Inferencia Estadística

Felipe José Bravo Márquez

19 de noviembre de 2013

Inferencia Estadística

- Para realizar conclusiones sobre una **población**, generalmente no es factible reunir todos los datos de ésta.
- Debemos realizar conclusiones razonables respecto a una población basado en la evidencia otorgada por **datos muestrales**.
- El proceso de realizar conclusiones sobre una población a partir de datos muestrales se conoce como **inferencia estadística**.

Inferencia Estadística (2)

- En inferencia estadística tratamos de **inferir** la distribución que genera los datos observados
- Ejemplo: Dado una muestra $X_1, \ldots, X_n \sim F$. ¿Cómo inferimos F?
- En algunos cosas sólo nos interesa inferir alguna propiedad de F como su **media**.
- Los modelos estadísticos que asumen que la distribución se puede modelar con un conjunto finito de parámetros $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ se llaman modelos **paramétricos**.
- Ejemplo: si asumimos que los datos vienen de una distribución normal $N(\mu, \sigma^2)$, μ y σ serían los parámetros del modelo.
- Un **estadístico** (muestral) es una medida cuantitativa calculada a partir de los datos.

Estimación Puntual

- La estimación puntual es el proceso de encontrar la **mejor aproximación** de una cantidad de interés a partir de una **muestra estadística**.
- La cantidad de interés puede ser: un parámetro en un modelo paramétrico, una CDF, una PDF, o una función de regresión.
- Por convención se denota a la estimación puntual del valor de interés θ como $\hat{\theta}$ o $\hat{\theta}_n$
- Es importante remarcar que mientras θ es un valor fijo desconocido, $\hat{\theta}$ depende de los datos y por ende es una variable aleatoria.

Estimación Puntual (2)

Definición Formal

- Sean $X_1, \ldots, X_n$ n observaciones IID de una distribución F
- Un estimador puntual $\hat{\theta}_n$ de un parámetro θ es una función de $X_1, \ldots, X_n$:

$$\hat{\theta}_n = g(X_1, \ldots, X_n)$$

- El **sesgo** (bias) de un estimador se define como:

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$$

- Un estimador es insesgado si $\mathbb{E}(\hat{\theta}_n) = \theta$ o $\text{bias}(\hat{\theta}_n) = 0$

Estimación Puntual (3)

- La distribución de $\hat{\theta}_n$ se conoce como la **distribución muestral**
- La desviación estándar de $\hat{\theta}_n$ se conoce como **error estándar** *se*:

$$se(\hat{\theta}_n) = \sqrt{\mathbb{V}(\hat{\theta}_n)}$$

- El error estándar nos habla sobre la variabilidad del estimador entre todas las posibles muestras de un mismo tamaño.

Estimación Puntual (4)

- Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población de media μ y varianza σ^2
- Se define la **media muestral** $\overline{X_n}$ o $\hat{\mu}$ como:

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Es un estimador insesgado:

$$\mathbb{E}(\overline{X_n}) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n} \times \mathbb{E}(\sum_{i=1}^{n} X_i) = \frac{1}{n}(n \times \mu) = \mu$$

- Su error estándar sería $se(\overline{X_n}) = \sqrt{\mathbb{V}(\overline{X_n})}$ donde

$$\mathbb{V}(\overline{X_n}) = \mathbb{V}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\mathbb{V}(\sum_{i=1}^{n} X_i) = \frac{n}{n^2}\mathbb{V}(X_i) = \frac{\sigma^2}{n}$$

- Entonces $se(\overline{X_n}) = \frac{\sigma}{\sqrt{n}}$

Ejemplos de Estimación Puntual (5)

- Por lo general no sabemos σ de la población.
- Cuando queremos estimar la varianza de una población a partir de una muestra hablamos de la **varianza muestral**:
- Existen dos estimadores comunes, una versión sesgada

$$s_n^2 = \frac{1}{n}\sum_i^n (X_i - \overline{X_n})^2$$

- Una versión sin sesgo

$$s^2 = \frac{1}{n-1}\sum_i^n (X_i - \overline{X_n})^2$$

- Cuando no sabemos la varianza de la población y queremos estimar la media, el error estándar es estimado:

$$\hat{se}(\overline{X_n}) = \frac{s}{\sqrt{n}}$$

Estimación Puntual (6)

