Propiedades de la funcion de distribucion empirica

Propiedades de la Función de distribución Empírica:

a. $\hat{F}_n$ es creciente de 0 hasta 1.
b. $\hat{F}_n$ es una función escalonada con saltos en los distintos valores de $X_1, X_2, ..., X_n$.
c. $E[\hat{F}_n(t)] = F(t)$.
d. $\text{var}[\hat{F}_n(t)] = F(t)\,[1 - F(t)] / n$.
e. $F_n(x) \rightarrow F(x)$ as $n \rightarrow \infty$ (con probabilidad 1). (Ley de los grandes números).

La Ley debil de los grandes numeros es similar a convergencia en probabilidad.
La ley fuerte de los grandes numeros es similar a convergencia casi en todas partes.
Según el teorema de Givlenko-Cantelli la convergencia es uniforme.

f. $\sqrt{n}\,\dfrac{F_n(x) - F(x)}{\sqrt{F(x)(1-F(x))}}$ se distribuye aproximadamente como una normal estándar cuando n es grande (Teorema del Limite Central).

Para las pruebas ver Rohatgi. “An introduction to probability and Mathematical Statistics”.

Propiedad: Sea g cualquier función de valor real y $\hat{F}_n$ la función de distribución empírica basada en la muestra $x_1, x_2, \ldots. x_n$. Entonces

$$\int g(x)d\hat{F}_n(x) = \frac{\sum_{i=1}^{n} g(x_i)}{n}$$

El método de estimación “plug-in” (por sustitución)

El estimador “plug-in” del parámetro θ=T(F) es definido por

$$\hat{\theta} = T(\hat{F}_n)$$

en otras palabras el estimador de la función θ=T(F) de la función de distribución F es la misma función evaluada en la distribución empírica.

Ejemplo. Probar que el estimador “plug-in” de $\theta = E_F(X)$ es $\overline{X}$.

Puesto que $\theta = E_F(X) = \int x dF(x)$ entonces el estimador “plug-in” será

$$\hat{\theta} = E_{\hat{F}_n}(x) = \int x d\hat{F}_n(x) = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$$

El estimador “plug-in” es bueno si la única información que se tiene de la distribución F es la muestra tomada. Si se tiene información adicional acerca de F, como por ejemplo que es Binomial, Poisson, Normal etc. entonces el estimador “plug-in” pierde algo su importancia.

El error estándar de la media muestral

Asumamos que la variable aleatoria X tiene una distribución F con valor esperado $\mu_F = E_F(X)$ y con varianza $\sigma_F^2 = VAR_F(X) = E_F[(x-\mu_F)^2]$
Si se toman la muestra aleatoria $X_1, \ldots, X_n$, entonces la media de la muestra

$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ tiene media $E(\bar{X}) = \mu_F$ y varianza $VAR(\bar{X}) = \frac{\sigma_F^2}{n}$.

La prueba se basa en la linealidad del valor esperado y en la independencia de las variables aleatorias X_i’s
El error estándar de la media muestral representado por $se_F(\bar{X})$, o simplemente $se(\bar{X})$, es la raíz cuadrada de la varianza de $\bar{X}$. Esto es,

$$se_F(\bar{X}) = \sigma_F / \sqrt{n}$$

En muchos textos el término “error estándar” es usado para representar un estimado de la desviación estándar de un estadístico.

El Teorema del Limite Central (TLC). Asumiendo ciertas condiciones bien generales acerca de la distribución F entonces la distribución de la media muestral $\bar{X}$ será aproximadamente normal cuando el tamaño de muestra n es bastante grande. Es decir,

$$\bar{X} \approx N(\mu_F, \sigma_F^2 / n)$$

Usando una tabla de la normal estándar se obtiene

$$\Pr ob(|\bar{X} - \mu_F| < \frac{\sigma_F}{\sqrt{n}}) \approx .683 \quad \text{y} \quad \Pr ob(|\bar{X} - \mu_F| < \frac{2\sigma_F}{\sqrt{n}}) \approx .954$$

Así que aproximadamente se espera que $\bar{X}$ se desvíe de la media poblacional μ_F a una distancia menor de una desviación estándar un 68.3% de las veces y se desvíe en menos de dos desviaciones estándar un 95.4% del tiempo.
El error estándar de la media y de un estimado en general da una buena idea de su precisión.

