Sobre los intervalos de confianza y de predicción

Javier Santibáñez

12 de septiembre de 2017

Intervalos de confianza

Se construyen intervalos de confianza para los parámetros poblacionales. Supongamos que tenemos una muestra aleatoria $\mathbf{X} = \{X_1, \ldots, X_n\}$ de una población $F(x|\theta)$, con θ fijo pero desconocido. Un intervalo de confianza $100(1-\alpha)\%$ para θ está formado por dos estadistícos $L(\mathbf{X})$ y $U(\mathbf{X})$ tales que

$$P\left(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})\right) \geq 1 - \alpha$$

Por ejemplo, considerar una muestra aleatoria $\mathbf{X} = \{X_1, \ldots, X_n\}$ de una población $N(\mu, \sigma^2)$, con μ y σ^2 desconocidos. De los cursos de inferencia estadística, se sabe que

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{S_n} \sim t_{n-1}$$

donde $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ y $S_n^2 = (n-1)^{-1}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2$. De lo anterior, se sigue que los estadísticos

$$\begin{aligned} L\left(\mathbf{X}\right) &= \bar{X}_n - t_{n-1}^{(\alpha/2)}\sqrt{S_n^2/n} \\ U\left(\mathbf{X}\right) &= \bar{X}_n + t_{n-1}^{(\alpha/2)}\sqrt{S_n^2/n} \end{aligned}$$

son tales que

$$P\left(L(\mathbf{X}) \leq \mu \leq U(\mathbf{X})\right) = 1 - \alpha$$

Ejemplo. Intervalo para la media de una población normal.

Consideremos una realización de una muestra aleatoria de tamaño 15 de una población $N(5, 1)$.

```
set.seed(1010)
x <- rnorm(15, 5, 1); x
```

```
##  [1] 5.131541 4.102090 6.351945 5.420075 4.711540 6.364561 6.938725
##  [8] 2.639017 4.405886 4.455447 4.804092 4.169927 3.564619 4.923929
## [15] 5.725868
```

```
xn <- mean(x); xn
```

```
## [1] 4.913951
```

```
sn <- var(x); sn
```

```
## [1] 1.283195
```

Calculamos los estadísticos $L(\mathbf{X})$ y $U(\mathbf{X})$ con $\alpha = 0.05$ para esta muestra en particular.

```
lx <- xn - qt(0.975, 14) * sqrt(sn / 15); lx
```

```
## [1] 4.286637
```

```
ux <- xn + qt(0.975, 14) * sqrt(sn / 15); ux
```

```
## [1] 5.541265
```

Los resultados son $L(\mathbf{x}) = 4.29$ y $U(\mathbf{x}) = 5.54$. ¿Qué significa esto?

Interpretación de los intervalos de confianza

Una vez que se observa la muestra $\mathbf{X} = \mathbf{x}$, el enunciado $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ ya no es probabilista, puesto que todas las cantidades son fijas. Según la interpretación frecuentista de la probabilidad, si repitieramos el experimento que generó los datos, un número grande de veces, el intervalo $(L(\mathbf{x}), U(\mathbf{x}))$ contendría a μ en el 95% de los casos.

Una vez que se observan los datos, el intervalo es fijo, por lo que de conocer el verdadero valor del parámetro, se podría decidir si está o no incluido en el intervalo. Sin embargo, el verdadero valor del parámetri no se conoce, pero se tiene *confianza* en que el intervalo observado sea uno de los que contienen al verdadero valor del parámetro.

Ejemplo (continuación)