- Sean $X_1, \ldots, X_n \sim$ Bernoulli(p) y sea $\hat{p}_n = \frac{1}{n}\sum_i X_i$
- Luego $\mathbb{E}(\hat{p}_n) = \frac{1}{n}\sum_i \mathbb{E}(X_i) = p$, entonces $\hat{p}_n$ es insesgado.
- El error estándar *se* sería

$$se = \sqrt{\mathbb{V}(\hat{p}_n)} = \sqrt{p(1-p)/n}$$

- El error estándar estimado $\hat{se}$:

$$\hat{se} = \sqrt{\hat{p}(1-\hat{p})/n}$$

Estimación Puntual (7)

- Se espera que un buen estimador sea insesgado y de mínima varianza.
- Un estimador puntual $\hat{\theta}_n$ de un parámetro θ es **consistente** si converge al valor verdadero cuando el número de datos de la muestra tiende a infinito.
- La calidad de un estimador se puede medir usando el **error cuadrático medio** (MSE)

$$MSE = \mathbb{E}_\theta(\hat{\theta}_n - \theta)^2$$

Estimación Puntual (8)

- Si para un estimador $\hat{\theta}_n$, su *bias* $\rightarrow 0$ y su *se* $\rightarrow 0$ cuando $n \rightarrow \infty$, $\hat{\theta}_n$ es un estimador consistente de θ.
- Por ejemplo, para la media muestral $\mathbb{E}(\overline{X_n}) = \mu$ lo que implica que el *bias* $= 0$ y $se(\overline{X_n}) = \frac{\sigma}{\sqrt{n}}$ que tiende a cero cuando $n \rightarrow \infty$. Entonces $\overline{X_n}$ es un estimador consistente de la media.
- Para el caso del experimento Bernoulli se tiene que $\mathbb{E}(\hat{p}) = p \Rightarrow$ *bias* $= 0$ y $se = \sqrt{p(1-p)/n} \rightarrow 0$ cuando $n \rightarrow \infty$. Entonces $\hat{p}$ es un estimador consistente de p.

Intervalo de Confianza

- Sabemos que el valor de un estimador puntual **varía** entre una muestra y otra
- Es más razonable encontrar un **intervalo** donde sepamos que valor **real del parámetro** se encuentra dentro del intervalo con una cierta **probabilidad**.
- La forma general de un intervalo de confianza en las siguiente:

$$\text{Intervalo de Confianza} = \text{Estadístico Muestral} \pm \text{Margen de Error}$$

- Entre más ancho el intervalo mayor incertidumbre existe sobre el valor del parámetro.

Intervalo de Confianza (2)

Definición

- Un **intervalo de confianza** para un parámetro poblacional desconocido θ con un **nivel de confianza** $1-\alpha$, es un intervalo $C_n = (a, b)$ donde:

$$\mathbb{P}(\theta \in C_n) = 1 - \alpha$$

- Además $a = a(X_1, \ldots, X_n)$ y $b = b(X_1, \ldots, X_n)$ son funciones de los datos
- El valor α se conoce como el nivel de **significancia**, generalmente se toma como 0,05 lo que equivale a trabajar con un nivel de confianza de 95 %
- La significancia se puede interpretar como la probabilidad de equivocarnos.

Intervalo de Confianza (3)

Interpretación

- Existe mucha **confusión** de como interpretar un intervalo de confianza
- Una forma de interpretarlos es decir que si repetimos **un mismo experimento** muchas veces, el intervalo contendrá el valor del parámetro el $(1-\alpha)$ % de las veces.
- Esta interpretación es correcta, pero rara vez repetimos un mismo experimento varias veces.
- Una interpretación mejor: un día recolecto datos creo un intervalo de 95 % de confianza para un parámetro θ_1. Luego, en el día 2 hago lo mismo para un parámetro θ_2 y así reiteradamente *n* veces. El 95 % de mis intervalos **contendrá** los valores reales de los parámetros.

