Selección de distribuciones de probabilidad

Patricia Kisbye

FaMAF

6 de mayo, 2010

Análisis estadístico de datos simulados

- Los sistemas reales tienen fuentes de aleatoriedad:

Tipo de sistema	**Fuente de aleatoriedad**
Fabricación	Tiempos de procesamiento Tiempos de falla Tiempos de reparación de máquinas
Defensa	Tiempos de arribo y carga útil de aviones o misiles. Errores de lanzamiento.
Comunicaciones	Tiempos entre llegadas de mensajes. Longitudes de mensajes.
Transporte	Tiempo de embarque Tiempos entre arribos a un subte...

Simulación a partir de los datos

Para simular un sistema real es necesario:

- Representar cada fuente de aleatoriedad de acuerdo a una distribución de probabilidad.
- Elegir adecuadamente la distribución, para no afectar los resultados de la simulación.

¿Cómo elegir una distribución? ¿Cómo simular un sistema a partir de un conjunto de observaciones?

- Utilizando los datos directamente.
- Realizando el muestreo a partir de la distribución *empírica* de los datos.
- Utilizando técnicas de inferencia estadística.

Elección de una distribución

Utilizar los datos directamente:

- Sólo reproduce datos históricos.
- En general es una información insuficiente para realizar simulaciones.
- Es útil para comparar dos sistemas, para hacer una validación del modelo existente con el simulado.

Distribución empírica:

- Reproduce datos intermedios (datos continuos).
- Es recomendable si no se pueden ajustar los datos a una distribución teórica.

Inferencia estadística

Inferencia estadística vs. distribución empírica:

- Las distribuciones empíricas pueden tener irregularidades si hay pocos datos, una distribución teórica suaviza los datos.
- Puede obtenerse información aún fuera del rango de los datos observados.
- Puede ser necesario imponer un determinado tipo de distribución, por el tipo de modelo que se desea simular.
- No es necesario almacenar los datos observados ni las correspondientes probabilidades acumuladas.
- Es fácil modificar los parámetros.
- Puede no existir una distribución adecuada.
- Generación de valores extremos no deseados.

Distribuciones de probabilidad más utilizadas

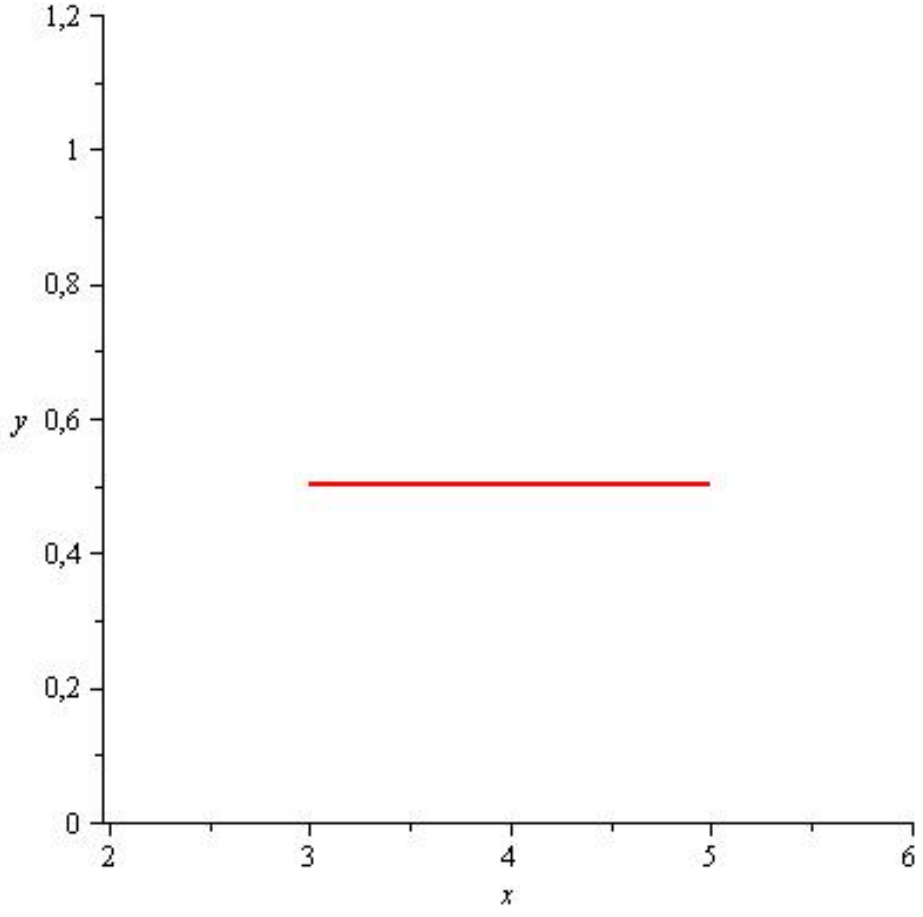
Continuas:

- Uniforme: Para cantidades que varían "aleatoriamente" entre valores *a* y *b*, y que no se conocen más datos.
- Exponencial: Tiempos entre llegadas de "clientes" a un sistema, y que ocurren a una tasa constante. Tiempos de falla de máquinas.
- Gamma, Weibull: Tiempo de servicio, tiempos de reparación.
- Normal: Errores. Sumas grandes $\rightarrow$ Teorema central del límite.
- Otras: (Law & Kelton, cap. 6)

Parámetros:

