Intervalo de confianza simétrico frecuentista para la media de n gaußianas

20 de mayo de 2019

El problema Supongamos que hay una cantidad q que deseamos medir. Por ejemplo q podría ser el número de autos que habían en circulación en Argentina el día 7 de mayo de 2019 a las 11:37 horas. Por nombrar otro ejemplo, q podría ser la edad del planeta Tierra. Lo importante es que q es un número que no es aleatorio, está bien definido y tiene un valor. Sólo que no lo conocemos.

Para intentar conocer a q realizamos n mediciones. Cada una de estas mediciones nos arroja un resultado de la forma $q_i \pm \sigma_i$ con $i \in \{1, \dots, n\}$. La pregunta es, dadas estas n mediciones ¿cuánto vale q?

Modelado de las mediciones Con el fin de estimar el valor de q generamos un conjunto de $n \in \mathbb{N}$ pares de datos (q_i, σ_i) tales que $q_i \in (q_1, \dots, q_n)$ y $\sigma_i \in (\sigma_1, \dots, \sigma_n)$ que representan, para cada $i \in \{1, \dots, n\}$, una realización q_i de una variable aleatoria $Q_i \sim \text{Gauß}\,(\mu = q, \sigma = \sigma_i)$. Es decir que asumimos que:

- La tupla $(q_1, \dots, q_n)$ es una realización de la "tupla aleatoria" $(Q_1, \dots, Q_n)$.
- Q_i es una variable aleatoria con distribución gaußiana

$$Q_i \sim \text{Gauß}\,(\mu = q, \sigma = \sigma_i)\,.$$

  Algunos comentarios sobre esta variable aleatoria:

  - La aleatoriedad tiene origen en nuestra incapacidad de realizar mediciones perfectas. Dijimos que q es un número no aleatorio. La aleatoriedad se introduce a la hora de medir.
  - La media de cada estimación se asumió que es $\mathbb{E}\,[Q_i] = q$ independiente de i (i.e. es igual para todas las mediciones). Esto refleja el hecho de que nuestros instrumentos de medición no tienen sesgo[1], sólo tienen fluctuaciones aleatorias.
  - La varianza de cada estimación es $\mathbb{V}\,[Q_i] = \sigma_i^2$ que sí depende de i. Esta dependencia podría originarse en el hecho de que "algunas mediciones fueron más complicadas que otras y entonces fueron de peor calidad"[2].
  - Hemos elegido una distribución gaußiana para Q_i, ¿por qué? Buena pregunta. La cuestión es que en el problema particular que analicé en su momento los datos q_i eran resultados de un instrumento de datación de piedras (o algo así) que decía que se asuma una distribución gaußiana. Probablemente promediaba internamente muchas mediciones y entonces, por Teorema Central del Límite, dicho promedio es gaußiano.

Box 1 - Ejemplo: Medición de temperatura ambiente

La cantidad q podría ser la temperatura ambiente. Queremos medir la temperatura ambiente. Supongamos que tenemos tres termómetros y cada uno de ellos nos arroja los siguientes resultados:

$$T = \begin{cases} (23 \pm 1)\ ^\circ\text{C} & \text{termómetro 1} \\ (22 \pm 3)\ ^\circ\text{C} & \text{termómetro 2} \\ (22 \pm 1)\ ^\circ\text{C} & \text{termómetro 3} \end{cases}.$$

En este caso $(q_i) = (23, 22, 22)\ ^\circ\text{C}$ y $(\sigma_i) = (1, 3, 1)\ ^\circ\text{C}$.

[1] O si lo tienen, es un sesgo constante y asumimos que lo hemos corregido. Si los instrumentos tienen sesgo que varía con el tiempo ya no se puede asumir que $\mathbb{E}\,[Q_i]$ es independiente de i.

[2] Véase que no siempre $\sigma_i \neq \sigma_j$. Supongamos que queremos medir el diámetro de un vaso con una regla. El diámetro del vaso es una cantidad q que es un número fijo, así que el método se puede aplicar. Le pedimos a n personas que lo midan, cada uno nos va a dar un q_i distinto. Sin embargo el σ_i corresponde a la graduación de la regla, i.e. $\sigma_i = 1$ mm. En este caso "todas las mediciones fueron igual de difíciles" y entones todas "tienen la misma calidad".