Intervalo de Confianza (4)

- Se tienen n observaciones independientes $X_1, \ldots, X_n$ IID de distribución $N(\mu, \sigma^2)$
- Supongamos que μ es **desconocido** pero σ^2 es **conocido**.
- Sabemos que $\overline{X_n}$ es un estimador insesgado de μ
- Por la ley de los grandes números sabemos que la distribución de $\overline{X_n}$ se concentra alrededor de μ cuando n es grande.
- Por el CLT sabemos que

$$Z = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

  cuando n es grande
- Despejando, tenemos que $\mu = \overline{X_n} - \frac{\sigma}{\sqrt{n}} Z$

Intervalo de Confianza (5)

- Queremos encontrar un intervalo intervalo $C_n = (\mu_1, \mu_2)$ con un nivel de confianza $1 - \alpha$:

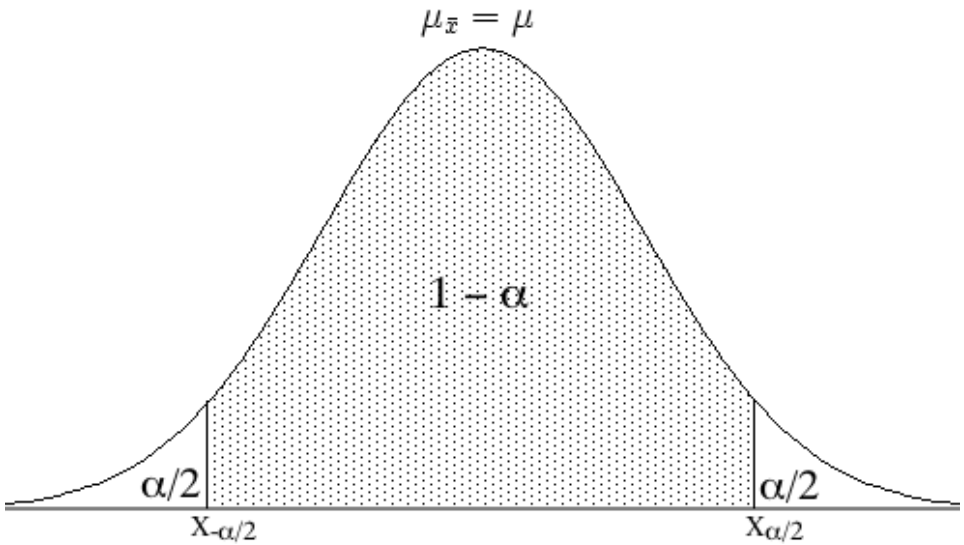
$$\mathbb{P}(\mu_1 \leq \mu \leq \mu_2) = 1 - \alpha$$

- Sea $z_a = \Phi^{-1}(1 - a)$, con $a \in [0, 1]$ donde Φ^{-1} es la función cuantía de una normal estandarizada
- Esto es equivalente a decir que z_a es el valor tal que $1 - \Phi(z_a) = \mathbb{P}(Z \geq z_a) = a$
- Por simetría de la normal $z_{\alpha/2} = -z_{(1-\alpha/2)}$

Intervalo de Confianza (6)

- Se tiene que

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$



Intervalo de Confianza (7)

- El intervalo de confianza para μ es:

$$C_n = (\overline{X_n} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X_n} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

- Entonces $z_{\alpha/2}$ nos dice cuantas veces tenemos que multiplicar el **error estándar** en el intervalo.
- Mientras menor sea α mayor será $z_{\alpha/2}$ y por ende más ancho será el intervalo.
- Demostración:

$$\begin{aligned}
\mathbb{P}(\mu \in C_n) &= \mathbb{P}(\overline{X_n} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X_n} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \\
&= \mathbb{P}(-z_{\alpha/2} < \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}) \\
&= \mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) \\
&= 1 - \alpha
\end{aligned}$$

Intervalo de Confianza (8)

- Como $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ podemos usar la función cuantía de la normal para calcular intervalos de confianza en R

```
> alpha <- 0.05
> xbar <- 5
> sigma <- 2
> n <- 20
> se <-sigma/sqrt(n)
> error <- qnorm(1-alpha/2)*se
> left <- xbar-error
> right <- xbar+error
> left
[1] 4.123477
> right
[1] 5.876523
>
```

Distribución T

- En la práctica, si no conocemos μ es poco probable que conozcamos σ
- Si estimamos σ usando s, los intervalos de confianza se construyen usando la distribución **T-student**