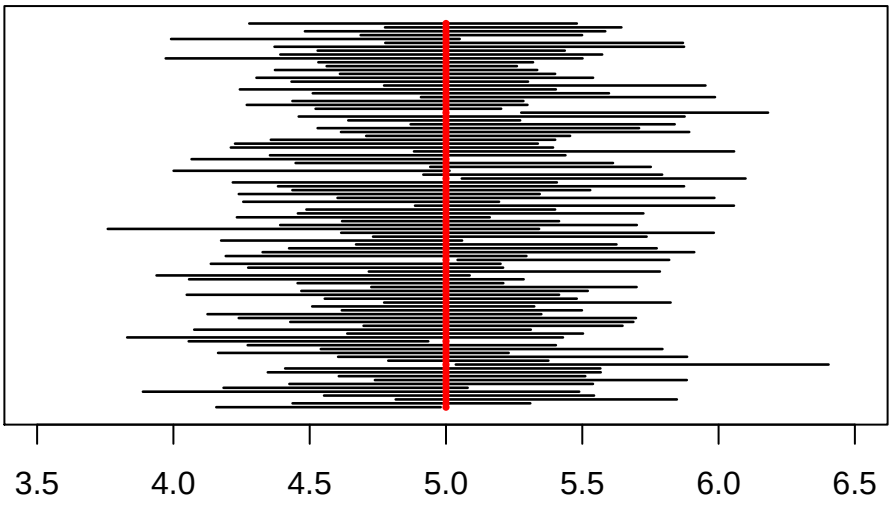
El enunciado μ está entre 4.29 y 5.54 con una confianza del 95%, es correcto. Sin embargo, esto significa que *confiamos* en que el intervalo (4.29, 5.54) sea uno de los que sí contienen a μ. Como conocemos el verdadero valor de μ para este ejercicio de simulación, sabemos que el intervalo observado sí contiene al parámetro.

Con el siguiente código se simulan 5000 muestras de la población $N(5, 1)$, con cada realización se calculan $L(\mathbf{x})$ y $U(\mathbf{x})$ y se verifica si contienen al parámetro μ.

```
vxn <- c(); vsn <- c()
for (i in 1:5000){
  muestra <- rnorm(15, 5, 1)
  vxn[i] <- mean(muestra)
  vsn[i] <- var(muestra)
}

vlxc <- vxn - qt(0.975, 14)*sqrt(vsn/15)
vuxc <- vxn + qt(0.975, 14)*sqrt(vsn/15)

auxc <- (vlxc <= 5) & (5 <= vuxc)
cobertura <- 100*(sum(auxc)/5000); cobertura
```

```
## [1] 94.62
```

El resultado es que en el 94.6 % de los casos el parámetro está contenido en el intervalo calculado, que es aproximadamente el 95% que se esperaba. A continuación se representan graficamente 100 de los intervalos simulados se puede ver como sólo en 7 casos el intervalo (segmento negro) no contiene a μ (línea roja), que dista poco de los 5 que se esperaban.



Intervalos de predicción

Se construyen intervalos de predicción para variables aleatorias. Supongamos que tenemos una muestra aleatoria $\mathbf{X}$ de una población $F(x|\theta)$ y se quiere predecir el valor de una nueva observación X_{new} a partir de la infromación de la muestra observada. Formalmente $X_{new} \perp \mathbf{X}$ por lo que toda la información sobre X_{new} se obtiene del hecho que viene de un población $F(x|\theta)$. Como θ es desconocido, es encontrar dos estadísticos, funciones de $\mathbf{X}$ cuya distribución no dependa de los parámetros desconocidos. Un intervalo de confianza está formado por dos estadísticos $L_p(\mathbf{X})$ y $U_p(\mathbf{X})$ tales que

$$P\left(L_p(\mathbf{X}) \leq X_{new} \leq U_p(\mathbf{X})\right) \geq 1-\alpha$$

Una vez observada la muestra, $\mathbf{X}=\mathbf{x}$, el enunciado $L_p(\mathbf{x}) \leq X_{new} \leq U_p(\mathbf{x})$ aún es probabilista, puesto que X_{new} es una variable aleatoria. Sin embargo, no necesariamente

$$P\left(L_p(\mathbf{x}) \leq X_{new} \leq U_p(\mathbf{x})\right) \geq 1-\alpha$$

Por lo que el intervalo observado $(L_p(\mathbf{x}), U_p(\mathbf{x}))$ no es un intervalo de probabilidad $1-\alpha$ para X_{new}.

