


Práctica 5. Distribuciones en el muestreo y estimación

Estadística

Grado en Ingeniería Informática

1. Distribuciones en el muestreo

Las distribuciones en el muestreo nos ayudan a entender cómo varían los resultados de ciertos estadísticos y/o estimadores cuando tomamos diferentes muestras de una misma población. En Estadística, no siempre podemos estudiar a toda una población, así que tomamos muestras y analizamos sus características. Si repitiéramos este proceso muchas veces, evaluásemos el estadístico de interés en cada una de ellas, los valores obtenidos formarían una distribución, llamada distribución en el muestreo. Esta nos permite hacer estimaciones y tomar decisiones sobre la población con base en los datos obtenidos de las muestras.

A lo largo de esta práctica nos centraremos en validar mediante simulación la distribución en el muestreo de los estadísticos que se emplean en las tareas de Inferencia Paramétrica. Además, veremos cómo resolver con  los problemas de estimación puntual y por intervalos que se trabajan en el Tema 5 de la asignatura.

Distribución de la media muestral

Supongamos que disponemos de X_1, \dots, X_n una muestra aleatoria simple de una variable aleatoria $X \in N(\mu, \sigma^2)$. Entonces la media muestral $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ verifica que:

$$\bar{X} \in N\left(\mu, \frac{\sigma^2}{n}\right),$$

es decir, \bar{X} se distribuye según una normal, centrada en la media real de la población μ , y cuya varianza disminuye a medida que aumenta el tamaño de la muestra. Veámoslo con la siguiente simulación.

Ejercicio. Generar 500 muestras de tamaño n (con $n \in \{10, 20, 100, 500\}$) de $X \in N(5, 1)$. Comprobar que la distribución de la media muestral sigue una distribución $N(5, 1/n)$.

```
ns<-500; n<-c(10,20,100,500); mu<-5; sigma<-1
par(mfrow=c(2,2))
for (i in 1:length(n)){
  x<-matrix(rnorm(n[i]*ns,mean=mu,sd=sigma),nrow=n[i],ncol=ns)
  media<-apply(x,2,mean)
  hist(media,freq=FALSE,col="grey",xlab=" ",ylab=" ",main=paste("n=",n[i]),
        xlim=c(4,6))
  lines(density(media),lwd=2,col=2)
  x<-seq(min(media),max(media),by=0.01)
  fx<-dnorm(x,mean=mu,sd=sqrt(sigma^2/n[i]))
  lines(x,fx,lwd=2,col=4)
}
```

Tal y como se ha visto en teoría, existen en Estadística algunas distribuciones que se derivan de la distribución Normal: la χ^2 (chi-cuadrado), la T de Student y la F de Snedecor. A continuación comprobaremos mediante simulación la construcción de estas distribuciones.

Distribución χ^2 (chi-cuadrado)

Si X_1, \dots, X_n es una muestra aleatoria simple de variables $N(0, 1)$, entonces

$$\sum_{i=1}^n X_i^2 \in \chi_n^2.$$

La media y la varianza de esta variable son, por tanto $\mathbb{E}(X) = n$ y $\text{Var}(X) = 2n$. Al parámetro n se le denomina grados de libertad. Además, si las $X_i \in N(\mu, \sigma^2)$, sería necesario tipificar dichas variables y por lo tanto tendríamos que:

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \in \chi_n^2.$$

Veamos un ejemplo práctico de cómo utilizar para generar valores de una χ^2 a partir de la distribución Normal, usando las fórmulas anteriores:

```
ns<-500; n<-c(5,10,15,50)
par(mfrow=c(2,2))
for (i in 1:length(n)){
```

```
x<-matrix(rnorm(n[i]*ns),nrow=n[i],ncol=ns)
chi<-apply(x^2,2,sum)
hist(chi,freq=FALSE,col="grey",xlab=,ylab=,xlim=c(0,80),
      main=paste("Chi cuadrado con n=",n[i]))
lines(density(chi),lwd=2,col=2)
x<-seq(min(chi),max(chi),by=0.01)
lines(x,dchisq(x,df=n[i]),lwd=2,col=3)
lines(x,dnorm(x,mean=n[i],sd=sqrt(2*n[i])),lwd=2,col=4)
}
```

Obsérvese que al aumentar n , como por construcción la distribución χ_n^2 es una suma de n variables χ_1^2 , para n suficientemente grande, puede utilizarse una aproximación normal con media n y varianza $2n$ (aplicación del Teorema Central del Límite).

