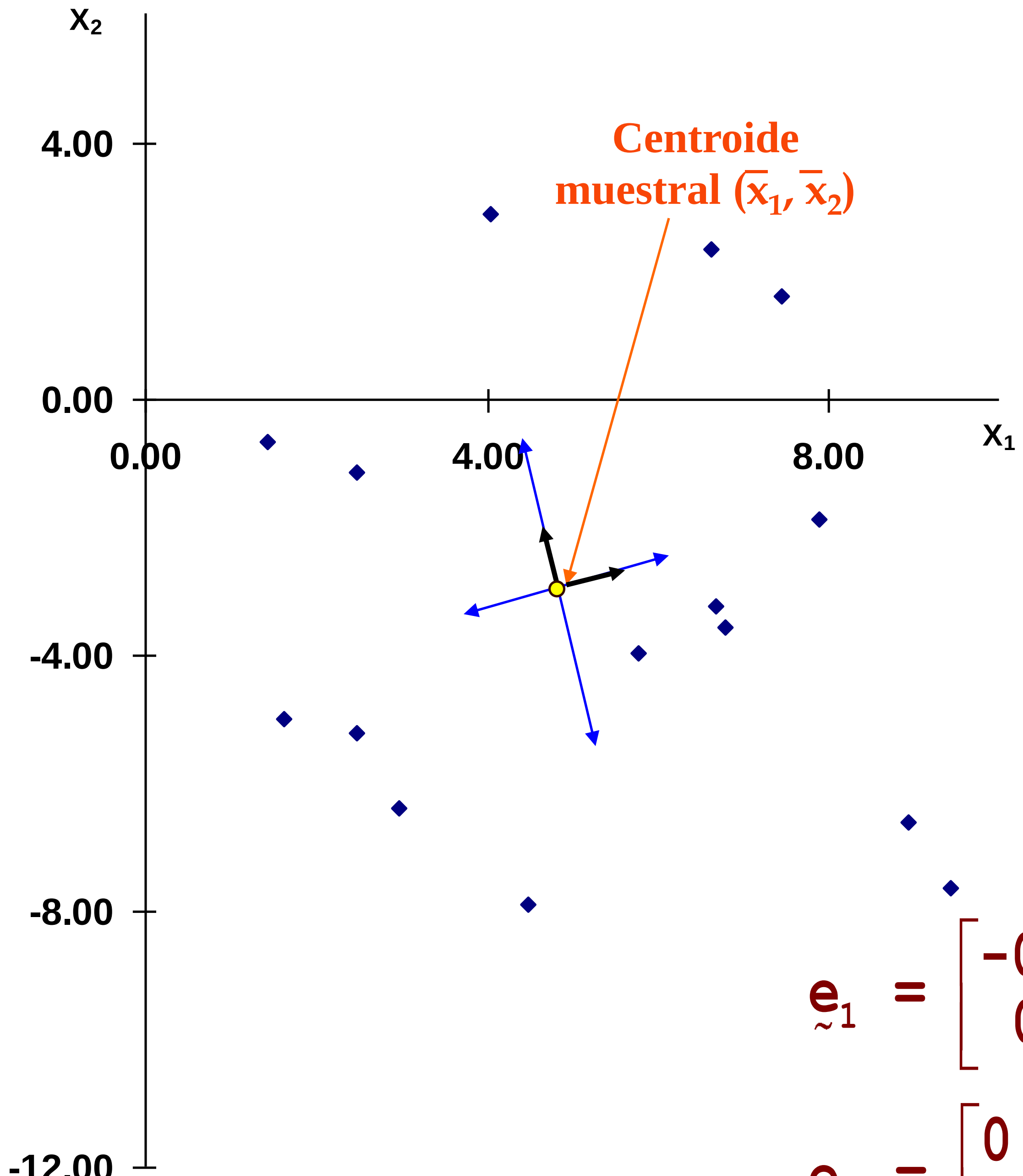
así que las mitades de las longitudes de los ejes mayor y menor están dadas por

$$\sqrt{\lambda_1}\sqrt{\frac{(n-1)p}{n(n-p)}F_{p,n-p}(\alpha)} = \sqrt{12.522}\sqrt{\frac{(15-1)2}{15(15-2)}5.95} = 3.7293$$

$$\sqrt{\lambda_2}\sqrt{\frac{(n-1)p}{n(n-p)}F_{p,n-p}(\alpha)} = \sqrt{7.024}\sqrt{\frac{(15-1)2}{15(15-2)}5.95} = 2.7931$$