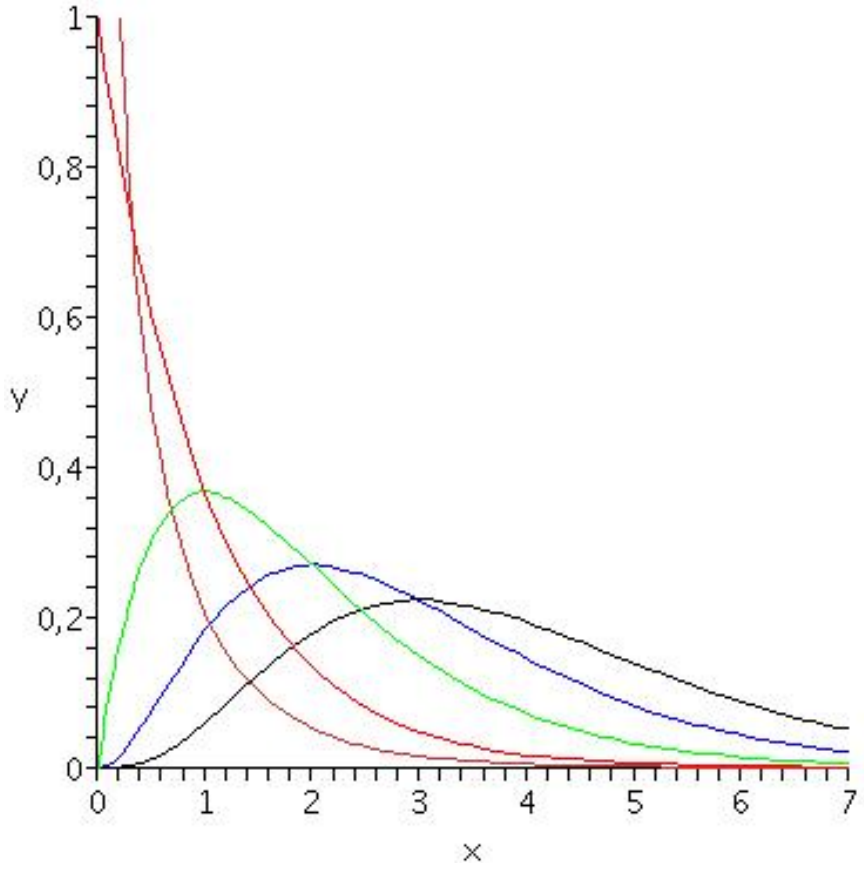
- de posición: (normal, uniforme)
- de escala: (normal, uniforme, exponencial, lognormal)
- de forma: (Gamma, Weibull, lognormal)

Distribución uniforme



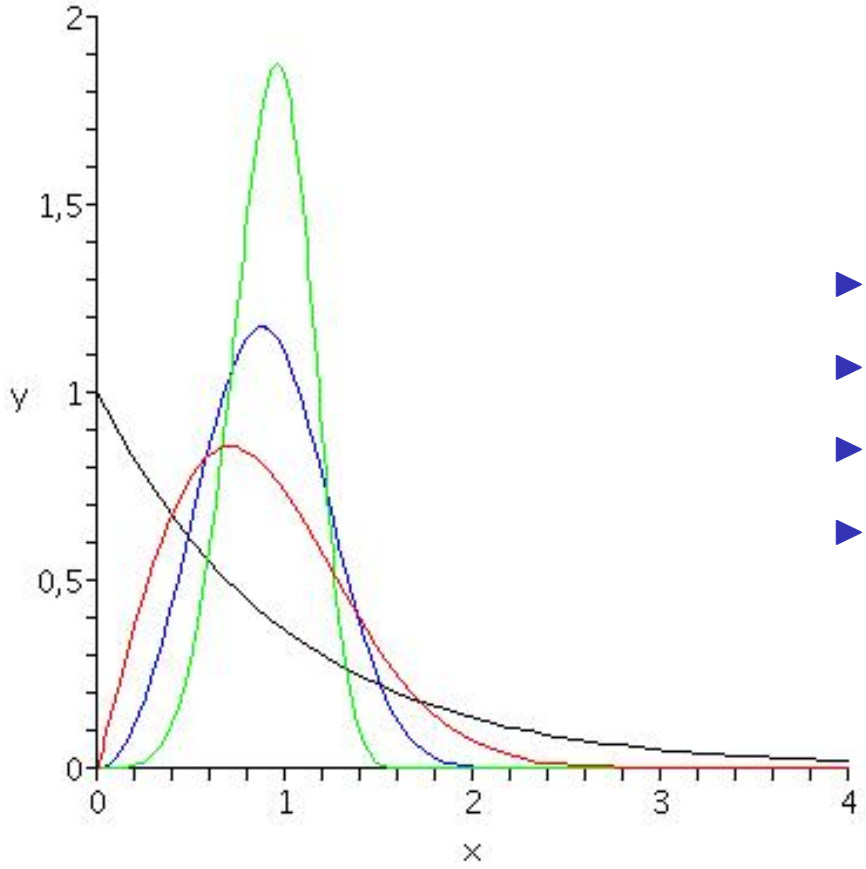
- $f(x) = \frac{1}{b-a}\mathbb{I}_{(a,b)}(x)$
- a: posición, $b - a$: escala.
- Rango: $a < x < b$.
- Media: $\frac{a+b}{2}$.
- Varianza: $\frac{(b-a)^2}{2}$.

