# Sobre los intervalos de confianza y de predicción

*Javier Santibáñez*

*28 de febrero de 2018*

## Intervalos de confianza

Se construyen intervalos de confianza para parámetros. Sea $\mathbf{X} = \{X_1, \ldots, X_n\}$ una muestra aleatoria de una población con distribución $F(x \,|\, \theta)$, donde $\theta$ desconocido. Un intervalo de confianza $100(1-\alpha)\%$ para $\theta$ se conforma de dos estadistícos $L(\mathbf{X})$ y $U(\mathbf{X})$ que cumplen con la siguiente condición

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

### Ejemplo 1

Sea una muestra aleatoria $\mathbf{X} = \{X_1, \ldots, X_n\}$ de una población con distribución $N(\mu, \sigma^2)$, con $\mu$ y $\sigma^2$ 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 $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$, $S_n^2 = (n-1)^{-1}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2$ y $t_{(n-1)}$ denota la distribución $t$ con $n-1$ grados de libertad. De lo anterior, se sigue que los estadísticos

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

con $t_{(n-1)}^{(1-\alpha/2)}$ el cuantil $1-\alpha/2$ de la distribución $t_{(n-1)}$, cumplen con

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

Consideremos la siguiente realización de una muestra aleatoria de tamaño $n = 15$ de una población con distribución $N(5, 1)$ con la que se quiere construir un intervalo de confianza 95%.

```
##  [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
```

Con las expresiones anteriores se calculan los estadísticos $L(\mathbf{X})$ y $U(\mathbf{X})$.

```
xn <- mean(x) # Media muestral
sn <- var(x)  # Varianza muestral
lx <- xn - qt(0.975, 14) * sqrt(sn / 15); lx # Límite inferior del IC
```

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

```
ux <- xn + qt(0.975, 14) * sqrt(sn / 15); ux # Límite superior del IC
```

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

Los resultados anteriores indican que el intervalo de confianza 95% calculado con esta muestra tiene extremos $L(\mathbf{x}) = 4.29$ y $U(\mathbf{x}) = 5.54$. ¿Cuál es la correcta interpretación de estos resultados?

## Interpretación de los intervalos de confianza

Una vez que se observa la muestra $\mathbf{X} = \mathbf{x}$, el enunciado $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ deja de ser probabilista, puesto que ya no involucra cantidades aleatorias. De conocer el verdadero valor de $\theta$, se podría decidir si está o no incluido en el intervalo calculado. Sin embargo el valor de $\theta$ es desconocido, es por ello que sólo se puede *confiar* en que el intervalo calculado contenga al parámetro.

La confianza de los intervalos se interpreta de la siguiente manera. Si fuera posible observar un número grande de muestras aleatorias de la misma población $\mathbf{x}_1, \ldots, \mathbf{x}_m$ y con cada una de ellas calcular el intervalo $(L(\mathbf{x}), U(\mathbf{x}))$, de acuerdo con la interpretación frecuentista de la probabilidad, aproximadamente el $100(1-\alpha)\%$, o bien una fracción $1-\alpha$, de los intervalos calculados contendría a $\theta$.

## Ejemplo 2

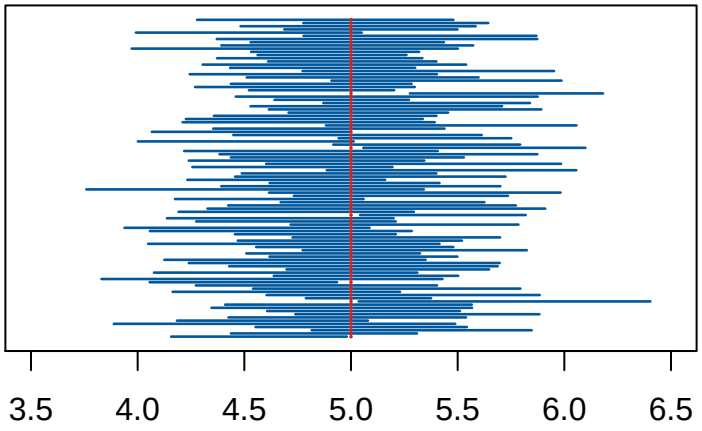
En este ejemplo se evaluará la cobertura del intervalo obtenido para la media de una población normal, como en el Ejemplo 1. Con el siguiente código se simulan 5000 muestras de tamaño $n = 15$ de una población con distribución $N(5, 1)$, con cada realización se calculan $L(\mathbf{x})$ y $U(\mathbf{x})$ y se verifica si contienen al parámetro $\mu$.

```
v_lx <- c()        # Vector para almacenar L(x)
v_ux <- c()        # Vector para almacenar U(x)
cobertura <- c()   # Vector para almacenar la cobertura de los IC
for (k in 1:5000)
{
  muestra <- rnorm(15, 5, 1)  # Generar la muestra N(5, 1) de tamaño 15
  xn_ <- mean(muestra); sn_ <- var(muestra)       # Estadísticos muestrales
  v_lx[k] <- xn_ - qt(0.975, 14) * sqrt(sn_/15)   # Límite inferior del IC
  v_ux[k] <- xn_ + qt(0.975, 14) * sqrt(sn_/15)   # Límite superior del IC
  cobertura[k] <- (v_lx[k] <= 5) & (5 <= v_ux[k]) # Verificar cobertura
}
confianza <- 100 * sum(cobertura) / 5000 # Porcentaje de cobertura
confianza
```

