

Análisis de Distribuciones de Probabilidad

Estudio de cuatro muestras discretas

Objetivo del análisis

Dadas cuatro muestras numéricas discretas:

- Identificar la distribución de probabilidad generadora
- Estimar sus parámetros
- Validar el ajuste mediante pruebas estadísticas

Se emplean las pruebas:

- χ^2 de bondad de ajuste
- Kolmogorov–Smirnov

Marco teórico

Sea X una variable aleatoria discreta.

Distribución Binomial

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$$\mathbb{E}[X] = np, \quad \text{Var}(X) = np(1 - p)$$

Distribución Poisson–Binomial

$$X = \sum_{i=1}^m \text{Bernoulli}(p_i)$$

$$\mathbb{E}[X] = \sum p_i, \quad \text{Var}(X) = \sum p_i(1 - p_i)$$

Muestra 1

Análisis análogo:

- Histograma
- Identificación de distribución
- Pruebas χ^2 y K-S

Descripción de la muestra 2

Características observadas:

- Variable discreta
- Valores enteros entre 0 y 14
- Tamaño muestral $n \approx 100\,000$

Momentos empíricos:

$$\mu = 6.206, \quad \sigma^2 = 3.784$$

La forma empírica sugiere una distribución unimodal y aproximadamente simétrica.

Frecuencias observadas

Se calcularon:

- Frecuencias absolutas O_i
- Frecuencias relativas $f_i = O_i/n$

El histograma de barras muestra una forma de campana discreta, con colas acotadas.

Ajuste binomial por momentos

Se igualan los momentos teóricos con los empíricos:

$$np = \mu$$

$$np(1 - p) = \sigma^2$$

Resolviendo:

$$n \approx 16, \quad p \approx 0.388$$

Esto define la binomial candidata:

$$X \sim \text{Bin}(16, 0.388)$$

Prueba χ^2 (binomial)

Se define:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Resultados:

$$\chi^2 \approx 184$$

$$p\text{-valor} \ll 0.05$$

Conclusión:

Se rechaza la hipótesis de binomial con ensayos idénticos.

Modelo Poisson–Binomial

Se considera un modelo más general:

$$X = \sum_{i=1}^m \text{Bernoulli}(p_i)$$

Estimación del número efectivo:

$$m_{\text{eff}} = \frac{\mu^2}{\mu - \sigma^2} \approx 15.9$$

Este valor es consistente con un proceso de aproximadamente 16 ensayos no idénticos.

Conclusión para la muestra 2

- La muestra no es binomial i.i.d.
- Es compatible con una Poisson–binomial
- La binomial sirve como buena aproximación descriptiva

Distribución seleccionada:

Poisson–Binomial

Muestra 3

(Contenido a completar)

Descripción de la muestra 4 (M16)

Características observadas:

- Variable continua
- Soporte estrictamente positivo
- Tamaño muestral: $n = 100\,000$

Momentos empíricos:

$$\mu = 10.33, \quad \sigma^2 = 137.93, \quad \sigma = 11.74$$

Valores extremos:

$$x_{\min} = 1.29 \times 10^{-13}, \quad x_{\max} = 41.67$$

La muestra presenta fuerte asimetría a la derecha.

Análisis exploratorio

Observaciones relevantes:

- Dispersión elevada: $\sigma > \mu$
- Cola derecha larga
- Alta concentración de valores pequeños

Estas características descartan distribuciones simétricas y sugieren modelos con soporte positivo como:

- Lognormal
- Gamma
- Weibull

Modelo lognormal candidato

Se propone inicialmente:

$$X \sim \text{LogNormal}(\mu_{\log}, \sigma_{\log})$$

donde:

$$\mu_{\log} = \mathbb{E}[\log X], \quad \sigma_{\log} = \text{Std}(\log X)$$

La distribución lognormal es adecuada para variables generadas por procesos multiplicativos y con colas largas.

Prueba χ^2 (lognormal)

Se agrupan los datos en clases de igual amplitud y se calcula:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Resultados:

$$\chi^2 \text{ grande}, \quad p\text{-valor} \ll 0.05$$

Conclusión:

Se rechaza la hipótesis exacta de lognormalidad.

Prueba Kolmogorov–Smirnov

Se compara la CDF empírica $F_n(x)$ con la CDF teórica $F(x)$. Resultados:

$$D \text{ elevado}, \quad p\text{-valor} \approx 0$$

El rechazo se debe en gran parte al tamaño muestral y a desviaciones localizadas en las colas.

Comparación con otras distribuciones

Se evaluaron modelos alternativos:

- Gamma
- Weibull
- Exponencial

Resultados comparativos:

- Ningún modelo supera completamente las pruebas
- La lognormal presenta el mejor ajuste visual global
- Mejor comportamiento en la región central

Conclusión para la muestra 4

- La muestra no sigue exactamente una lognormal
- Las pruebas formales rechazan el ajuste exacto
- La lognormal es la mejor aproximación empírica global

Distribución seleccionada:

Lognormal (aproximación empírica)

Conclusiones generales

- Cada muestra presenta un mecanismo generador distinto
- Las pruebas estadísticas son determinantes
- La similitud gráfica no implica igualdad de distribuciones
- Distribuciones complejas describen mejor datos reales