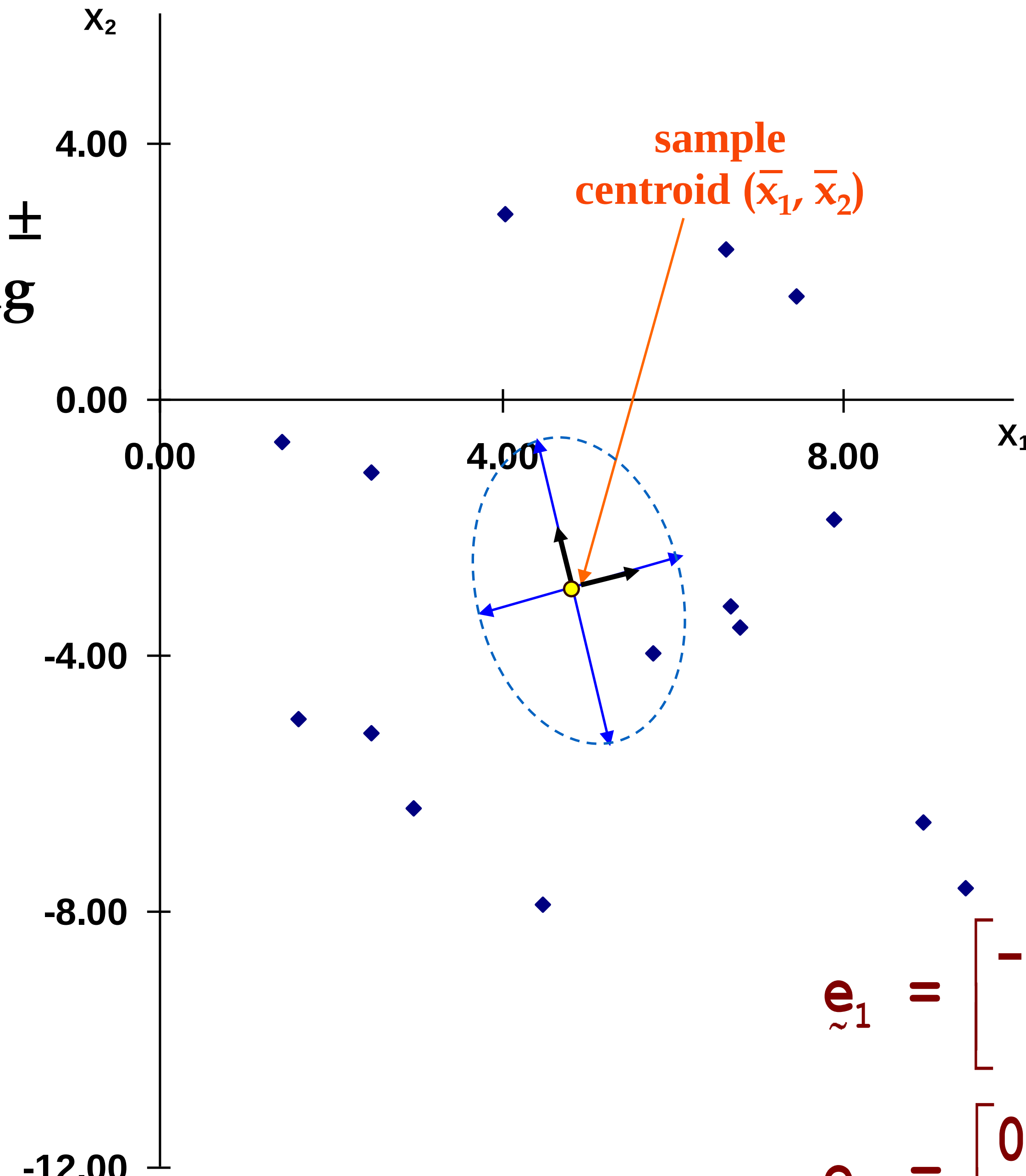
Los ejes descansan a lo largo de los correspondientes vectores propios e_i:

$$\underset{\sim}{e}_1 = \begin{bmatrix} -0.13316 \\ 0.99109 \end{bmatrix}, \quad \underset{\sim}{e}_2 = \begin{bmatrix} 0.99109 \\ 0.13316 \end{bmatrix}$$



$$\underset{\sim}{e}_1 = \begin{bmatrix} -0.13316 \\ 0.99109 \end{bmatrix}, \qquad \underset{\sim}{e}_2 = \begin{bmatrix} 0.99109 \\ 0.13316 \end{bmatrix}$$

