FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

EYP1113 - Probabilidad y Estadística

Laboratorio 05

Pilar Tello Hernández
pitello@uc.cl

Facultad de Matemáticas
Departamento de Estadística
Pontificia Universidad Católica de Chile

Segundo Semestre 2021

set.seed()

La función `set.seed()` permite fijar una semilla que establece el número inicial utilizado para generar una secuencia de números aleatorios, esto sirve para asegurar obtener el mismo resultado si se comienza con la misma semilla cada vez que ejecuta el mismo proceso.

Ejemplo:

```
set.seed(1113)
x <- rnorm(10,mean=10,sd=2)
x
```

Distribución Hipergeométrica

En un lote de tamaño N tengo m objetos defectuosos y $N - m$ que no son defectuosos, obtengo una muestra aleatoria de tamaño n y luego la probabilidad de que x objetos sean defectuosos está dada por la función de probabilidad de la distribución hipergeométrica. Donde:

- X: cantidad de objetos defectuosos de la muestra.
- X=0,1,...,min(m,n)

La función de probabilidad de esta función está dada por:

$$p_X(x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$$

Distribución Hipergeométrica

En `R`

En R se define como una urna con `m` bolas blancas y `n` bolas negras. Se realiza una extracción de tamaño `k` y `x` representa el número de bolas blancas extraídas. En este caso:

- $N = \texttt{m+n}$
- $n = \texttt{k}$

Aquí el X: cantidad de bolas blancas que obtengo y $X = 0, 1, ..., min(\texttt{m,k})$. Los comandos en `R` correspondientes a esta distribución son:

```
dhyper(x,m,n,k)
phyper(q,m,n,k)
qhyper(p,m,n,k)
rhyper(nn,m,n,k)
```

La media teórica en este caso es $E(X) = \texttt{k} \cdot \texttt{p}$, con $p = \frac{\texttt{m}}{\texttt{m+n}}$ y la varianza teórica es $Var(X) = \texttt{k} \cdot \texttt{p} \cdot \texttt{(1-p)} \cdot \frac{\texttt{m+n-k}}{\texttt{m+n-1}}$.

Medidas Descriptivas Teóricas vs Empíricas

Una variable aleatoria puede ser descrita totalmente por su **función de distribución de probabilidad o de densidad**, o bien por su **función de distribución de probabilidad acumulada**.

Sin embargo, en la práctica la forma exacta puede no ser totalmente conocida.

En tales casos se requieren ciertas "medidas" para tener una idea de la forma de la distribución:

- Medidas Centrales
- Medidas de Posición
- Medidas de Dispersión
- Medidas de Asimetrías y Forma

Medidas Descriptivas Teóricas vs Empíricas

Una variable aleatoria puede ser descrita totalmente por su **función de distribución de probabilidad o de densidad**, o bien por su **función de distribución de probabilidad acumulada**.

Sin embargo, en la práctica la forma exacta puede no ser totalmente conocida.

En tales casos se requieren ciertas “medidas” para tener una idea de la forma de la distribución:

- Medidas Centrales
- Medidas de Posición
- Medidas de Dispersión
- Medidas de Asimetrías y Forma

Medidas Descriptivas Teóricas vs Empíricas

Para este laboratorio trabajaremos con el siguiente ejemplo:

Hay una urna con 17 bolas blancas y 23 negras, si se extraen 15 bolas al azar, ¿cuál es la distribución de las bolas blancas extraídas?

$$X \sim \text{Hipergeométrica}(m = 17, n = 23, k = 15)$$

Vamos a simular una muestra aleatoria de tamaño $n = 120$ en R.

```
nmuestra=120
m=17
n=23
k=15
set.seed(1113)
X=rhyper(nn=nmuestra,m=m,n=n,k=k)
?rhyper
```

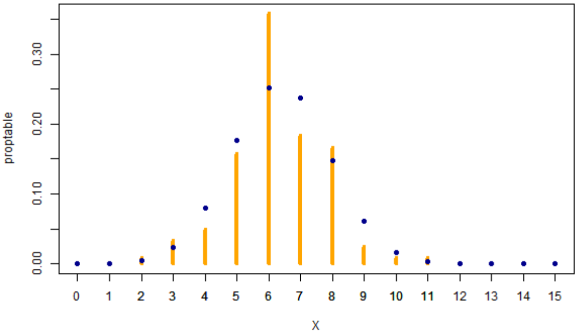
Medidas Descriptivas Teóricas vs Empíricas

En primera instancia vamos a graficar la distribución empírica vs la teórica de esta variable aleatoria discreta. Esto se hace de manera distinta a lo visto para las distribuciones contínuas

```
maximo=min(m,k);maximo
table(X)
prop.table(table(X))
proptable=prop.table(table(X))
sum(proptable)
plot(proptable,xlim=c(0,maximo),col="orange",lwd=4)
axis(side=1,at=x)
x=0:maximo;x
dhyper(x,m=m,n=n,k=k)
sum(dhyper(x,m=m,n=n,k=k))
points(x,dhyper(x,m=m,n=n,k=k),lwd=10,pch=16,col="darkblue")
```

La función `prop.table(X)` divide a la tabla por la suma total de ésta. Así en este ejemplo `sum(proptable)` debe ser 1, obteniendo las probabilidades empíricas.