Distribución T de Student

Si consideramos una variable aleatoria $X \in N(0, 1)$ y otra variable aleatoria $Y \in \chi_n^2$ independientes entre sí, el cociente

$$\frac{X}{\sqrt{\frac{Y}{n}}} \in T_n$$

sigue una distribución t de Student con n grados de libertad. Veamos a continuación como generar valores de una T de Student con 5 grados de libertad a partir de una normal y una χ_5^2 .

```
n<-c(50,100,1000,5000); gl<-5
par(mfrow=c(2,2))
for (i in 1:length(n)){
  est<-rnorm(n[i])/sqrt(rchisq(n[i],df=gl)/gl)
  hist(est,freq=FALSE,col="grey",xlab=" ",ylab=" ",ylim=c(0,0.4),
        main=paste("T de Student con n=",n[i],sep=),xlim=c(-6,6))
  lines(density(est),lwd=2,col=2)
  x<-seq(min(est),max(est),by=0.01)
  fx<-dt(x,df=gl)
  lines(x,fx,lwd=2,col=4)
}
```

Cuando n es suficientemente grande, la distribución T de Student se aproxima a una $N(0, 1)$, como podemos contrastar con el siguiente código:

```
x<-seq(-4,4,by=0.001)
plot(x,dnorm(x,mean=0,sd=1),lwd=2,xlim=c(-4,4),type="l",xlab=,ylab=,
      main="Distribución N(0,1) y t de student")
lines(x,dt(x,df=1),lwd=2,lty=3,col=2)
lines(x,dt(x,df=3),lwd=2,lty=3,col=3)
lines(x,dt(x,df=5),lwd=2,lty=3,col=4)
lines(x,dt(x,df=10),lwd=2,lty=3,col=5)
legend(x="topright",legend=paste("n=",c(1,3,5,10)),fill=2:5)
```

Distribución F de Snedecor

Sea X una variable aleatoria con distribución χ_n^2 e Y otra variable con distribución χ_m^2 , independientes entre sí. Entonces el cociente:

$$\frac{X/n}{Y/m} \in F_{n,m}$$

sigue una distribución F de Snedecor con n y m grados de libertad. Podemos representar las funciones de densidad de las F de Snedecor con distintos grados de libertad:

```
ns<-500; n<-c(5,20); m<-c(5,20)
par(mfrow=c(2,2))
for (i in 1:length(n)){
  for (j in 1:length(m)){
    x<-rchisq(ns,df=n[i])
    y<-rchisq(ns,df=m[j])
    est<-(x/n[i])/(y/m[j])
    hist(est,freq=FALSE,col="grey",xlab=" ",ylab=" ",
          main=paste("F de Snedecor con n=",n[i],"y m=",m[j],sep=" "),
          xlim=c(0,10),ylim=c(0,0.9))
    lines(density(est),lwd=2,col=2)
    x<-seq(min(est),max(est),by=0.01)
    fx<-df(x,df1=n[i],df2=m[j])
    lines(x,fx,lwd=2,col=4)
  }
}
```

Una vez validada mediante simulación la construcción de estas distribuciones, vamos a ver su aplicación a una serie de estadísticos que luego se emplean en las tareas de Inferencia.

Distribución de la media muestral, cuando la varianza poblacional no es conocida

La distribución t de Student nos describe la distribución de la media muestral cuando la varianza poblacional no es conocida. El estadístico T de Student sigue esta distribución:

$$\frac{\bar{X} - \mu}{S/\sqrt{n-1}} \in T_{n-1}, \quad \text{o bien} \quad \frac{\bar{X} - \mu}{S_c/\sqrt{n}} \in T_{n-1},$$

donde S^2 representa la varianza muestral y S_c^2 representa la cuasivarianza.