Distribución T

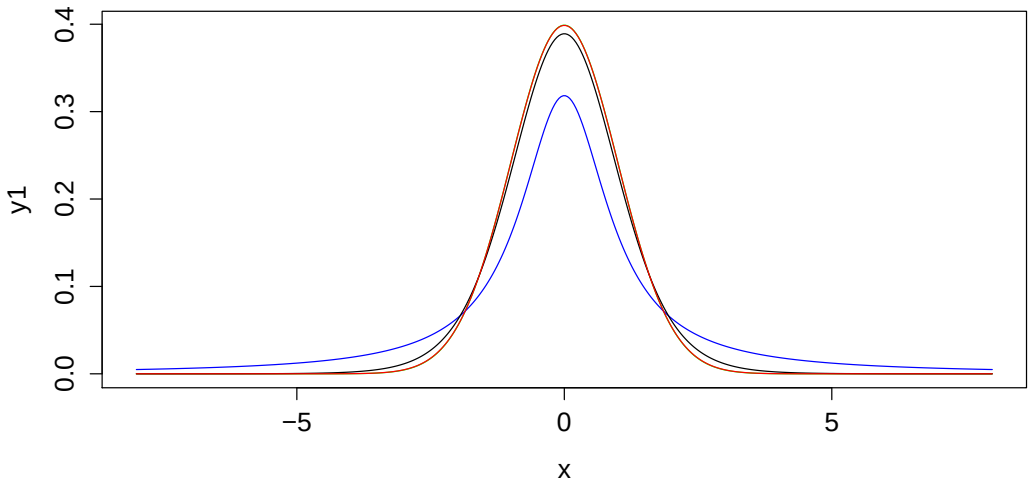
- Una V.A tiene distribución t con k grados de libertad cuando tiene la siguiente PDF:

$$f(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})(1+\frac{t^2}{k})^{(k+1)/2}}$$

- Cuando $k = 1$ se le llama distribución de **Cauchy**
- Cuando $k \to \infty$ converge a una distribución normal estandarizada
- La distribución t tiene colas más anchas que la normal cuando tiene pocos grados de libertad

Distribución T (2)

```
x<-seq(-8,8,length=400)
y1<-dnorm(x)
y2<-dt(x=x,df=1)
y3<-dt(x=x,df=10)
y4<-dt(x=x,df=350)
plot(y1~x,type="l",col="green")
lines(y2~x,type="l",col="blue")
lines(y3~x,type="l",col="black")
lines(y4~x,type="l",col="red")
```



Intervalo de Confianza (9)

- Sea $s^2 = \frac{1}{n-1}\sum_i^n (X_i - \overline{X_n})^2$ tenemos:

$$T = \frac{\overline{X_n} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

- Sea $t_{n-1,a} = \mathbb{P}(T > a)$, equivalente a la función cuantía qt evaluada en $(1-a)$
- El intervalo de confianza resultante es:

$$C_n = (\overline{X_n} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \overline{X_n} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}})$$

- Como las colas de la distribución t son más anchos cuando n es pequeño, los intervalos de confianza resultantes son más anchos

Intervalo de Confianza (10)

- Calculemos un intervalo de confianza para la media de `Petal.Length` de los datos del **Iris** con 95 % de confianza

```
>data(iris)
>alpha<-0.05
>n<-length(iris$Petal.Length)
>xbar<-mean(iris$Petal.Length)
>xbar
[1] 3.758
>s<-sd(iris$Petal.Length)
>se<-s/sqrt(n)
>error<-qt(p=1-alpha/2,df=n-1)*se
>left<-xbar-error
>left
[1] 3.473185
>right<-xbar+error
>right
[1] 4.042815
```

- Otra forma:

```
>test<-t.test(iris$Petal.Length,conf.level=0.95)
>test$conf.int
[1] 3.473185 4.042815
```

Test de Hipótesis

- Cuando queremos probar si alguna **propiedad** asumida sobre una población se contrasta con una muestra estadística usamos un **Test de Hipótesis**
- El test se compone de las siguientes hipótesis:
  - **Hipótesis Nula** H_0: Simboliza la situación actual. Lo que se ha considerado real hasta el presente.
  - **Hipótesis Alternativa** H_a: es el modelo alternativo que queremos considerar.
- La idea es encontrar suficiente **evidencia estadística** para rechazar H_0 y poder concluir H_a
- Si no tenemos suficiente evidencia estadística **fallamos en rechazar** H_0

Test de Hipótesis (2)

Metodología para Realizar un Test de Hipótesis

- Elegir una hipótesis nula H_0 y alternativa H_a
- Fijar un nivel de significancia α del test
- Calcular un estadístico T a partir de los datos
- El estadístico T es generalmente un valor estandarizado que podemos chequear en una tabla de distribución
- Definir un criterio de rechazo para la hipótesis nula. Generalmente es un valor crítico c.

