Estadística Inferencial

Clifford Torres

Facultad de Ciencias (UNI)

October 3, 2019

Introducción

- Considere la determinación de la distribución de muestreo de la media muestral $\bar{X}$.
- Suponga que se toma una muestra aleatoria de tamaño n de una población normal con media μ, y varianza σ^2.
- Cada observación en esta muestra (por ejemplo, X_1, X_2,, X_n) es una variable aleatoria distribuida normal e independiente, con media μ y varianza σ^2.
- Entonces, por la propiedad reproductiva de la distribución normal, se concluye que la media muestral

$$\bar{X} = \frac{X_1 + X_2 + \cdots X_n}{n}$$

tiene una distribución normal con media

$$\mu_{\bar{X}} = E[\bar{X}] = \frac{\mu + \mu + \cdots + \mu}{n} = \mu$$

y varianza

$$\sigma^2_{\bar{X}} = var[\bar{X}] = \frac{\sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Introducción

- Muchos procedimientos estadísticos comunes requieren que los datos sean aproximadamente normales, por lo que en muchos casos se acude al **Teorema del Límite Central** para tratar de dar una justificación al hecho de asumir que ciertos datos son aproximadamente normales.
- Si se muestrea una población que tiene una distribución de probabilidad desconocida o incluso no sigue una distribución normal, la distribución de muestreo de la media muestral será aproximadamente normal con media μ y varianza $\frac{\sigma^2}{n}$, siempre y cuando el tamaño de la muestra n sea grande.

Teorema del Límite Cental (TLC)

- Es uno de los resultados más importantes en estadística.
- Bajo ciertas condiciones nos permite usar la distribucióon normal para estudiar otras distribuciones más generales.

Teorema (Teorema del Límite Central)

Si $X_1, \cdots, X_n$ son variables aleatorias i.i.d. con media μ y desviación estándar σ, entonces $\displaystyle\sum_{i=1}^{n} X_i$ tiene asintóticamente (en el infinito) una distribución normal de media $n\mu$ y desviación estándar $\sqrt{n}\sigma$. Por las propiedades de la normal, tenemos que:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{n \to \infty} N(0,1) \equiv \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} N(0,1)$$

Teorema del Límite Cental (TLC)

- Este teorema nos dice que cuando los resultados de un experimento son debidos a un conjunto muy grande de causas independientes que actúan sumando sus efectos, entonces esos resultados siguen una distribución aproximadamente normal.
- Dicho de otra forma, aunque cada uno de los efectos sea raro o difícil de estudiar, si lo que queremos estudiar es la suma de los mismos sabemos que, bajo ciertas condiciones, esta se comportaría de modo normal.
- El Teorema Central del Límite sirve para dar una explicación al hecho constatado de que muchas distribuciones de variables observadas en la naturaleza o en experimentos físicos sean aproximadamente normales.

Ejemplos (binomial)

- Veamos el Teorema Central del Límite en el caso de la binomial.
- Como sabemos la binomial de parámetros n y p es la suma de n Bernoullis de parámetro p.
- Tomamos, $n = 40$ y $p = 0,4$.
- El valor esperado y la desviación estándar de la binomial

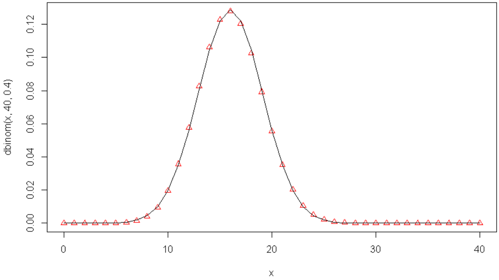
  $$np = 16 \quad \text{y} \quad \sqrt{np(1-p)} = \sqrt{40 \times 0,4 \times 0,6} = 3,0983$$

  respectivamente.

Ejemplos (binomial)

- Comparamos ambas distribuciones.

```
> x=seq(0,40,by=1)
> plot(x,dbinom(x,40,0.4),pch=2,col=2)
> lines(x,dnorm(x,16,3.0983))
```



- Como se puede comprobar la aproximación de la binomial por la normal es realmente buena.