La verosimilitud de la realización $(q_1, \ldots, q_n)$ es, por definición,

$$\begin{aligned} L &= f_{Q_1,\ldots,Q_n}(q_1, \ldots, q_n) \\ \text{Asumimos } Q_i \text{ independientes} \rightarrow &= \prod_{i=1}^{n} f_{Q_i}(q_i) \\ \text{Son gaußianas} \rightarrow &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2}\left(\frac{q_i - q}{\sigma_i}\right)^2\right) \end{aligned}$$

donde $f_{Q_1,\ldots,Q_n}$ es la función de densidad conjunta de las Q_i y f_{Q_i} es la función de densidad de Q_i.

Un estimador puntual para la cantidad q A partir de esta verosimilitud se puede obtener el estimador de máxima verosimilitud[3]

$$\hat{q} = \frac{\sum \frac{Q_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}. \rightarrow \text{MLE}$$

(MLE es *Maximum Likelihood Estimator*.) La notación $\hat{q}$ con el sombrerito es estándar para denotar que $\hat{q}$ es el estimador MLE del parámetro q. Aquí q es el parámetro de la distribución de las Q_i, que casualmente coincide con la cantidad que queremos medir. Si quisiéramos calcular el valor numérico de $\hat{q}$ en función de nuestros datos simplemente deberíamos reemplazar las variables aleatorias por las observaciones, i.e. $Q_i \to q_i$, y entonces obtenemos

$$\hat{q}_{\text{observado}} = \frac{\sum \frac{q_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}.$$

Este valor de $\hat{q}_{\text{observado}}$ es el que nos conviene elegir tal que la probabilidad de nuestra observación sea máxima. Lo malo es que no nos da un intervalo con un error $\pm\sigma$, es sólo un número.

Un intervalo para q El estimador $\hat{q}$ nos dice cuál es el valor de q que maximiza la probabilidad de nuestra observación. Sin embargo, no nos da un intervalo de la forma $q \pm \sigma$. Para obtener un intervalo podemos proceder mediante estadística bayesiana o mediante estadística frecuentista. Los pros y contras de cada tipo de estadística se pueden googlear. Los resultados obtenidos mediante cada tipo de estadística en general no son comparables entre sí aunque uno espera que den parecidos. Yo voy a proceder con la estadística frecuentista, que es menos intuitiva, pero no tiene que hacer suposiciones *ad hoc*.

La forma de obtener un intervalo $q \pm \sigma$ en la estadística frecuentista es mediante la construcción de un *intervalo de confianza frecuentista*. Un intervalo frecuentista[4] es un "algoritmo" que como entrada recibe:

1. La realización $(q_1, \ldots, q_n)$. (O sea, las mediciones.)
2. Los parámetros $(\sigma_1, \ldots, \sigma_n)$. (O sea, los errores de las mediciones.)
3. Un parámetro adicional $CL \in (0, 1)$ que es el nivel de confianza (*confidence level* en Google). El nivel de confianza es, por definición,
$$CL \stackrel{\text{def}}{=} \min_q \mathbb{P}(q \in \text{intervalo de confianza}).$$
E.g. si $CL = 68\ \%$ entonces al menos el 68 % de los intervalos construidos con este algoritmo van a contener al q verdadero.

El resultado de este algoritmo son dos números que forman el afamado intervalo:

$$I = [q_{\text{mín}}, q_{\text{máx}}]. \rightarrow \text{Intervalo de confianza}$$

Por supuesto que este intervalo puede expresarse en la forma clásica $q \pm \sigma$ mediante una simple transformación:

$$I = \frac{q_{\text{mín}} + q_{\text{máx}}}{2} \pm \frac{q_{\text{máx}} - q_{\text{mín}}}{2}.$$

Hay varios métodos que permiten relacionar a $q_{\text{mín}}$ y $q_{\text{máx}}$ con los datos q_i y σ_i. El método que voy a usar acá es el del "cinturón de Neyman" (*Neyman belt* en Google). La construcción del cinturón de Neyman para un "intervalo de confianza simétrico[5]" consiste en hallar $\hat{q}_{\text{mín}}$ y $\hat{q}_{\text{máx}}$ (notar que $\hat{q}_{\text{mín}} \neq q_{\text{mín}}$ y lo mismo para $\hat{q}_{\text{máx}}$) tales que

$$\text{Construcción} \rightarrow \begin{cases} \mathbb{P}(\hat{q} < \hat{q}_{\text{mín}} | q = q_0) = \dfrac{1 - \text{CL}}{2} \\ \mathbb{P}(\hat{q} > \hat{q}_{\text{máx}} | q = q_0) = \dfrac{1 - \text{CL}}{2} \end{cases}$$

[3] Es un procedimiento puramente matemático que consiste en estimar q como aquel que maximiza la probabilidad de lo observado. Lo observado es la realización $\{q_i\}$. El resultado de este procedimiento es el estimador de máxima verosimilitud.