Distribución Gamma(α, β)



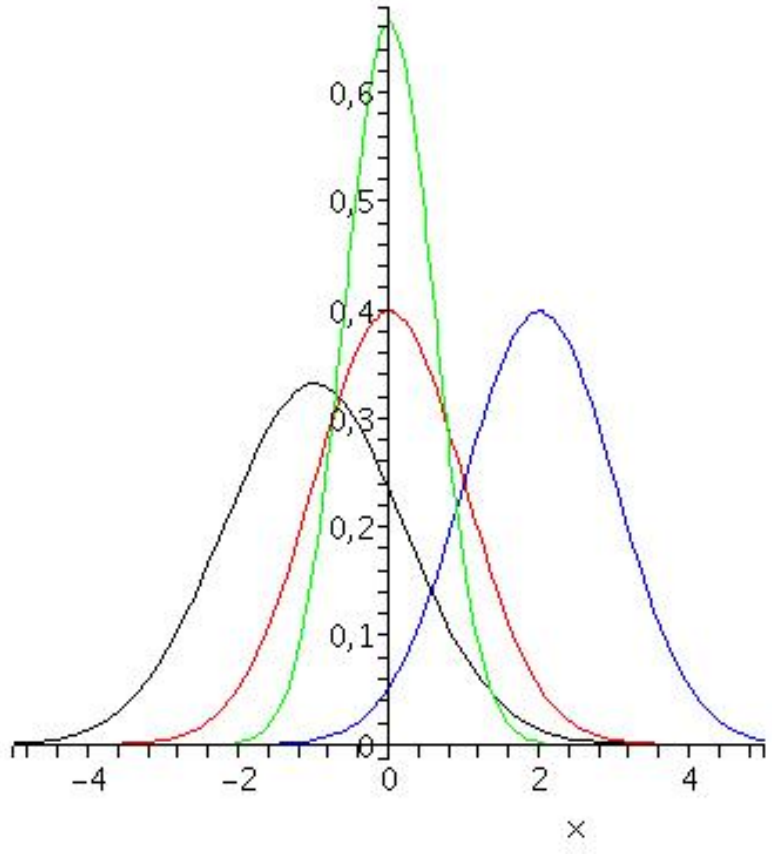
- $f(x) = \frac{\beta^{-\alpha} x^{\alpha-1} \exp(-x/\beta)}{\Gamma(\alpha)}$
- α: forma, β: escala.
- Rango: $x > 0$.
- Media: $\alpha\beta$.
- Varianza: $\alpha\beta^2$.
- NOTACIÓN para β
- $\alpha = 1 \Rightarrow$ *Exponencial*

Distribución Weibull (α, β)



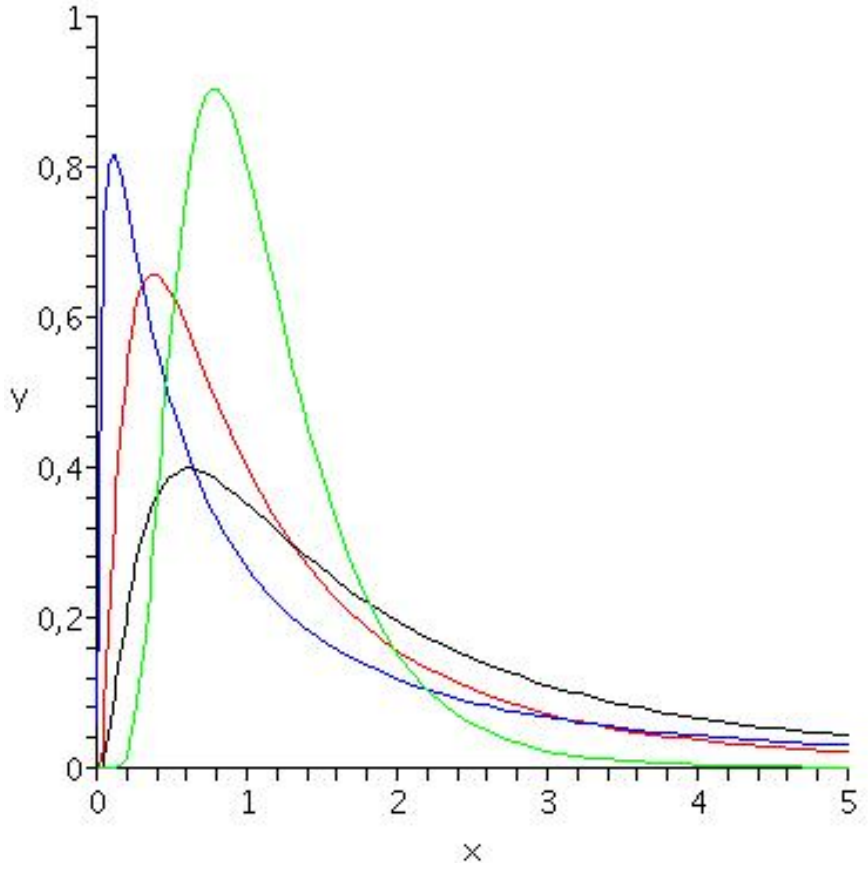
- $f(x) = \alpha\beta^{-\alpha}x^{\alpha-1}\, e^{-(x/\beta)^{\alpha}}$
- α: forma, β: escala.
- Rango: $x > 0$.
- Media: $\frac{\beta}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)$.

Distribución Normal(μ, σ^2)



- $f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-(x-\mu)^2/(2\sigma^2))$
- μ: posición, σ: escala.
- Rango: $\mathbb{R}$.
- Media: μ.
- Varianza: σ^2.