Test de Hipótesis (3)

- Ejemplo: Se sabe que la cantidad de horas promedio de uso de Internet mensual en Chile país es de 30 horas
- Supongamos que queremos demostrar que el promedio es distinto a ese valor.
- Tendríamos que $H_0 : \mu = 30$ y $H_a : \mu \neq 30$
- Fijamos $\alpha = 0{,}05$ y recolectamos 100 observaciones
- Supongamos que obtenemos $\overline{X_n} = 28$ y $s = 10$
- Una forma de hacer el test es construir un intervalo de confianza para μ y ver si H_0 está en el intervalo.

```
> 28-qt(p=0.975,99)*10/sqrt(100)
[1] 26.01578
> 28+qt(p=0.975,99)*10/sqrt(100)
[1] 29.98422
```

- El intervalo sería la zona de aceptación de H_0 y todo lo que esté fuera de éste será mi región de rechazo.
- Como 30 está en la región de rechazo, rechazo mi hipótesis nula con un 5 % de confianza.

Test de Hipótesis (4)

- Otra forma de realizar el test es calcular el estadístico $T = \frac{\overline{X_n} - \mu_o}{\frac{s}{\sqrt{n}}}$
- En este caso sería

$$T = \frac{28 - 30}{\frac{10}{\sqrt{100}}} = -2$$

- Como $H_a : \mu \neq 30$, tenemos un test de dos lados, donde la región de aceptación es

$$t_{n-1,1-\alpha/2} < T < t_{n-1,\alpha/2}$$

```
 > qt(0.025,99)
[1] -1.984217
> qt(0.975,99)
[1] 1.984217
```

- Como T está en la región de rechazo, rechazamos la hipótesis nula.

Test de Hipótesis (5)

- Generalmente, además de saber si rechazamos o fallamos en rechazar una hipótesis nula queremos saber la evidencia que tenemos en contra de ella.
- Se define un **p-valor** como la probabilidad de obtener un resultado al menos tan extremo como el observado en los datos dado que la hipótesis nula es verdadera.
- "Extremo" significa lejos de la hipótesis nula.
- Si el **p-valor** es menor que el nivel de significancia α, rechazamos H_0
- Ejemplo:

```
> data(iris)
> mu<-3 # La hipótesis nula
> alpha<-0.05
> n<-length(iris$Petal.Length)
> xbar<-mean(iris$Petal.Length)
> s<-sd(iris$Petal.Length)
> se<-s/sqrt(n)
> t<-(xbar-mu)/(s/sqrt(n))
> pvalue<-2*pt(-abs(t),df=n-1)
> pvalue
[1] 4.94568e-07 # es menor que 0.05 entonces rechazamos H0
```

Test de Hipótesis (6)

- La forma elegante de hacerlo en R:

```
> t.test(x=iris$Petal.Length,mu=3)

One Sample t-test

data:  iris$Petal.Length
t = 5.2589, df = 149, p-value = 4.946e-07
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 3.473185 4.042815
sample estimates:
mean of x
    3.758
```

Test de Hipótesis (7)

- Tenemos dos tipos de errores cuando realizamos un test de hipótesis
- Error tipo I: es cuando rechazamos la hipótesis nula cuando ésta es cierta.
- Este error es equivalente al nivel de significancia α
- Error tipo II: es cuando la hipótesis nula es falsa pero no tenemos evidencia estadística para rechazarla.
- Para mitigar los errores tipo I generalmente usamos valores de α más pequeños.
- Para mitigar los errores tipo II generalmente trabajamos con muestras más grandes.
- Existe un trade-off entre los errores tipo I y tipo II.

	Retener H_0	Rechazar H_0
H_0 es verdadera	✓	error tipo I
H_1 es verdadera	error tipo II	✓

Bilbiografía I


L. Wasserman *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics, 2005.