[4] *Intervalo frecuentista* es lo mismo que *intervalo de confianza frecuentista*.

[5] Además hay distintos tipos de intervalos. El de tipo *simétrico* es el que nos da un $q \pm \sigma$. Otros posibles intervalos son e.g. una cota superior o una cota inferior.

donde q_0 es un valor que se asume conocido (aunque en verdad no se conoce). Lo anterior es el "modo construcción" del cinturón de confianza. Para obtener un intervalo de confianza debemos "invertirlo" y expresarlo en "modo uso". Esto es

$$\text{Utilización} \rightarrow \begin{cases} \mathbb{P}\left(\hat{q} < \hat{q}_{\text{observado}} | q = q_{\text{máx}}\right) = \dfrac{1 - CL}{2} \\ \mathbb{P}\left(\hat{q} > \hat{q}_{\text{observado}} | q = q_{\text{mín}}\right) = \dfrac{1 - CL}{2} \end{cases}$$

donde $\hat{q}_{\text{observado}}$ es el que se encontró previamente. De aquí podemos despejar $q_{\text{mín}}$ y $q_{\text{máx}}$ para obtener el tan preciado intervalo. Los detalles de por qué esto es así y por qué esto funciona exceden el alcance de este breve documento. Lo importante es que el problema ya "está resuelto", ahora sólo hay que despejar $q_{\text{máx}}$ y $q_{\text{mín}}$ en términos de los datos y hacer la "cuentita". Procedamos, la probabilidad de la primera ecuación se puede desarrollar del siguiente modo:

$$\begin{aligned} \mathbb{P}\left(\hat{q} < \hat{q}_{\text{observado}} | q = q_{\text{máx}}\right) &= \mathbb{P}\left(\frac{\sum \frac{Q_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}} < \hat{q}_{\text{observado}} | q = q_{\text{máx}}\right) \\ &= \mathbb{P}\left(\sum \frac{Q_i}{\sigma_i^2} < \hat{q}_{\text{observado}} \sum \frac{1}{\sigma_i^2} | q = q_{\text{máx}}\right). \end{aligned}$$

Recordemos ahora que $Q_i \sim \text{Gauß}\left(\mu = q, \sigma = \sigma_i\right)$ por hipótesis a la hora de modelar nuestras mediciones, por lo tanto

$$\frac{Q_i}{\sigma_i^2} \sim \frac{\text{Gauß}\left(\mu = q, \sigma = \sigma_i\right)}{\sigma_i^2} \equiv \text{Gauß}\left(\mu = \frac{q}{\sigma_i^2}, \sigma = \frac{1}{\sigma_i}\right).$$

Esto implica que

$$\sum_{i=1}^{n} \frac{Q_i}{\sigma_i^2} \sim \sum_{i=1}^{n} \text{Gauß}\left(\mu = \frac{q}{\sigma_i^2}, \sigma = \frac{1}{\sigma_i}\right) \equiv \text{Gauß}\left(\mu = q \sum_{i=1}^{n} \frac{1}{\sigma_i^2}, \sigma = \sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}\right).$$

Entonces

$$\mathbb{P}\left(\hat{q} < \hat{q}_{\text{observado}} | q = q_{\text{máx}}\right) = \mathbb{P}\left(\text{Gauß}\left(\mu = q_{\text{máx}} \sum_{i=1}^{n} \frac{1}{\sigma_i^2}, \sigma = \sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}\right) < \hat{q}_{\text{observado}} \sum \frac{1}{\sigma_i^2}\right).$$