Como se asume que X_{new} es independiente de $\mathbf{X}$, se sigue que

$$X_{new} - \bar{X}_n \sim N\left(0, \sigma^2\left(1+\frac{1}{n}\right)\right).$$

Entonces,

$$\frac{X_{new}-\bar{X}_n}{\sqrt{S_n^2\left(1+\frac{1}{n}\right)}} \sim t_{n-1}$$

Por lo tanto, los estadísticos

$$L_p(\mathbf{X}) = \bar{X}_n - t_{n-1}^{(\alpha/2)}\sqrt{S_n^2\left(1+\frac{1}{n}\right)}$$

$$U_p(\mathbf{X}) = \bar{X}_n + t_{n-1}^{(\alpha/2)}\sqrt{S_n^2\left(1+\frac{1}{n}\right)}$$

son tales que

$$P\left(L_p(\mathbf{X}) \leq X_{new} \leq U_p(\mathbf{X})\right) = 1 - \alpha$$

Ejercicio de simulación

Con la muestra del ejemplo anterior se calcula un intervalo de confianza para una nueva observación como sigue:

```
lx <- xn - qt(0.975, 14)*sqrt(sn*(1+1/15)); lx
```

```
## [1] 2.404696
```

```
ux <- xn + qt(0.975, 14)*sqrt(sn*(1+1/15)); ux
```

```
## [1] 7.423206
```

El intervalo que resulta es $L_p(\mathbf{x}) = 2.4$ y $U(\mathbf{x}) = 7.42$. ¿Qué significan estos resultados?

Interpretación de los resultados

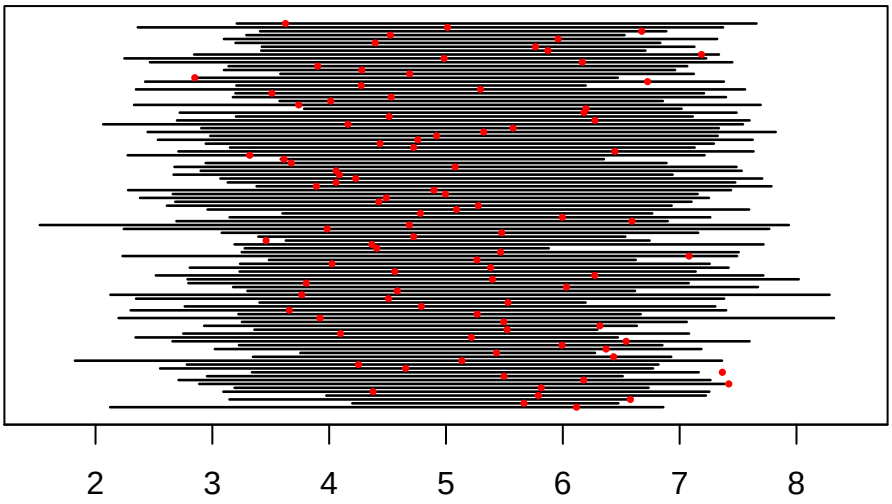
Como se mencionó anteriormente, el intervalo no es de probabilidad para X_{new}. El enunciado, una nueva observación X_{new} estará entre 2.4 y 7.42 con probabilidad 0.95 es falso, porque como $X_{new} \sim N(5, 1)$, tal probabilidad es 0.99. La interpretación correcta es en términos de la *confianza*. El enunciado una nueva observación X_{new} estará entre 2.4 y 7.42 con 95% de confianza es la interpretación correcta. ¿Qué significa la *emph* en predicción? la interpretación es la siguiente, si repetimos el experimento que generó a $\mathbf{x}$ un número grande de veces, para cada repetición observamos X_new, y verificamos si x_{new} está contenida o no en el intervalo $(L_p(\mathbf{x}), U_p(\mathbf{x}))$; se espera que en el 95% de las repeticiones el intervalo de predicción cubra a x_{new}. Por lo tanto, los intervalos de predicción también son intervalos de confianza.