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

Los resultados anteriores muestran que el 94.62% de los intervalos contienen al parámetro, que en este ejercicio de simulación tiene un valor conocido $\mu = 5$. El porcentaje calculado no dista demasiado del 95% que se esperaba, de hecho, si el código anterior se ejecuta repetidamente, se puede observar como el procentaje de cobertura oscila alrededor de 95%. El porcentaje de cobertura mejora si se consideran más repeticiones, por ejemplo $m = 50,000$.

A continuación se representan graficamente los 100 primeros intervalos calculados. En 7 casos $\mu = 5$ no está contenido en el intervalo.



# Intervalos de predicción

Se construyen intervalos de predicción para variables aleatorias. Sea $\mathbf{X} = \{X_1, \ldots, X_n\}$ una muestra aleatoria de una población con distribución $F(x|\theta)$, donde $\theta$ desconocido y que se tiene interés en predecir el valor de una nueva observación $X_*$ (independiente de la muestra).

El objetivo de hacer predicciones (por intervalos) es determinar con rango de posibles valores entre los cuáles sea razonable suponer estará la nueva observación. En el caso en que $\theta$ fuera conocido, se podrían tomar los cuantiles $\alpha/2$ y $1-\alpha/2$ de $F(x\,|\,\theta)$, $F^{(\alpha/2)}$ y $F^{(1-\alpha/2)}$ respectivamente, y así proponer $\left\{F^{(\alpha/2)}, F^{(1-\alpha/2)}\right\}$ como intervalo de posibles valores para $X_*$. En este caso es razonable suponer que $X_*$ tomé algún valor de intervalo ya que

$$P\left\{F^{(\alpha/2)} \leq X_* \leq F^{(1-\alpha/2)}\right\} = 1-\alpha$$

En este caso se dice que $\left\{F^{(\alpha/2)}, F^{(1-\alpha/2)}\right\}$ conforman un intervalo de probabilidad $1-\alpha$ para $X_*$.

El problema radica en que $\theta$ es desconocido y tiene que se estimado a partir de $\mathbf{X}$ para poder hacer predicciones. Un intervalo de predicción para una nueva observación $X_*$ de confianza $100(1-\alpha)\%$ se conforma de dos estadísticos $L_p(\mathbf{X})$ y $U_p(\mathbf{X})$ que satisfacen

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

Suponer que se tiene una muestra aleatoria $\mathbf{X} = \{X_1, \ldots, X_n\}$ de una población con distribución $N(\mu, \sigma^2)$, con ambos parámetros desconocidos y que se requiere construir un intervalo de predicción de confianza $100(1-\alpha)\%$ para una nueva observación independiente $X_*$.

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

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

De lo anterior se sigue que

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

donde $\bar{X}_n$ y $\S_n^2$ se calculan sólo con $\mathbf{X}$. Por lo tanto, los estadísticos

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

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

cumplen con la condición

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

Así, $(L_p(\mathbf{X}), U_p(\mathbf{X}))$ es el intervalo buscado.

## Ejemplo 3

Con la muestra del Ejemplo 1 se calcula un intervalo de confianza para una nueva observación como sigue

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

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

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

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

Estos resultados indican que el intervalo de predicción para una nueva observación $X_*$ de confianza 95% calculado con esta muestra tiene extremos $L(\mathbf{x}) = 2.4$ y $U(\mathbf{x}) = 7.42$. Pero, ¿cuál es la correcta interpretación de estos resultados?

## Interpretación de los intervalos de predicción

Una vez que se observa la muestra $\mathbf{X} = \mathbf{x}$ y que se calculan los estadísticos $L_p(\mathbf{x})$ y $U_p(\mathbf{x})$, el enunciado $L_p(\mathbf{x}) \leq X_* \leq U_p(\mathbf{x})$ aún es probabilista, puesto que $X_*$ es una variable aleatoria. Sin embargo, no necesariamente se cumple

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

Sin embargo, no es posible saberlo ya que para ello es necesario conocer a $\theta$. Con esto se justifica que los intervalos de predicción no son intervalos de probabilidad y por ello su interpretación está dada en términos de confianza.

La confianza de los intervalos de predicción se explica con el siguiente experimento mental. Suponer que es posible observar una nueva muestra aleatoria $\mathbf{x}_s$ de la población y que con ella se calcula el intervalo $(L_p(\mathbf{x}_1), U_p(\mathbf{x}_1))$; posteriormente, suponer que se observa una nueva $x_s$ independiente de la muestra $\mathbf{x}_s$ y se verifica si esta observación adicional está contenida en $(L_p(\mathbf{x}_1), U_p(\mathbf{x}_1))$; pensar que es posible repetir lo anterior un número grande de veces. La confianza de los intervalos de predicción es el porcentaje de intervalos que sí continen a la nueva observación.