Medidas Descriptivas Teóricas vs Empíricas



Medidas Descriptivas Teóricas vs Empíricas

Valor esperado (media)

Para una variable aleatoria X se define el valor esperado, μ_x, como:

$$\mu_x = E(X) = \begin{cases} \displaystyle\sum_{x \in \Theta_X} x \cdot p_X(x), \text{ caso discreto} \\ \displaystyle\int_{-\infty}^{\infty} x \cdot f_X(x) dx, \text{ caso contínuo} \end{cases}$$

En R, la función `mean(,na.rm=TRUE)` la calcula de manera empírica.

```
# Media muestral
mean(X)
abline(v=mean(X),col="red",lty=2,lwd=2)
# Media teórica
p=m/(m+n)
k*p
abline(v=k*p,col="darkgreen",lty=2,lwd=2)
```

Medidas Descriptivas Teóricas vs Empíricas

Otras medidas de centro son:

- **La Moda**: Valor más frecuente o con mayor probabilidad.

```
# Moda muestral
library(modeest)
mlv(X)
# Moda teórica
dhyper(x,m=m,n=n,k=k)==max(dhyper(x,m=m,n=n,k=k))
x[dhyper(x,m=m,n=n,k=k)==max(dhyper(x,m=m,n=n,k=k))]
```

- **La Mediana**: Sea X_{med} el valor que toma la mediana, entonces:

$$F_X(X_{med}) = 0{,}5$$

```
# Mediana muestral
median(X)
# Mediana teórica
qhyper(0.5,m=m,n=n,k=k)
```

Medidas Descriptivas Teóricas vs Empíricas

Esperanza Matemática

La noción del valor esperado como un promedio ponderado puede ser generalizado para funciones de la variable aleatoria X.

Dada una función $g(X)$, entonces el valor esperado de esta puede ser obtenido como:

$$E(g(X)) = \begin{cases} \sum_{x \in \Theta_X} g(x) \cdot p_X(x), \text{ caso discreto} \\ \int_{-\infty}^{\infty} g(x) \cdot f_X(x) dx, \text{ caso contínuo} \end{cases}$$

```
# Esperanza matemática de g(X)=X^2
g=function(X){
X^2
}
mean(g(X))
```

Medidas Descriptivas Teóricas vs Empíricas

Percentil: Valor en los reales, llamemos X_p, que es superior al $p \times 100\,\%$ de la información.

$$F_X(x_p) = p$$

En `R` las siguientes funciones entregan percentiles empíricos.

```
‘‘quantile": Percentil
‘‘min": Mínimo
‘‘max": Máximo
```

```
# Percentiles muestrales
quantile(X,seq(from=0,to=1,by=0.1))
# Percentiles teóricos
qhyper(seq(from=0,to=1,by=0.1),m=m,n=n,k=k)
```

Medidas Descriptivas Teóricas vs Empíricas

Varianza y desviación estándar

$$\sigma_X^2 = Var(X) = E[(X-\mu_X)^2] = \begin{cases} \sum_{x\in\Theta_X}(x-\mu_X)^2\cdot p_X(x), \text{ caso discreto} \\ \int_{-\infty}^{\infty}(x-\mu_X)^2\cdot f_X(x)dx, \text{ caso contínuo} \end{cases}$$

En `R`, la función **var()** la calcula.