Distribución Lognormal(μ, σ^2)



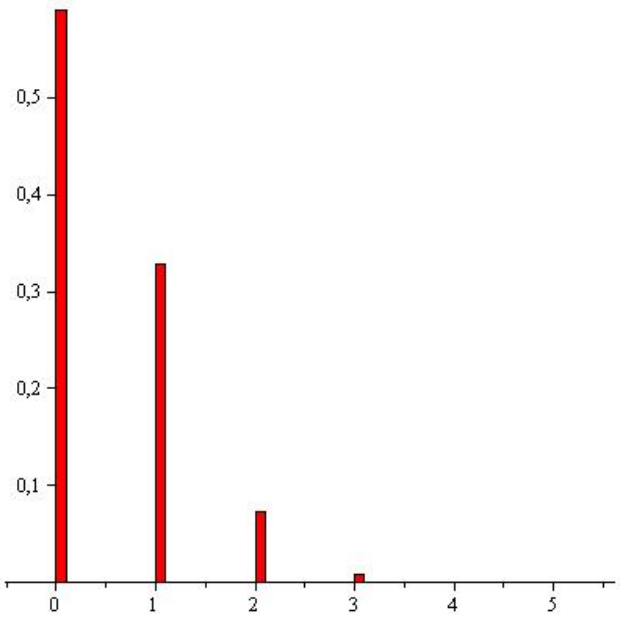
- $f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-(\log(x)-\mu)^2/(2\sigma^2)}$
- σ: forma, μ: escala.
- Rango: $x > 0$.
- Media: $e^{\mu+\sigma^2/2}$.
- Varianza: $e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$.

Distribuciones de probabilidad más utilizadas

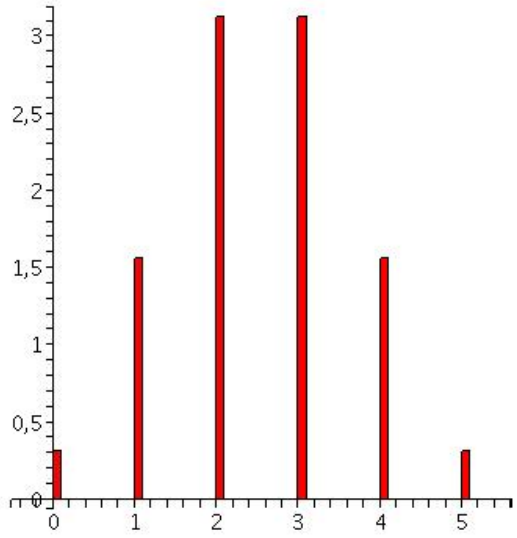
Discretas:

- Bernoulli.
- Uniforme discreta.
- Geométrica: número de observaciones hasta detectar el primer error.
- Binomial negativa: número de observaciones hasta detectar el n-ésimo error.
- Poisson: Número de eventos en un intervalo de tiempo, si ocurren a tasa constante.