Con el siguiente código se simulan 5000 muestras de la población $N(5, 1)$, con cada realización se calculan $L_p(\mathbf{x})$ y $U_p(\mathbf{x})$, se simualan 5000 nuevas observaciones de la población y se verifica si cada realización de $L_p(\mathbf{x})$ y $U_p(\mathbf{x})$ contiene a la observación de X_{new} que correspnde.

```
vxn <- c(); vsn <- c()
for (i in 1:5000){
  muestra <- rnorm(15, 5, 1)
  vxn[i] <- mean(muestra)
  vsn[i] <- var(muestra)
}

vlxp <- vxn - qt(0.975, 14)*sqrt(vsn*(1+1/15))
vuxp <- vxn + qt(0.975, 14)*sqrt(vsn*(1+1/15))

xnew <- rnorm(5000, 5, 1)

auxp <- (vlxp <= xnew) & (xnew <= vuxp)
cobertura <- 100*(sum(auxp)/5000); cobertura
```

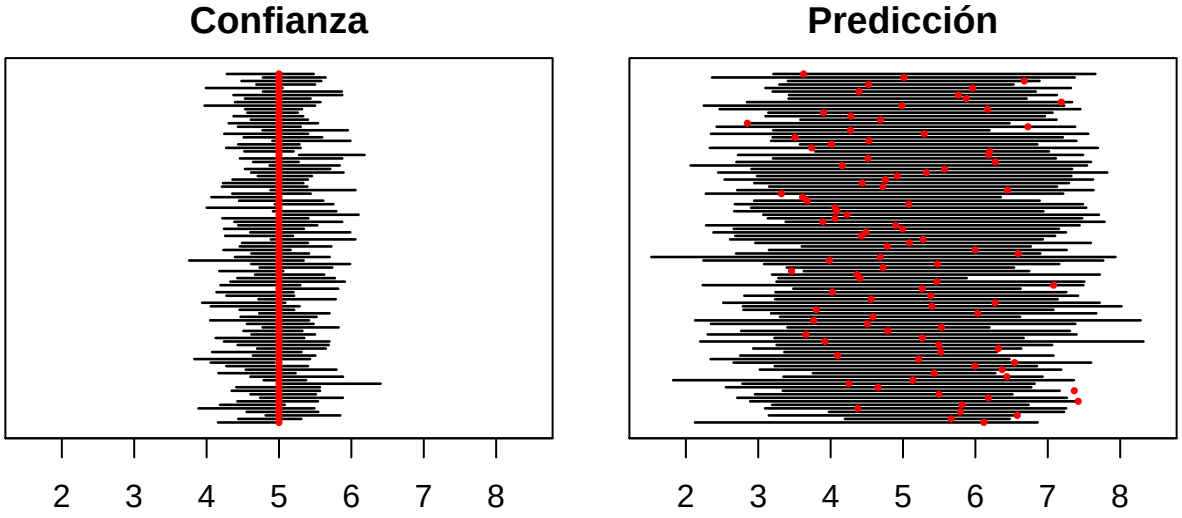
```
## [1] 95.1
```

El resultado es que en el 95.1 % de los casos el parámetro está contenido en el intervalo calculado, que es aproximadamente el 95% que se esperaba. A continuación se representan graficamente 100 de los intervalos simulados se puede ver como sólo en 4 casos el intervalo (segmento negro) no contiene a la nueva observación (punto rojo), que dista poco de los 5 que se esperaban.



Difencias entre predicción y confianza

Lo primero que se debe notar es que los intervalos de predicción para una nueva observación son más amplios que los intervalos de confianza para los parámetros desconocidos. ¿Por qué? El tamaño del intervalo de confianza para el parámetro θ depende de la incertidumbre de la estimación que hacemos a partir de una muestra. Mientras que el tamaño del intervalo de predicción para una nueva observación tiene dos fuentes de incertidumbre, una debida a la estimación de los parámentros desconocidos y la otra es propia de la aleatoriedad que suponemos, proque se debe recordar que esa nueva observación es una variable aleatoria!.



Para entender mejor la diferencia entre cada tipo de intervalo, consideremos el caso extremo en que conocemos los verdaderos parámetros de la población. En tal caso, se elimina completamente la incertidumbre sobre μ, por lo que no tendría sentido constuir un intervalo de confianza para este parámetro. Mientras que una nueva observación aún es aleatoria, porque ese es nuestro supuesto, entonces aún podríamos construir un intervalo de predicción.