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

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

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

Distribución Poisson–Binomial

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

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

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:

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

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:

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

Esto define la binomial candidata:

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

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:

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

Valores extremos:

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

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:

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

donde:

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

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:

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

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:

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

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)

Descripción de la muestra (m5)

- Variable discreta (valores enteros)
- Tamaño muestral: $n = 100\,000$
- Valores: $x_{\text{min}} = 3$, $x_{\text{max}} = 17$

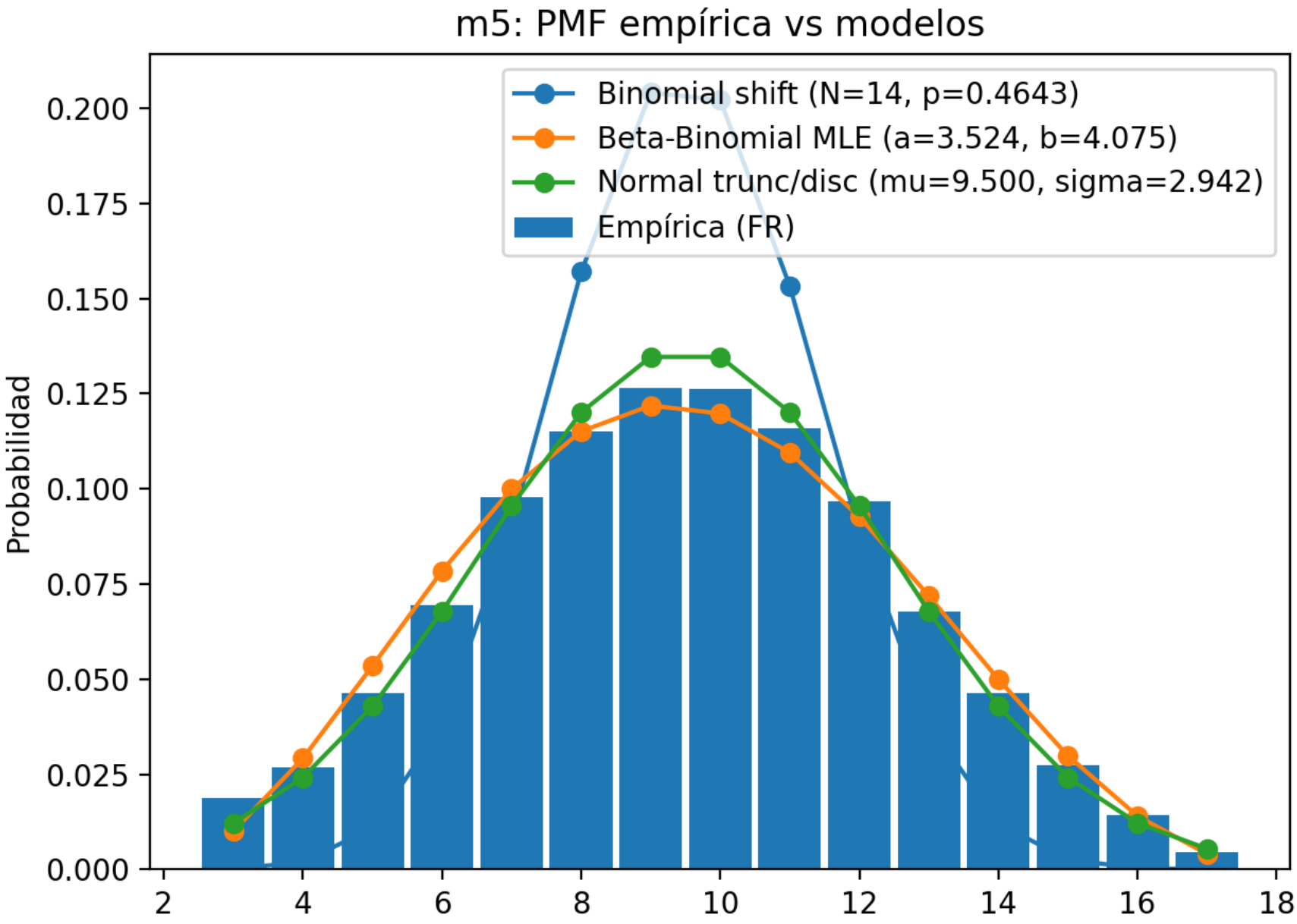
Momentos empíricos:

$$\mu = 9.49973, \qquad \sigma^2 = 8.65502, \qquad \sigma = 2.94194$$

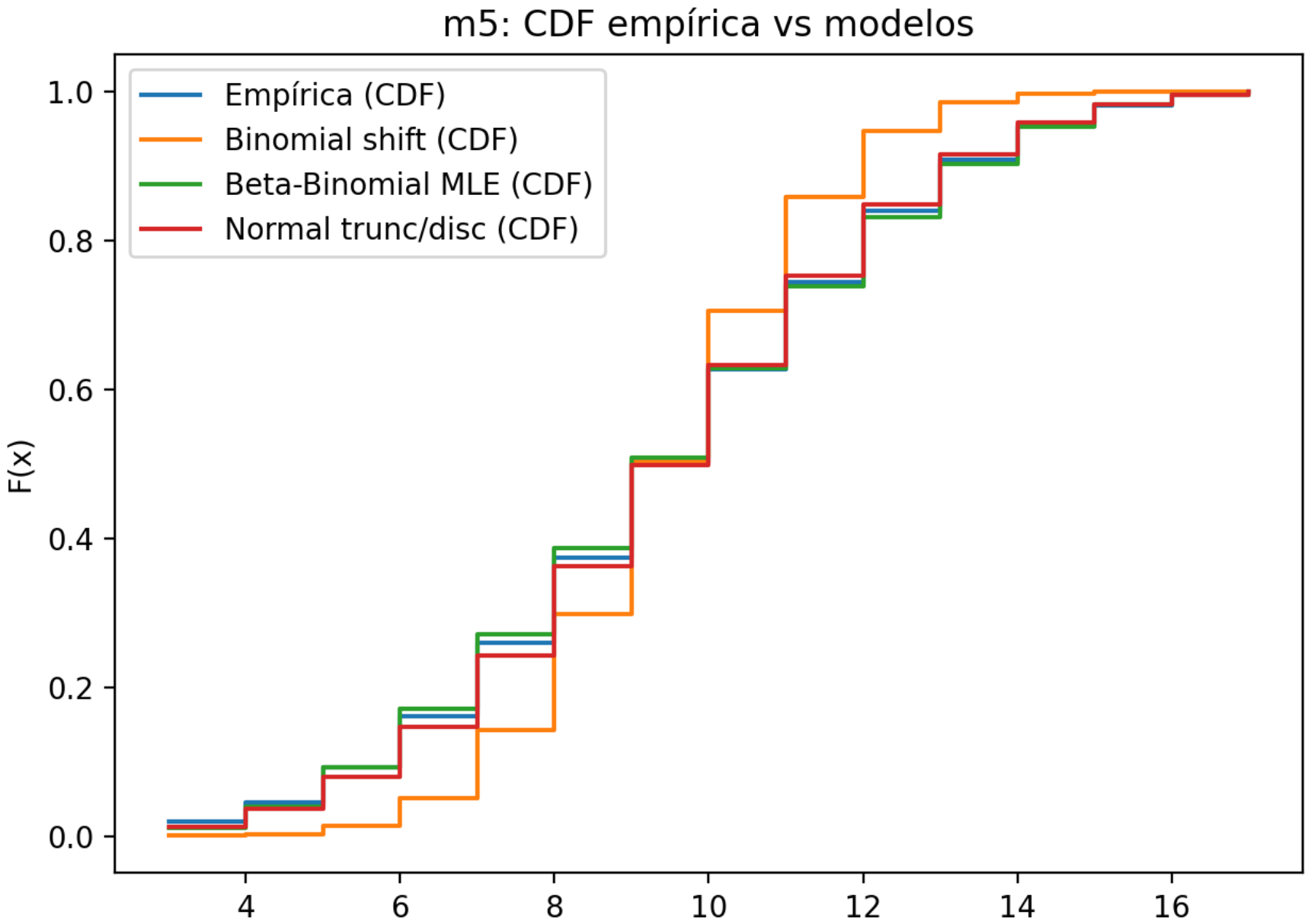
Frecuencias observadas (FO y FR)

x	FO	FR
3	1875	0.01875
4	2697	0.02697
5	4634	0.04634
6	6942	0.06942
7	9784	0.09784
8	11508	0.11508
9	12654	0.12654
10	12633	0.12633
11	11588	0.11588
12	9661	0.09661
13	6764	0.06764
14	4621	0.04621
15	2745	0.02745
16	1429	0.01429
17	465	0.00465

Comparación visual: PMF empírica vs modelos



Comparación visual: CDF empírica vs modelos



Modelos evaluados

- **Binomial desplazada:** $Y = X - 3 \sim \mathrm{Bin}(14, p)$, $p = 0.46427$.
- **Beta-Binomial (MLE):** $Y \sim \mathrm{BetaBin}(14, \alpha, \beta)$,

$$\alpha = 3.52441, \qquad \beta = 4.07544.$$

- **Normal truncada/discretizada:**

$$\mu = 9.49973, \qquad \sigma = 2.94194.$$

Prueba χ^2 (bondad de ajuste)

- **Binomial desplazada:**

$$\chi^2 = 459708.23, \quad df = 13, \quad p \approx 0 \Rightarrow \textbf{Rechazado.}$$

- **Beta-Binomial (MLE):**

$$\chi^2 = 1146.67, \quad df = 12, \quad p \approx 0 \Rightarrow \textbf{Rechazado.}$$

- **Normal truncada/discretizada:**

$$\chi^2 = 679.74, \quad df = 12, \quad p \approx 0 \Rightarrow \textbf{Rechazado.}$$

Nota

Con $n = 100\,000$, incluso discrepancias pequeñas suelen dar p-valores muy bajos.

Prueba Kolmogorov–Smirnov

- **Binomial desplazada:**

$$D = 0.23075, \quad D_{\text{crit}} = 0.00430 \Rightarrow \textbf{Rechazado.}$$

- **Beta-Binomial (MLE):**

$$D = 0.13389, \quad D_{\text{crit}} = 0.00430 \Rightarrow \textbf{Rechazado.}$$

- **Normal truncada/discretizada:**

$$D = 0.01895, \quad D_{\text{crit}} = 0.00430, \quad p \approx 1.28 \times 10^{-31} \Rightarrow \textbf{Rechazado.}$$

Conclusión (m5)

- Todas las hipótesis fueron rechazadas por χ^2 y K–S (alto poder por n grande).
- Se selecciona el modelo con mejor semejanza visual y menores estadísticos:

Mejor aproximación: Normal truncada/discretizada en $[3, 17]$ ($\mu = 9.49973$,

- Segundo mejor: Beta-Binomial (MLE).
- Peor ajuste: Binomial desplazada.

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