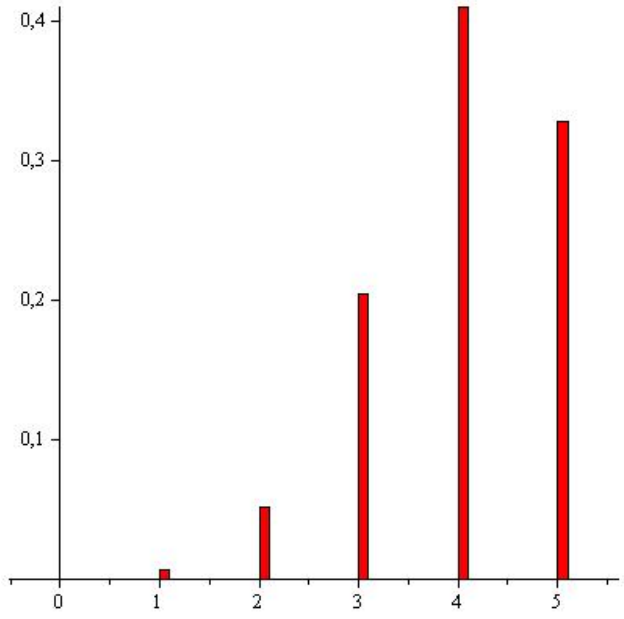
Distribución Binomial



$n = 5,\ p = 0.1$

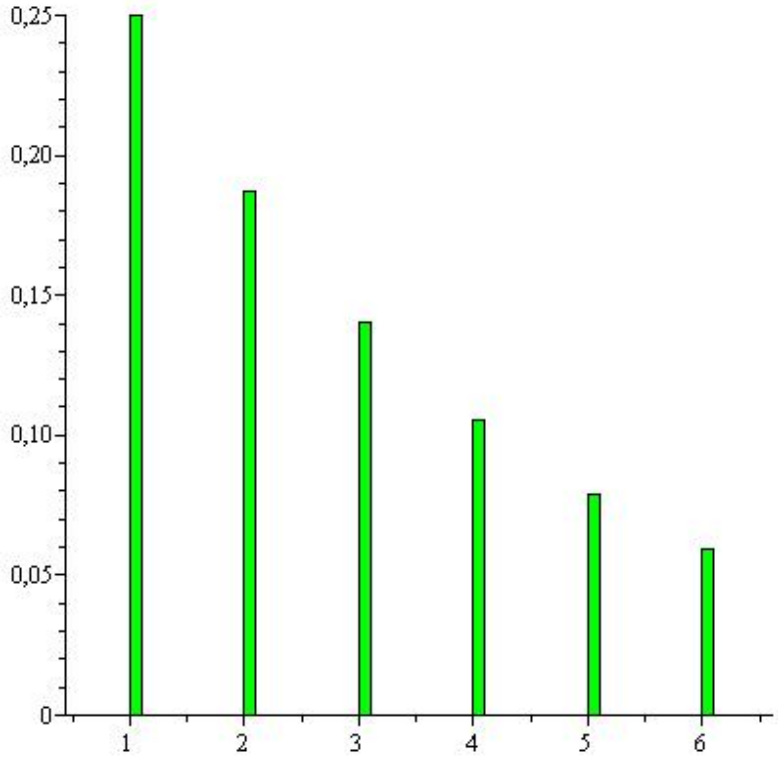


$n = 5,\ p = 0.5$

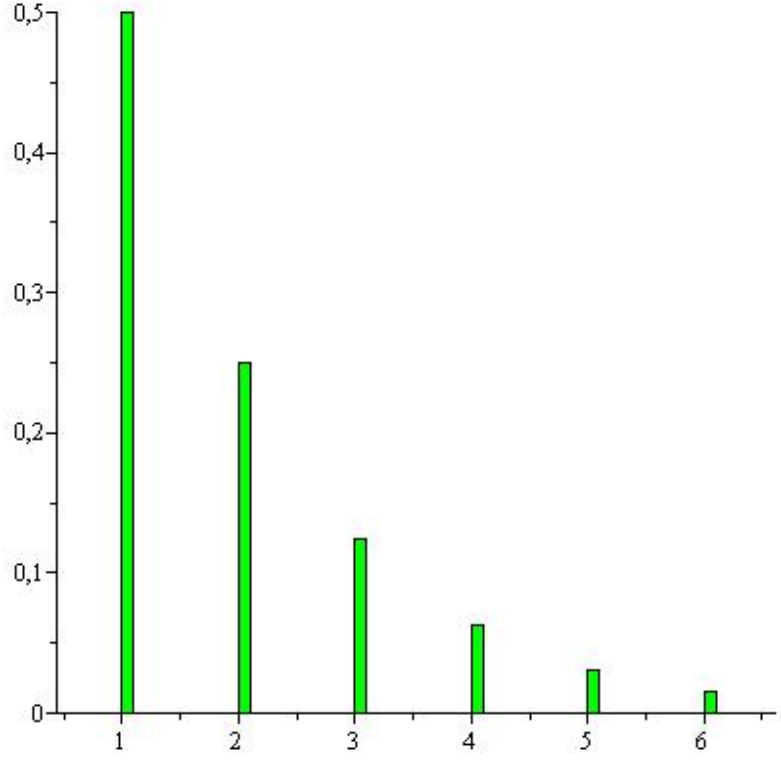


$n = 5,\ p = 0.8$

Distribución Geométrica

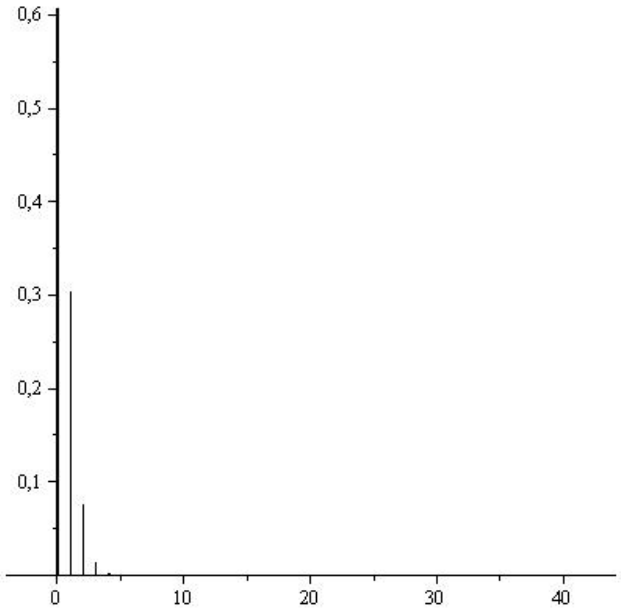


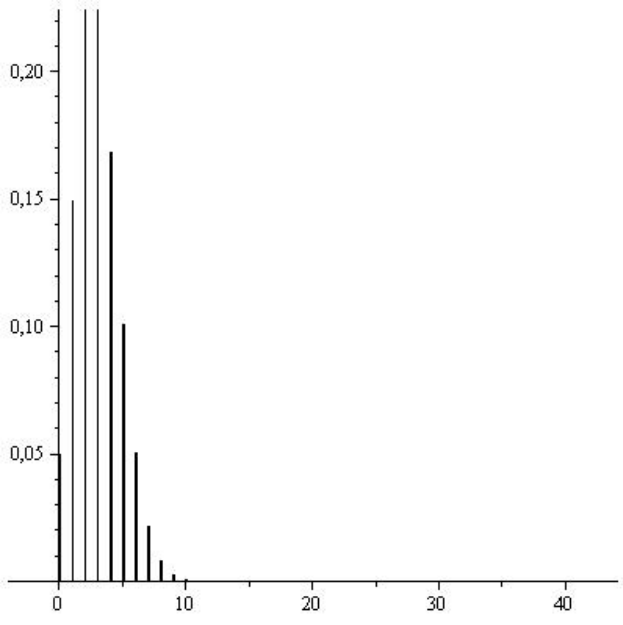
$p = 0.25$