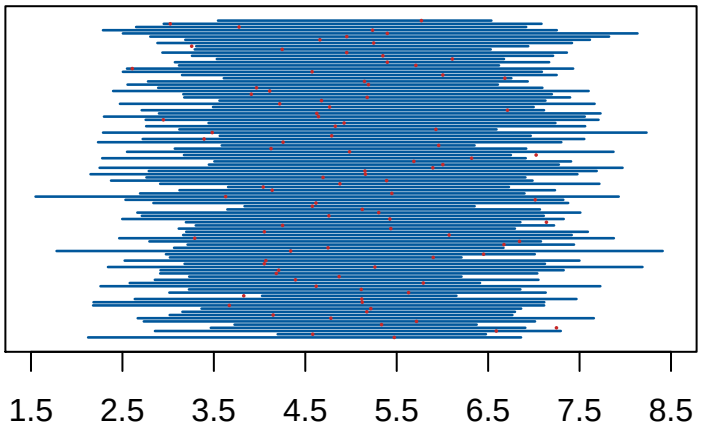
## Ejemplo 4

En este ejemplo se evaluará la cobertura del intervalo de predicción de confianza 95% obtenido para una nueva observación de una población normal, como en el Ejemplo 2. Con el siguiente código se simulan 5000 muestras de tamaño $n = 15$ de una población con distribución $N(5, 1)$ y para cada una se genera una nueva observación $x_s$, con cada realización se calculan $L_p(\mathbf{x})$ y $U_p(\mathbf{x})$, posteriormente se verifica si el intervalo calculado contiene a la observación adicional.

```
xs <- c()          # Vector para almacenar las nuevas observaciones
v_lx_p <- c()      # Vector para almacenar L_p(x)
v_ux_p <- c()      # Vector para almacenar U_p(x)
cobertura <- c()   # Vector para almacenar la cobertura
for (k in 1:5000)
{
  muestra <- rnorm(15, 5, 1)  # Generar la muestra N(5, 1) de tamaño 15
  xs[k] <- rnorm(1, 5, 1)     # Observación adicional
  xn_ <- mean(muestra); sn_ <- var(muestra) # Estadísticos muestrales
  v_lx_p[k] <- xn_ - qt(0.975, 14) * sqrt(sn_*(1+1/15)) # Límite inferior del IC
  v_ux_p[k] <- xn_ + qt(0.975, 14) * sqrt(sn_*(1+1/15)) # Límite superior del IC
  cobertura[k] <- (v_lx_p[k] <= xs[k]) & (xs[k] <= v_ux_p[k]) # Verificar cobertura
}
confianza <- 100 * sum(cobertura) / 5000 # Porcentaje de cobertura
confianza
```

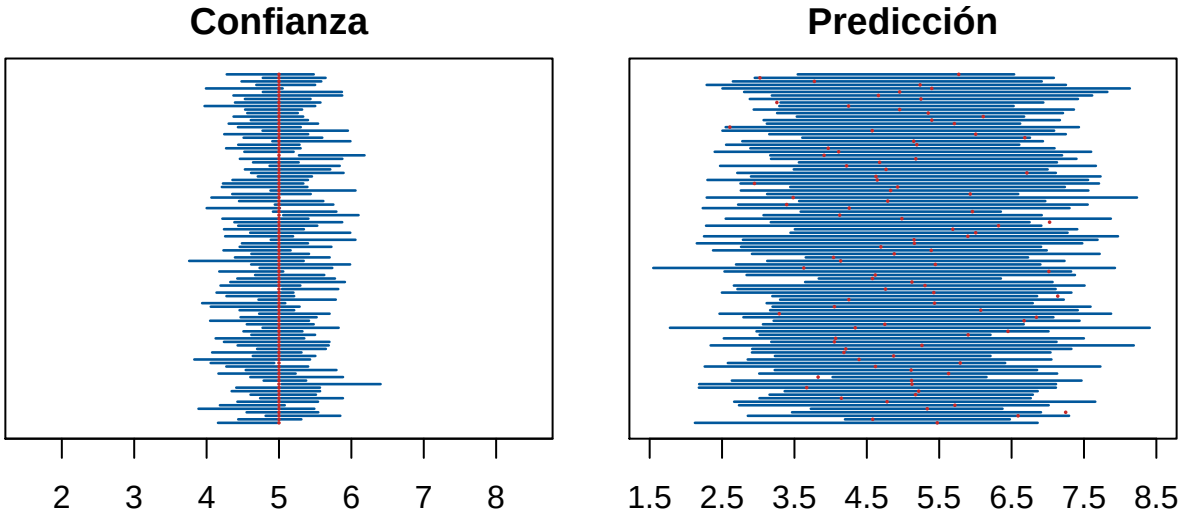
```
## [1] 95.04
```

Los resultados muestran que el 95.04% de los intervalos contienen a la nueva observación para la que fueron calculados. El porcentaje de cobertura calculado no dista mucho del 95% que se esperaba. A continuación se representan graficamente los 100 primeros intervalos calculados. En 5 casos la nueva observación no está contenida en el intervalo.



## Diferencias 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 $\theta$ 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, porque 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 $\mu$, 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.