Ejercicio. Generar 500 muestras de tamaño $n \in \{10, 20, 100, 500\}$ de $X \in N(5, 1)$. Para cada muestra calcular $\frac{\bar{X} - \mu}{S_c/\sqrt{n}}$ y comprobar que este estadístico sigue una distribución t de Student con $n - 1$ grados de libertad.

```
ns<-500; n<-c(10,20,100,500); mu<-5; sigma<-1
par(mfrow=c(2,2))
for (i in 1:length(n)){
  x<-matrix(rnorm(n[i]*ns,mean=mu,sd=sigma),nrow=n[i],ncol=ns)
  media<-apply(x,2,mean)
  cvar<-apply(x,2,var)
  est<-(media-mu)/sqrt(cvar/n[i])
  hist(est,freq=FALSE,col="grey",xlab=" ",ylab=" ",main=paste("n=",n[i]),
        xlim=c(-6,6))
  lines(density(est),lwd=2,col=2)
  x<-seq(min(est),max(est),by=0.01)
  fx<-dt(x,df=n[i]-1)
  lines(x,fx,lwd=2,col=4)
}
```

Distribución de la varianza muestral y de la cuasi-varianza

El Teorema de Fisher establece que si X_1, \dots, X_n es una muestra aleatoria simple de variables normales con varianza σ^2 , entonces \bar{X} y S^2 son independientes y además:

$$\frac{nS^2}{\sigma^2} \in \chi_{n-1}^2.$$

En términos de la cuasivarianza S_c^2 (recordar que $S_c^2 = \frac{n}{n-1} S^2$), tendríamos que:

$$\frac{(n-1)S_c^2}{\sigma^2} \in \chi_{n-1}^2.$$

Ejercicio. Generar 500 muestras de tamaño $n \in \{5, 10, 20, 50\}$ de $X \in N(3, 0.5^2)$. Calcular el estadístico $\frac{(n-1)S_c^2}{\sigma^2}$ para cada una de las 500 muestras y comprobar que sigue una distribución χ_{n-1}^2 .

```
ns<-500; n<-c(5,10,20,50); mu<-3; sigma<-0.5
par(mfrow=c(2,2))
for (i in 1:length(n)){
  x<-matrix(rnorm(n[i]*ns,mean=mu,sd=sigma),nrow=n[i],ncol=ns)
  cvar<-apply(x,2,var)
  est<-(n[i]-1)*cvar/sigma^2
  hist(est,freq=FALSE,col="grey",xlab=" ",ylab=" ",main=paste("n=",n[i]),
        xlim=c(0,90))
  lines(density(est),lwd=2,col=2)
  x<-seq(min(est),max(est),by=0.01)
  fx<-dchisq(x,df=n[i]-1)
  lines(x,fx,lwd=2,col=4)
}
```

Distribución del cociente de varianzas

Al igual que comparamos medias, también podemos comparar varianzas. En el caso de las medias, por ser parámetros de localización, la comparación se hacía por medio de una diferencia. En el caso de las varianzas, que son parámetros de escala, su comparación se realizará por medio de un cociente.

Dadas dos variables X e Y , por el Teorema de Fisher teníamos que:

$$\frac{(n-1)S_{c,X}^2}{\sigma_X^2} \in \chi_{n-1}^2 \quad \text{y} \quad \frac{(m-1)S_{c,Y}^2}{\sigma_Y^2} \in \chi_{m-1}^2$$

Entonces, el cociente de varianzas tiene distribución F de Snedecor:

$$\frac{S_{c,X}^2/\sigma_X^2}{S_{c,Y}^2/\sigma_Y^2} \in F_{n-1,m-1}.$$

Ejercicio. Generar 500 muestras de tamaño $n = 100$ de $X \in N(0, 1)$ y 500 muestras de tamaño $m = 200$ de $Y \in N(1, 0.5^2)$. Comprobar que $\frac{S_{c,X}^2/\sigma_X^2}{S_{c,Y}^2/\sigma_Y^2}$ sigue una distribución $F_{n-1,m-1}$.

```
ns<-500; n<-100; muX<-0; sigmaX<-1; m<-200; muY<-1; sigmaY<-0.5
x<-matrix(rnorm(n*ns,mean=muX,sd=sigmaX),nrow=n,ncol=ns)
```

```
cvarX<-apply(x,2,var)
y<-matrix(rnorm(m*ns,mean=muY,sd=sigmaY),nrow=m,ncol=ns)
cvarY<-apply(y,2,var)
est<-(cvarX/sigmaX^2)/(cvarY/sigmaY^2)
hist(est,freq=FALSE,col="grey",xlab=" ",ylab=" ",ylim=c(0,3),
      main="Comparación de varianzas",xlim=c(0.5,2))
lines(density(est),lwd=2,col=2)
x<-seq(min(est),max(est),by=0.01)
fx<-df(x,df1=n-1,df2=m-1)
lines(x,fx,lwd=2,col=4)
```

Otro aspecto importante que estudiamos en la teoría es la aproximación asintótica de ciertas distribuciones mediante la aplicación del Teorema Central del Límite. Una de las aproximaciones más empleadas es la de la distribución Binomial mediante la distribución Normal. En el siguiente ejercicio ilustramos esta cuestión mediante simulación.