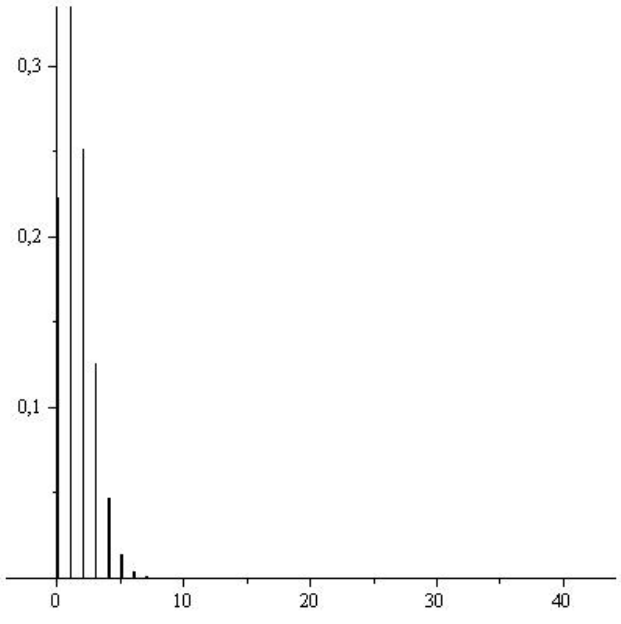


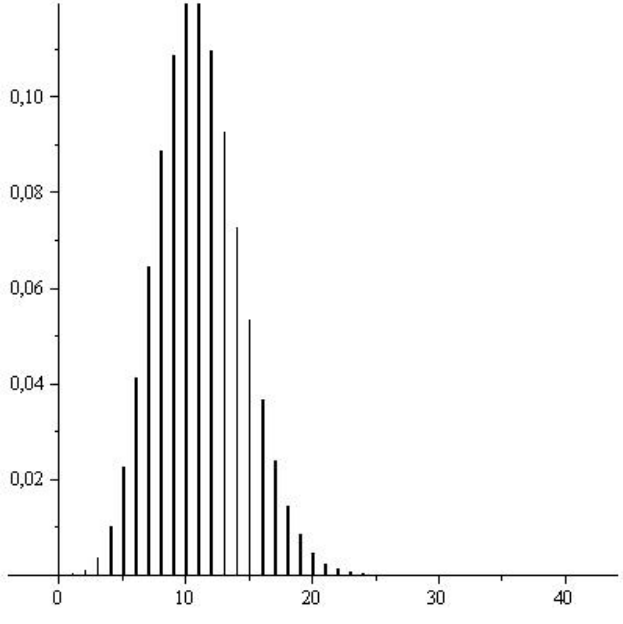
$p = 0.5$

Distribución Poisson

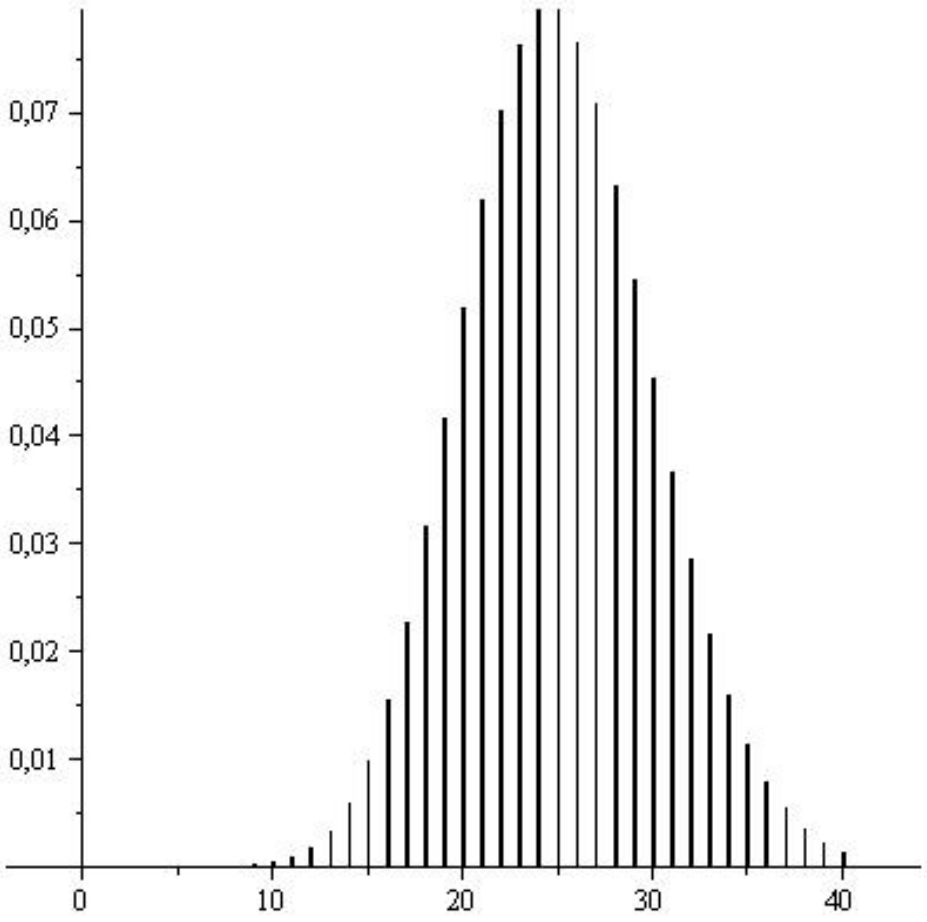








Distribución Poisson



Corresponde a $\lambda = 25$.

Distribución empírica

Datos continuos

- Datos observados: $X_1, X_2, \ldots, X_n$ disponibles.
- Datos agrupados en un intervalo: histograma.