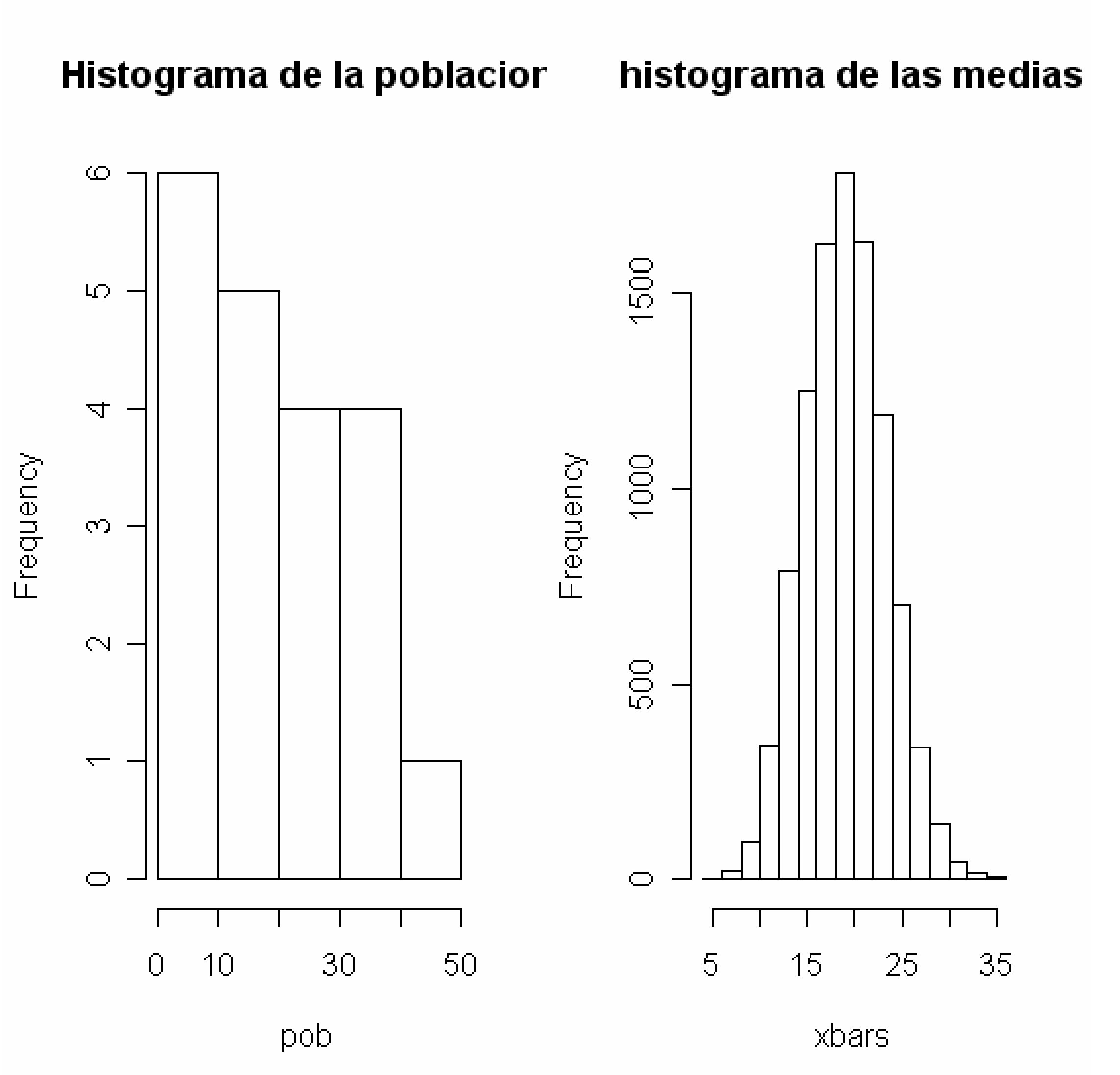
Ejemplo: El siguiente programa en R ilustra el Teorema del limite central usando muestras de tamaño 9 de una población consistente de 20 elementos

```
> pob=c(2,5,9,12,17,21,24,33,37,45,39,34,27,23,15,13,12,8,4,2)
> mean(pob)
```

```
[1] 19.1
> var(pob)
[1] 173.8842
> muestras=matrix(0,10000,9)
> for(i in 1:10000){muestras[i,]=sample(pob,9,replace=T)}
> xbars=apply(muestras,1,mean)
> mean(xbars)
[1] 19.06872
> var(xbars)
[1] 18.434
> var(pob)/9
[1] 19.32047
> # Haciendo los histogramas de la poblacion y de las medias muestrales
> par(mfrow=c(1,2))
> hist(pob,main="Histograma de la poblacion")
> hist(xbars,main="histograma de las medias")
> pob=c(2,5,9,12,17,21,24,33,37,45,39,34,27,23,15,13,12,8,4,2)
> mean(pob)
[1] 19.1

> var(pob)
[1] 173.8842
>  muestras=matrix(0,1000,9)
> for(i in 1:1000){muestras[i,]=sample(pob,9,replace=T)}
> xbars=apply(muestras,1,mean)
> hist(xbars)
> means(xbars)
 [1] 19.21922
> var(xbars)
[1] 17.75190
> var(pob)/9
[1] 19.32047
> # Haciendo los histogramas de la poblacion y de las medias muestrales
> par(mfrow=c(1,2))
> hist(pob,main="Histograma de la poblacion")
> hist(xbars,main="histograma de las medias")
```