De manera completamente análoga la probabilidad de la segunda ecuación se puede reescribir según

$$\begin{aligned} \mathbb{P}\left(\hat{q} > \hat{q}_{\text{observado}} | q = q_{\text{mín}}\right) &= \mathbb{P}\left(\text{Gauß}\left(\mu = q_{\text{mín}} \sum_{i=1}^{n} \frac{1}{\sigma_i^2}, \sigma = \sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}\right) > \hat{q}_{\text{observado}} \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\right) \\ &= 1 - \mathbb{P}\left(\text{Gauß}\left(\mu = q_{\text{mín}} \sum_{i=1}^{n} \frac{1}{\sigma_i^2}, \sigma = \sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}\right) \leq \hat{q}_{\text{observado}} \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\right). \end{aligned}$$

Entonces el sistema a resolver queda expresado del siguiente modo:

$$\text{Con esto calculamos } q_{\text{máx}} \text{ y } q_{\text{mín}} \rightarrow \boxed{\begin{cases} \mathbb{P}\left(\text{Gauß}\left(\mu = q_{\text{máx}} \sum_{i=1}^{n} \frac{1}{\sigma_i^2}, \sigma = \sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}\right) < \hat{q}_{\text{observado}} \sum \frac{1}{\sigma_i^2}\right) = \dfrac{1 - CL}{2} \\ 1 - \mathbb{P}\left(\text{Gauß}\left(\mu = q_{\text{mín}} \sum_{i=1}^{n} \frac{1}{\sigma_i^2}, \sigma = \sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}\right) \leq \hat{q}_{\text{observado}} \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\right) = \dfrac{1 - CL}{2} \end{cases}}.$$

Este sistema así como está ya se puede usar para calcular $q_{\text{mín}}$ y $q_{\text{máx}}$. Obviamente a mano es imposible, pero en una computadora no es difícil.

Implementación en Python El siguiente script recibe como entrada un archivo CSV con los datos q_i y σ_i y realiza el cálculo del estimador de máxima verosimilitud, i.e. $\hat{q}_{\text{observado}}$, y también del intervalo frecuentista $[q_{\text{mín}}, q_{\text{máx}}]$. Luego imprime en pantalla los resultados.

```
1 from scipy.stats import norm
2 import numpy as np
3
4 CONFIDENCE_LEVEL = .95
```

```
5
6  def freqconfint_gaussian(x, sigma=1, clevel=.68, step=.01):
7      """
8      This function calculates the frequentist symmetric interval for the
9      mean of a sample <x> of gaussian variables, assumed all with the same
10     mean. Different values for each sigma_i is admited if passed as a numpy
11     array in the argument <sigma>.
12     """
13     x_obs = (x/sigma**2).sum() / (1/sigma**2).sum()
14     just_a_number = (sigma**-2).sum()
15     x_max = 0
16     while norm.cdf(just_a_number*x_obs, loc=x_max*just_a_number, scale=
       just_a_number**.5) > (1-clevel)/2:
17         x_max += step
18     x_min = 0
19     while norm.cdf(just_a_number*x_obs, loc=x_min*just_a_number, scale=
       just_a_number**.5) > 1-(1-clevel)/2:
20         x_min += step
21     return x_min, x_max
22
23 data = np.genfromtxt(input('Tell␣me␣the␣name␣of␣the␣file␣with␣the␣data,␣please...\
       n-->␣'), delimiter=',', skip_header = 0) # Read the data.
24 data = data.transpose() # Accomodate the data.
25 q_i = data[0] # Get data.
26 sigma_i = data[1] # Get data.
27 q_min, q_max = freqconfint_gaussian(q_i, sigma = sigma_i, clevel =
       CONFIDENCE_LEVEL) # Calculation of the confidence interval.
28 q_obs = (q_i/sigma_i**2).sum() / (1/sigma_i**2).sum() # Calculation of the MLE
       estimator.
29 print('q_obs␣=␣' + str(q_obs) + '␣is␣the␣MLE')
30 print('[q_min,␣q_max]␣=␣[' + str(q_min) + ',␣' + str(q_max) + ']␣is␣the␣' + str(
       CONFIDENCE_LEVEL*100) + '␣%␣confidence␣level␣interval␣frequentist␣interval')
31 print('The␣previous␣frequentist␣interval␣can␣be␣equivalently␣written␣as␣' + str((
       q_max+q_min)/2) + '␣+-␣' + str((q_max-q_min)/2))
```