Si los datos están disponibles, puede construirse una función de distribución lineal a trozos:

- Ordenando los datos de menor a mayor: $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$
- Definiendo $F(x)$ como:

$$F(x) = \begin{cases} 0 & x < X_{(1)} \\ \frac{i-1}{n-1} + \frac{x - X_{(i)}}{(n-1)(X_{(i+1)} - X_{(i)})} & X_{(i)} \leq x < X_{(i+1)} \quad 1 \leq i < n \\ 1 & X_{(n)} \leq x \end{cases}$$

$F(X_{(i)}) = (i-1)/(n-1) \approx$ proporción de X_j menores que $X_{(i)}$.

Si los datos están agrupados en intervalos:

$$[a_0, a_1), [a_1, a_2), \ldots, [a_{k-1}, a_k)$$

y cada intervalo $[a_{j-1}, a_j)$ contiene n_j observaciones:

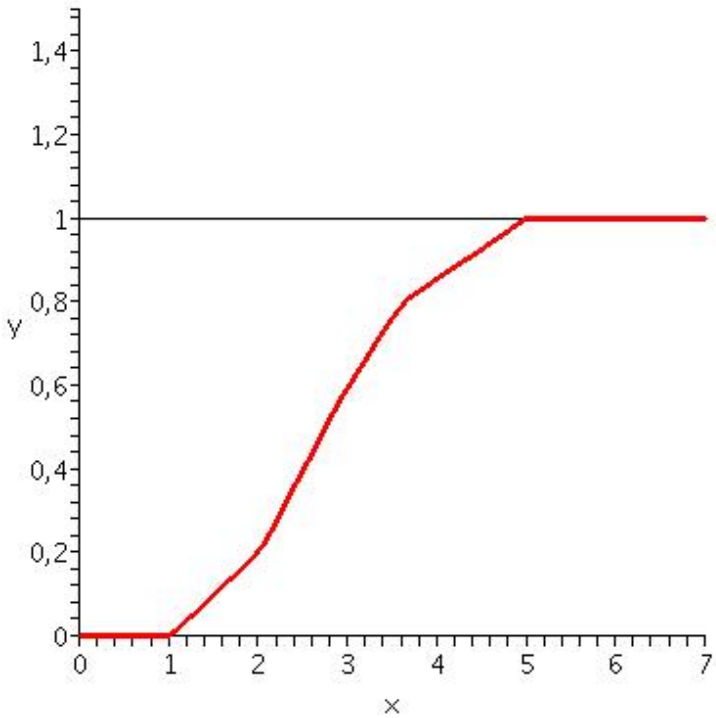
$$n_1 + n_2 + \cdots + n_k = n$$

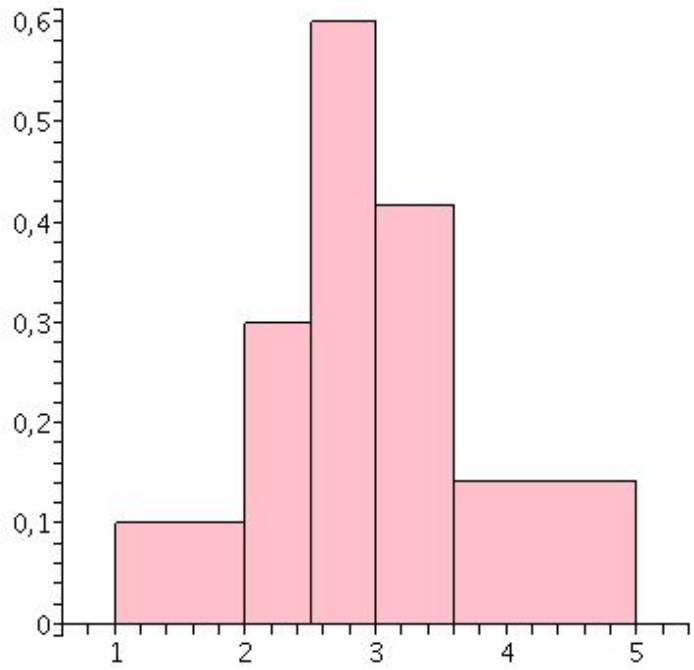
es razonable definir la distribución empírica como:

$$G(a_0) = 0, \qquad G(a_j) = \frac{n_1 + n_2 + \cdots + n_j}{n}$$

e interpolando linealmente estos puntos.
$G(a_j) =$ proporción de X_j menores que a_j.

Distribución empírica





Caso discreto

- Si los datos $X_1, X_2, \ldots, X_n$ están disponibles se define la función de masa empírica:

$$p(x) = \frac{\#\{i \mid X_i = x\}}{n}$$

- Si los datos están agrupados, se define la función de masa $p(x)$ de modo que la suma en cada intervalo sea igual a la proporción de X_i´s en dicho intervalo.
  La definición de $p(x)$ dentro del intervalo es esencialmente arbitraria.

Técnicas de prueba de independencia

Para ciertos tests, es necesario asumir independencia de los datos observados.
Ejemplos de no independencia:

- Datos de temperaturas a lo largo de un día.
- Tiempos de demora en una cola de espera.

Técnicas

- Gráficos de correlación: ρ_j.
- Diagramas de dispersión (scattering): (X_i, X_{i+1}).
- Tests no paramétricos.

Inferencia estadística

1. Elegir una o más distribuciones apropiadas.
2. Estimación de parámetros de la distribución elegida.
3. Pruebas (tests) de bondad de ajuste.
4. Si es necesario, corregir la distribución adoptada.

Elegir una distribución

- Conocer el origen de los datos.
- Estimar algunas medidas a partir de los datos:
  - Media, mediana, máximo y mínimo, coeficiente de variación, desviación estándar, coeficiente de asimetría.
- Histograma.
- q-cuantiles, diagramas de caja (box-plots)
- $Q - Q$ plots y $P - P$ plots.