Ejercicio. Generar 500 muestras de tamaño 100 de una $Bi(1, 0.7)$ y calcular las 500 proporciones muestrales asociadas. Dibujar el histograma obtenido a partir de las proporciones calculadas y compararlo con la función de densidad de una distribución $N(0.7, \frac{0.7 \cdot 0.3}{100})$.

```
ns<-500; n<-100; p<-0.7
x<-matrix(rbinom(n=n*ns,size=1,prob=p),nrow=n,ncol=ns)
prop<-apply(x,2,mean)
hist(prop,freq=FALSE,col="grey",xlab=" ",ylab=" ",main="Proporción muestral")
lines(density(prop),lwd=2,col=2)
x<-seq(min(prop),max(prop),by=0.01)
fx<-dnorm(x,mean=p,sd=sqrt(p*(1-p)/n))
lines(x,fx,lwd=2,col=4)
```

2. Estimación puntual y por intervalos de confianza

En Estadística, cuando queremos aproximar un valor desconocido de una población, utilizamos la estimación puntual y/o por intervalos de confianza. La estimación puntual consiste en obtener un único valor a partir de una muestra para aproximar un parámetro poblacional desconocido, como la media o la proporción. Sin embargo, proporcionar un único valor sin ninguna medida del posible error, puede no ser siempre lo más adecuado, por lo que se usa la estimación por intervalos de confianza, que proporciona un rango de valores dentro del cual es probable que se encuentre el ver-

dadero parámetro poblacional, con un cierto nivel de confianza. Esto nos ayuda a tomar decisiones con mayor seguridad al considerar la incertidumbre de la muestra.

A lo largo de esta sección emplearemos comandos que se encuentran en el paquete de  LearningStats, por lo que es necesario que se instale y se cargue en la sesión actual:

```
install.packages("LearningStats")  
library(LearningStats)
```

Ejercicio. En un estudio sobre el tiempo de respuesta de un sistema informático, se ha registrado una muestra aleatoria de 10 mediciones (en milisegundos):

12.5 13.2 11.8 12.9 13.5 12.2 11.9 12.7 13.1 12.4

Sabiendo que los tiempos de respuesta siguen una distribución Normal, obtén un intervalo de confianza para el tiempo medio de respuesta a un nivel de confianza del 92 %.

En primer lugar definimos la variable de interés X = “Tiempo de respuesta de un sistema informático” de la que sabemos que $X \in N(\mu, \sigma^2)$ donde media y varianza poblacional son desconocidas.

```
x=c(12.5, 13.2, 11.8, 12.9, 13.5, 12.2, 11.9, 12.7, 13.1, 12.4)  
n=length(x)  
med=mean(x)  
vari=var(x)*(n-1)/n  
cuasivar=var(x)  
Mean.CI(x,conf.level = 0.92)
```

Ejercicio. Un equipo de analistas de sistemas desea comparar los tiempos de respuesta de dos servidores distintos. Se toman dos muestras aleatorias de tiempos de respuesta (en milisegundos):

- **Servidor A:** $n_A = 15$, media muestral $\bar{x}_A = 120$ ms, varianza muestral $s_A^2 = 25$ ms².
- **Servidor B:** $n_B = 12$, media muestral $\bar{x}_B = 130$ ms, varianza muestral $s_B^2 = 30$ ms².

Asumiendo que los tiempos de respuesta siguen distribuciones normales e independientes, construye un intervalo de confianza del 95 % para la diferencia de los tiempos medios de respuesta de ambos servidores.

En este caso tenemos dos variables, X_A = “Tiempo de respuesta del servidor A”, que sigue una distribución Normal de parámetros desconocidos, es decir, $X_A \in N(\mu_A, \sigma_A^2)$, y X_B = “Tiempo de respuesta del servidor B”, que sigue una distribución Normal de parámetros desconocidos, es decir, $X_B \in N(\mu_B, \sigma_B^2)$. Además nos indican que ambas variables son independientes entre sí.

```
nA=15; medA=120; sA2=25  
nB=12; medB=130; sB2=30
```



```
diffmean.CI(x1=medA,x2=medB,n1=nA,n2=nB,s1=sqrt(sA2),s2=sqrt(sB2),  
conf.level=0.95)
```

Ejercicio. Una empresa de *software* ha desarrollado una nueva aplicación móvil y quiere estimar la proporción de usuarios que están satisfechos con el producto. Se realiza una encuesta a 200 usuarios seleccionados aleatoriamente y 150 de ellos indican estar satisfechos con la aplicación. Obtén un intervalo de confianza del 95 % para la proporción poblacional de usuarios satisfechos.