La desviación estándar es la raíz de la varianza que vendría siendo σ_X y en `R` se calcula con `sd()`.

```
# Varianza muestral
var(X)
# Varianza teórica
k*p*(1-p)*(m+n-k)/(m+n-1)

# Desviación estándar muestral
sd(X)
# Desviación estándar teórica
sqrt(k*p*(1-p)*(m+n-k)/(m+n-1))
```

Medidas Descriptivas Teóricas vs Empíricas

Rango: Min - Max

```
# Rango muestral
Rango=function(X){
max(X)-min(X)
}
Rango(X)
range(X)
range(X)[2]-range(X)[1]
# Rango teórico
maximo-0
```

Medidas Descriptivas Teóricas vs Empíricas

Rango Intercuartil: $X_{0,75} - X_{0,25}$

```
# Rango intercuartílico muestral
IQR=function(X){
quantile(X,0.75)-quantile(X,0.25)
}
IQR(X)
# Rango intercuartílico teórico
qhyper(0.75,m=m,n=n,k=k)-qhyper(0.25,m=m,n=n,k=k)
```

Medidas Descriptivas Teóricas vs Empíricas

En términos de dimensionalidad, es conveniente utilizar la desviación estándar, es decir,

$$\sigma_X = \sqrt{Var(X)}$$

Ahora, si $\mu_X > 0$, una medida adimensional de la variabilidad es el coeficiente de variación (COV):

$$\delta_X = \frac{\sigma_X}{\mu_X}$$

```
# Coeficiente de variación=sigma/mu
# Coeficiente de variación muestral
sd(X)/mean(X)
# Coeficiente de variación teórico
sqrt(k*p*(1-p)*(m+n-k)/(m+n-1))/(k*p)
```

Medidas Descriptivas Teóricas vs Empíricas

Se define una medida de asimetría (skewness) como al tercer momento central:

$$E[(X-\mu_X)^3] = \begin{cases} \sum_{x_i \in \Theta_X} (x_i - \mu_X)^3 \cdot p_X(x_i), \text{ caso discreto} \\ \int_{-\infty}^{\infty} (x - \mu_X)^3 \cdot f_X(x) dx, \text{ caso contínuo} \end{cases}$$

Una medida conveniente es el coeficiente de asimetría que se define como:

$$\theta_X = \frac{E[(X-\mu_X)^3]}{\sigma_X^3}$$

Para el cálculo de skewness en `R` se utilizará la función `skewness` de la librería `moments`.

```
install.packages("moments")
library(moments)
# Coeficiente de asimetría muestral
skewness(X)
```

Medidas Descriptivas Teóricas vs Empíricas

El cuarto momento central se conoce como la kurtosis:

$$E[(X-\mu_X)^3] = \begin{cases} \sum_{x_i \in \Theta_X} (x_i - \mu_X)^4 \cdot p_X(x_i), \text{ caso discreto} \\ \int_{-\infty}^{\infty} (x - \mu_X)^4 \cdot f_X(x) dx, \text{ caso contínuo} \end{cases}$$

que es una medida del "apuntamiento" o "achatamiento" de la distribución de probabilidad o de densidad. Usualmente se prefiere el coeficiente de kurtosis:

$$K_X = \frac{E[(X-\mu_X)^4]}{\sigma_X^4} - 3$$

Para el cálculo de kurtosis en `R` se utilizará la función `kurtosis` de la librería `moments` a la que posteriormente hay que restarle 3 por definición.

```
install.packages("moments")
library(moments)
# Kurtosis muestral
kurtosis(X)-3
```

Medidas Descriptivas Teóricas vs Empíricas

Cuando hay dos variables aleatorias X e Y, puede haber una relación entre las variables.

En particular, la covarianza definida como:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(X \cdot Y) - \mu_X \cdot \mu_Y$$

mide el grado de asociación lineal entre dos variables, pero es preferible su normalización llamada correlación para poder cuantificar la magnitud de la relación:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Este coeficiente toma valores en el intervalo (-1,1).

En `R`, las funciones `cov()` y `cor()` entregan ambas medidas.

Medidas Descriptivas Teóricas vs Empíricas

Para esto último necesitamos una segunda variable que definiremos como $Y \sim \text{Binomial}(k, p)$

```
set.seed(1113)
Y=rbinom(120,size=k,prob=p)
maximo=k;maximo
table(Y)
prop.table(table(Y))
proptable=prop.table(table(Y))
sum(proptable)
plot(proptable,xlim=c(0,maximo),col="orange",lwd=4)
axis(1,at=x)
x=0:maximo;x
dbinom(x,size=k,prob=p)
sum(dbinom(x,size=k,prob=p))
points(x,dbinom(x,size=k,prob=p),lwd=10,pch=16,col="darkblue")
plot(X,Y,pch=16)

# Covarianza muestral
cov(X,Y)

# Correlación muestral
cor(X,Y)
cov(X,Y)/(sd(X)*sd(Y))
```

Actividad

Replique los ejercicios realizados con las simulando las siguientes muestras:

- Muestra de tamaño $n = 200$ proveniente de una distribución $\mathit{Binomial}(n = 10, p = 0{,}3)$
- Muestra de tamaño $n = 400$ proveniente de una distribución $\mathit{Normal}(\mu = 650, \sigma = 40)$