Medidas útiles

Función	Estimador	Estima
Min, Max	$X_{(1)}, X_{(n)}$	rango
Media μ	$\overline{X}(n)$	Tendencia central
Mediana	$\hat{m} = \begin{cases} X_{(n+1)/2} \\ \frac{1}{2}(X_{n/2} + X_{(n/2+1)}) \end{cases}$	Tendencia central.
Varianza σ^2	$S^2(n)$	Variabilidad
c.v.=$\frac{\sigma}{\mu}$	$\hat{cv}(n) = \frac{\sqrt{S^2(n)}}{\overline{X}(n)}$	Variabilidad
τ	$\hat{\tau} = \frac{S^2(n)}{\overline{X}(n)}$	Variabilidad
Asimetría $\nu = \frac{E[(X-\mu)^3]}{(\sigma^2)^{3/2}}$	$\hat{\nu}(n) = \frac{\sum_i (X_i - \overline{X}(n))^3/n}{[S^2(n)]^{3/2}}$	Simetría

Estimación de parámetros

Dada una muestra de n datos observados, se llama estimador $\hat{\theta}$ del parámetro θ a cualquier función de los datos observados.

Propiedades de un buen estimador

- Insesgabilidad: se dice que el estimador es insesgado si $E[\hat{\theta}] = \theta$.
- Consistencia: si al aumentar la muestra, el estimador se aproxima al parámetro.
- Eficiencia: se calcula comparando su varianza con la de otro estimador.
- Suficiencia: utiliza toda la información obtenida de la muestra.

Estimador de máxima verosimilitud

Supongamos que se tiene la hipótesis de una distribución discreta para los datos observados, y se desconoce un parámetro θ.
Sea $p_\theta(x)$ la probabilidad de masa para dicha distribución.

Dado que se han observado datos $X_1, X_2, \ldots, X_n$, se define la función de máxima verosimilitud $L(\theta)$ como sigue:

$$L(\theta) = p_\theta(X_1) \cdot p_\theta(X_2) \cdots p_\theta(X_n).$$

El estimador de máxima verosimilitud es el valor $\hat{\theta}$ que maximiza $L(\theta)$:

$$L(\hat{\theta}) \geq L(\theta), \qquad \theta \text{ valor posible.}$$

Si la distribución supuesta es continua, y $f_\theta(x)$ es la densidad para dicha distribución, se define:

$$L(\theta) = f_\theta(X_1) \cdot f_\theta(X_2) \cdots f_\theta(X_n).$$

Estimador de máxima verosimilitud

El estimador de máxima verosimilitud tiene, en general, las siguientes propiedades:

1. Es único: $L(\hat{\theta}) > L(\theta)$ para cualquier otro valor de θ.
2. La distribución asintótica de $\hat{\theta}$ tiene media θ.
3. Es invariante: $\phi = h(\theta)$, entonces $\hat{\phi} = h(\hat{\theta})$.
4. Su distribución asintótica está normalmente distribuida.
5. Es fuertemente consistente: $\lim_{n\to\infty} \hat{\theta} = \theta$.

Ejemplo

Para la distribución exponencial, $\theta = \beta$ ($\beta > 0$) y $f_\beta(x) = \frac{1}{\beta} e^{-x/\beta}$ para $x \geq 0$.

$$\begin{aligned} L(\beta) &= \left(\frac{1}{\beta} e^{-x_1/\beta}\right)\left(\frac{1}{\beta} e^{-x_2/\beta}\right) \cdots \left(\frac{1}{\beta} e^{-x_n/\beta}\right) \\ &= \beta^{-n} \exp\left(-\frac{1}{\beta}\sum_{i=1}^{n} x_i\right) \end{aligned}$$

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}(n) = \text{ Media muestral.}$$

Ejemplo

Para la distribución geométrica, $\theta = p$ $(0 < p < 10)$ y $p_p(x) = p(1-p)^{x-1}$ para $x = 1, 2, \ldots$.

$$\begin{aligned} L(p) &= p^n (1-p)^{\sum_{i=1}^n (x_i - 1)} \\ &= \left(\frac{p}{1-p}\right)^n (1-p)^{\sum_{i=1}^n x_i} \end{aligned}$$

$$\hat{p} = \left(\frac{1}{n}\sum x_i\right)^{-1}$$

Estimadores de máxima verosimilitud:

Distribuciones continuas:

- Uniforme: $\hat{a} = \min\{X_i\}$, $\hat{b} = \max\{X_i\}$.
- Exponencial: $\hat{\beta} = \overline{X}(n)$.
- Gamma, Weibull: $\hat{\alpha}$ y $\hat{\beta}$ se resuelven numéricamente.
- Normal:

$$\hat{\mu} = \overline{X}(n), \quad \hat{\sigma} = \left[\frac{n-1}{n} S^2(n)\right]^{1/2}.$$

- Lognormal:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \log(X_i)}{n}, \qquad \hat{\sigma} = \left[\frac{\sum_{i=1} n(\log(X_i) - \hat{\mu})^2}{n}\right]^{1/2}.$$

Estimadores de máxima verosimilitud

Distribuciones discretas:

- Binomial (t, p): si t es conocido, $\hat{p} = \overline{X}(n)/t$.
- Bernoulli: Caso binomial con $t = 1$ e igual p.
- Geométrica: $\hat{p} = \frac{1}{\overline{X}(n)}$.
- Binomial negativa (s, p): número de ensayos hasta el s-ésimo éxito. Si s es conocido: $\hat{p} = \frac{s}{\overline{X}(n)}$.
- Poisson: $\hat{\lambda} = \overline{X}(n)$.