En primer lugar tenemos que darnos cuenta que se trata de un problema sobre la proporción. La variable de interés es $X = \text{“¿Está satisfecho?”}$ que sigue una distribución $Ber(p)$, donde p es la proporción poblacional desconocida que queremos estimar.

```
n=200  
 exitos=150  
 hatp= exitos/n  
 proportion.CI(x= exitos, n=n, conf.level=0.95)
```

Ejercicio. Un departamento de ciberseguridad está evaluando la efectividad de dos métodos de autenticación en la detección de accesos no autorizados. Se probaron dos sistemas diferentes en condiciones controladas y se registró cuántos accesos sospechosos fueron correctamente detectados en cada caso:

- **Método 1:** se realizaron $n_1 = 400$ intentos de acceso, de los cuales $x_1 = 280$ fueron correctamente detectados como sospechosos.
- **Método 2:** se realizaron $n_2 = 350$ intentos de acceso, de los cuales $x_2 = 245$ fueron correctamente detectados como sospechosos.

Construye un intervalo de confianza a un nivel de confianza del 95 % para la diferencia de proporciones entre el Método 1 y el Método 2.

En este ejercicio volvemos a tener dos variables, pero que nada tienen que ver con las del ejercicio anterior: $X_1 = \text{“Hubo acceso sospechoso según el Método 1”}$, que sigue una distribución Bernoulli de parámetro p_1 desconocido, esto es, $X_1 \in Ber(p_1)$, y $X_2 = \text{“Hubo acceso sospechoso según el Método 2”}$, y de manera análoga a la primera, $X_2 \in Ber(p_2)$, con p_2 la proporción poblacional desconocida.

```
n1=400  
 exitos1=280  
 n2=350  
 exitos2=245  
 diffproportion.CI(x1= exitos1, x2= exitos2, n1=n1, n2=n2, conf.level = 0.95)
```

Ejercicio. Un equipo de ingenieros informáticos está evaluando la variabilidad en los tiempos de ejecución de un algoritmo. Se toma una muestra aleatoria de 12 ejecuciones (en segundos), obteniendo los siguientes tiempos:

2.1 2.4 2.0 2.5 2.3 2.2 2.6 2.1 2.3 2.4 2.2 2.5

Se asume que los tiempos de ejecución siguen una distribución Normal, calcula una estimación puntual de la media y de la varianza y obtén además los correspondientes intervalos de confianza a un nivel de confianza del 97 %.

En primer lugar definimos la variable de interés, X ="Tiempo de ejecución del algoritmo" que nos dicen que sigue una distribución Normal, de la cual no conocemos ninguno de sus parámetros teóricos, es decir, $X \in N(\mu, \sigma^2)$. Queremos estimar tanto puntualmente como mediante intervalos ambos parámetros.

```
x=c(2.1, 2.4, 2.0, 2.5, 2.3, 2.2, 2.6, 2.1, 2.3, 2.4, 2.2, 2.5)
n=length(x)
med=mean(x)
vari=var(x)*(n-1)/n
cuasivar=var(x)
# IC para la media
Mean.CI(x,conf.level = 0.97)
# IC para la varianza
variance.CI(x,conf.level = 0.97)
```

Ejercicio. Una empresa de desarrollo de software está probando dos algoritmos de búsqueda en una base de datos grande. Para el algoritmo A, se midieron los tiempos de respuesta (en milisegundos):

120, 130, 115, 122, 128, 118, 125, 121

Para el algoritmo B, se obtuvieron:

110, 150, 135, 140, 138, 132, 147, 130, 145, 136

Se asume que los tiempos de respuesta de ambos algoritmos siguen una distribución normal. Construye un intervalo de confianza del 99 % para el cociente de varianzas de los tiempos entre ambos algoritmos.

```
x1=c(120, 130, 115, 122, 128, 118, 125, 121)
x2=c(110, 150, 135, 140, 138, 132, 147, 130, 145, 136)
diffvariance.CI(x1=x1,x2=x2,conf.level = 0.99)
```