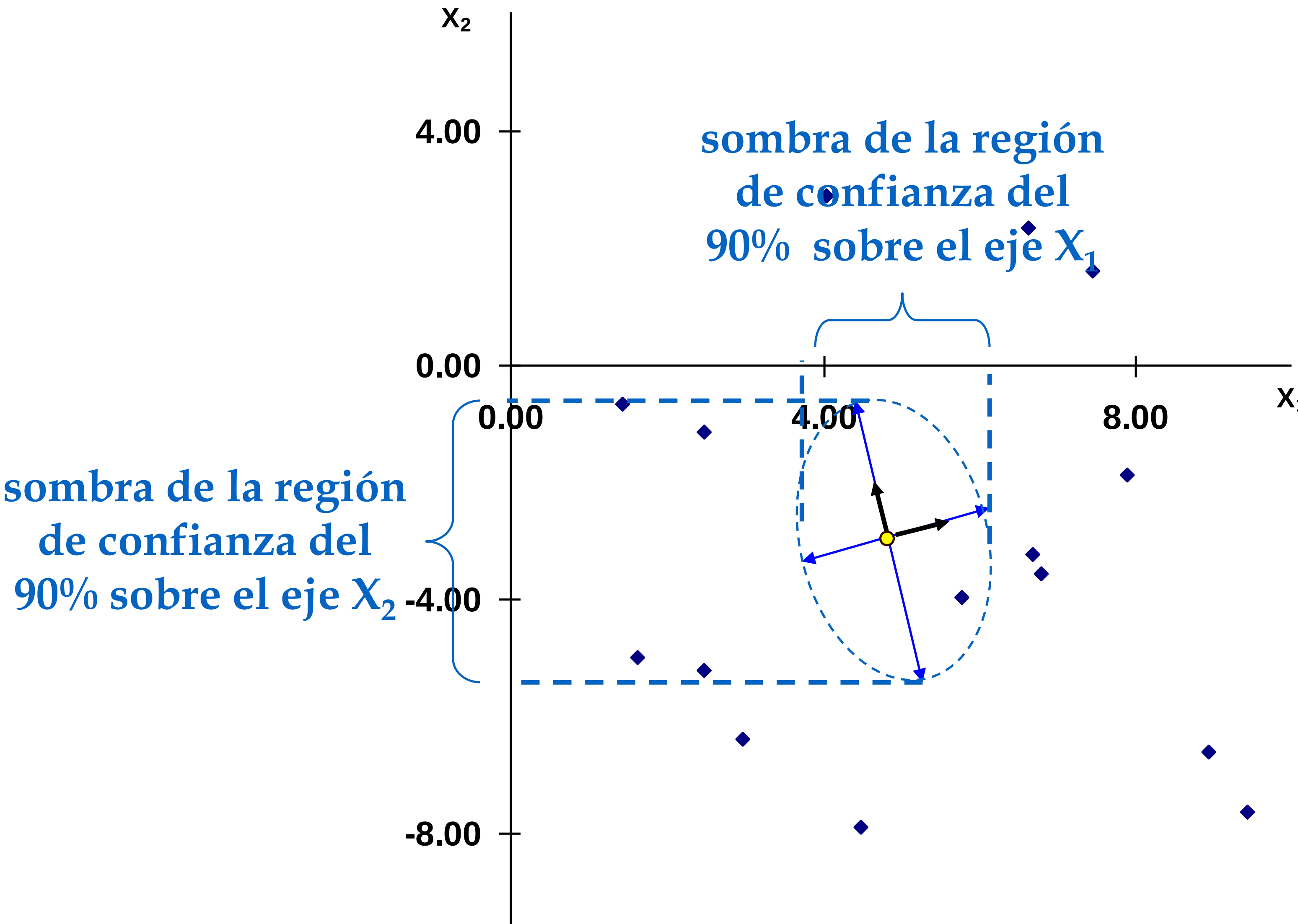
Now we move ± 3.73 units along the vector $\underset{\sim}{e}_1$



and ± 2.793 units along the vector $\underset{\sim}{e}_2$.

$$\underset{\sim}{e}_1 = \begin{bmatrix} -0.13316 \\ 0.99109 \end{bmatrix},$$

$$\underset{\sim}{e}_2 = \begin{bmatrix} 0.99109 \\ 0.13316 \end{bmatrix}$$



También podemos proyectar la elipse sobre los ejes para formar intervalos de confianza simultáneos.