Los histogramas aparecen en la siguiente figura

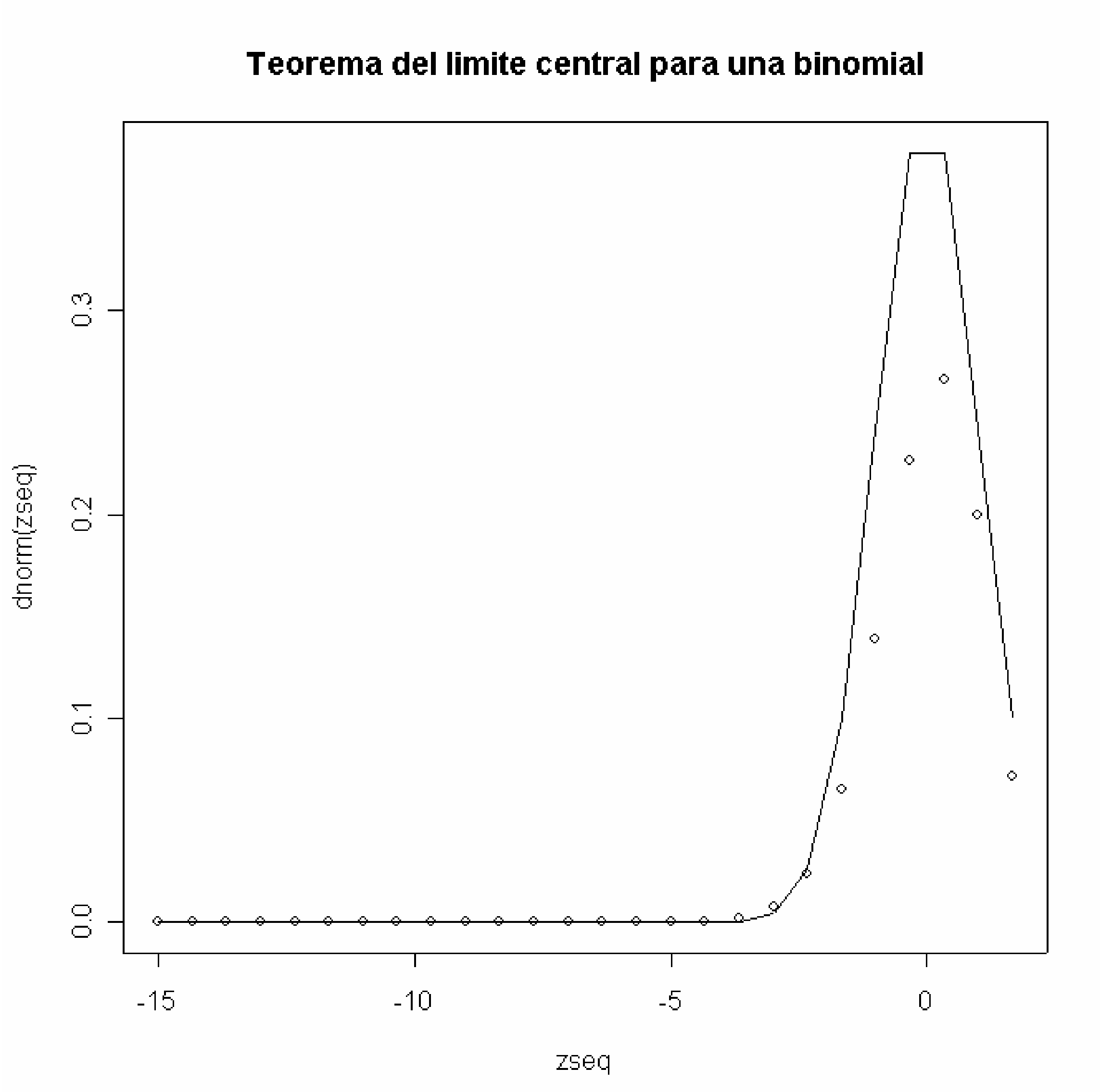


Ejemplo. El TLC aplicado a experimentos de Bernoulli, es decir experimentos con solo dos resultados posibles: Éxito (1) y fracaso (0) puede dar malos resultados si la probabilidad de éxito está cerca de 0 o cerca de 1.

En este caso $\Pr ob(X=1)=p$ y $\Pr ob(X=0)=1-p$. Luego, $\mu_F = p$ y $\sigma_F^2 = p(1-p)$. Así que por el TLC $\hat{p}=\frac{\sum_{i=1}^{n} X_i}{n}$ se distribuye aproximadamente como una normal con media p y varianza p(1-p)/n.

```
> seqx=0:25
> dbinom(seqx,25,.9)
 [1] 1.000000e-25 2.250000e-23 2.430000e-21 1.676700e-19 8.299665e-18
 [6] 3.137273e-16 9.411820e-15 2.299173e-13 4.655826e-12 7.914904e-11
```

```
[11] 1.139746e-09 1.398779e-08 1.468718e-07 1.321846e-06 1.019710e-05
[16] 6.730087e-05 3.785674e-04 1.803762e-03 7.215049e-03 2.392358e-02
[21] 6.459368e-02 1.384150e-01 2.264973e-01 2.658881e-01 1.994161e-01
[26] 7.178980e-02
> zseq=(seqx-22.5)/1.5
> zseq
 [1] -15.0000000 -14.3333333 -13.6666667 -13.0000000 -12.3333333 -11.6666667
 [7] -11.0000000 -10.3333333  -9.6666667  -9.0000000  -8.3333333  -7.6666667
[13]  -7.0000000  -6.3333333  -5.6666667  -5.0000000  -4.3333333  -3.6666667
[19]  -3.0000000  -2.3333333  -1.6666667  -1.0000000  -0.3333333   0.3333333
[25]   1.0000000   1.6666667
> plot(zseq,dnorm(zseq),type="l")
> points(zseq,dbinom(seqx,25,.9))
> plot(zseq,dnorm(zseq),type="l")
> title("Teorema del limite central para una binomial")
> points(zseq,dbinom(seqx,25,.9))
>
```


Teorema del limite central para una binomial
dnorm(zseq)
0.0
0.1
0.2
0.3
-15
-10
-5
0
zseq