Ejemplos (binomial)

- Podemos comparar numéricamente estos valores:

```
> for (i in 0:40) cat(i,"\t",dbinom(i,40,0.4),"\t",
dnorm(i,16,3.0983),"\n")
```

0	1.336749e-09	2.083865e-07
1	3.564665e-08	1.047402e-06
2	4.634065e-07	4.743683e-06
3	3.91321e-06	1.935871e-05
4	2.413146e-05	7.118619e-05
5	0.000115831	0.0002358704
6	0.00045054	0.000704222
7	0.001458613	0.001894542
8	0.004011185	0.004592586

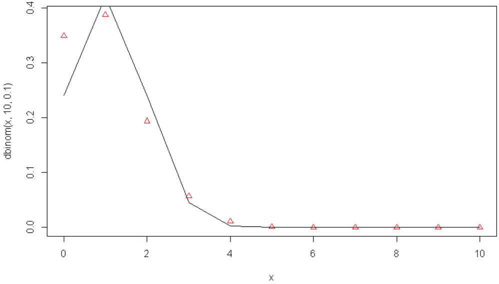
Ejemplos (binomial)

- En la práctica se utiliza la aproximación cuando $n \geq 30$, $np \geq 5$ y $n(1-p) \geq 5$.
- En nuestro caso se verifican todas las restricciones ya que $n = 40$, $np = 16$ y $n(1-p) = 24$.
- Para comprobar que estas restricciones son importantes, repetimos el experimento tomando valores que no las verifican. Por ejemplo, tomamos $n = 10$ y $p = 0,1$. Entonces, $np = 1$ y $n(1-p) = 9$. Vemos el efecto:

  ```
  > x=seq(0,10,by=1)
  > plot(x,dbinom(x,10,0.1),pch=2,col=2)
  > lines(x,dnorm(x,1,0.9486))
  ```

Ejemplos (binomial)

- Como se puede comprobar la aproximación de la binomial por la normal no es tan buena como en el caso anterior.



- Podemos comparar numéricamente estos valores:

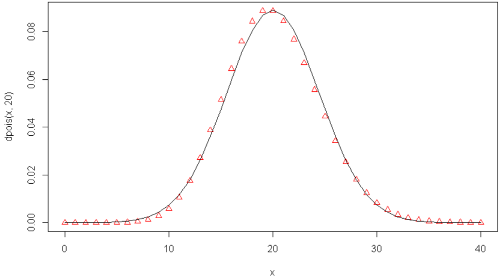
```
> for (i in 0:10) cat(i,"\t",dbinom(i,10,0.1),"\t",
dnorm(i,1,0.9486),"\n")
```

Ejemplos (Poisson)

- Supongamos que tenemos que el número de entradas en una página web por hora tenía una distribución Poisson de parámetro $\lambda = 20$. Esta variable puede considerarse como la suma de 60 variables que midan el número de entradas en una página web por minuto, que también son variables Poisson de parámetro $\frac{20}{60} = 0,3333$.
- El TLC aplicado a variables con distribución Poisson dice que cuanto mayor sea el número de variables Poisson que sumemos, la variable resultante (que también es Poisson) se parecerá cada vez más a una normal.
- Se considera que si $\lambda > 5$, la aproximación a la normal es buena. En ese caso, se puede considerar que, aproximadamente $N(\lambda, \sqrt{\lambda})$.

Ejemplos (Poisson)

```
> x=seq(0,40,by=1)
> plot(x,dpois(x,20),pch=2,col=2)
> lines(x,dnorm(x,20,4.4721))
```



Ejemplos (Poisson)

- Como se puede comprobar la aproximación de la poisson por la normal es bastante buena. Podemos comparar numéricamente estos valores:

```
> for (i in 0:40) cat(i,"\t",dpois(i,20),"\t",
  dnorm(i,20,4.4721),"\n")
```

0	2.061154e-09	4.049337e-06
1	4.122307e-08	1.073564e-05
2	4.122307e-07	2.707427e-05
3	2.748205e-06	6.49487e-05
4	1.374102e-05	0.0001482072
5	5.49641e-05	0.0003217013
6	0.0001832137	0.0006642344
7	0.0005234676	0.001304592
8	0.001308669	0.002437324
9	0.002908153	0.004331483

Referencia

- Leer a sección 7.3 y 7.4 del libro base.
- Revisar los ejercicios propuestos en dichas secciones.