Análisis y Visualización de Datos

Diplomatura CDAAyA 2020

Teoría, datos, experimentos, simulación...
¿Que es todo esto y cómo se combinan?

Variable Aleatoria (repetición del experimento)

X= cantidad de caras en 3 tiradas de moneda.

```
C, S = 'c', 's'
SAMPLE_SPACE = ['-'.join(x) for x in
                itertools.product([C, S], repeat=3)]
SAMPLE_SPACE
```

```
['c-c-c', 'c-c-s', 'c-s-c', 'c-s-s', 's-c-c', 's-c-s', 's-s-c', 's-s-s']
```

```
sampled_values = [
    x.count(C) for x in numpy.random.choice(SAMPLE_SPACE, 1000)]
```

Proporción de resultados tal que X=k :

```
result = numpy.unique(sampled_values, return_counts=True)
[(label, count/1000.0) for label, count in zip(*result)]
```

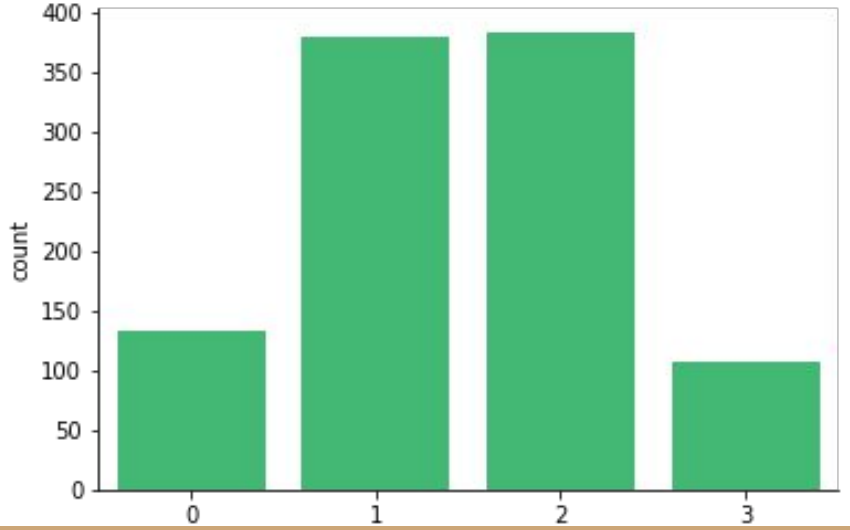
```
[(0, 0.132), (1, 0.379), (2, 0.383), (3, 0.106)]
```

Variable Aleatoria (repetición del experimento)

```
result = numpy.unique(sampled_values, return_counts=True)
[(label, count/1000.0) for label, count in zip(*result)]
```

```
[(0, 0.132), (1, 0.379), (2, 0.383), (3, 0.106)]
```

la Proporción de la muestra tal que X=k, estima la probabilidad P(X=k), p/ k=0,1,2,3



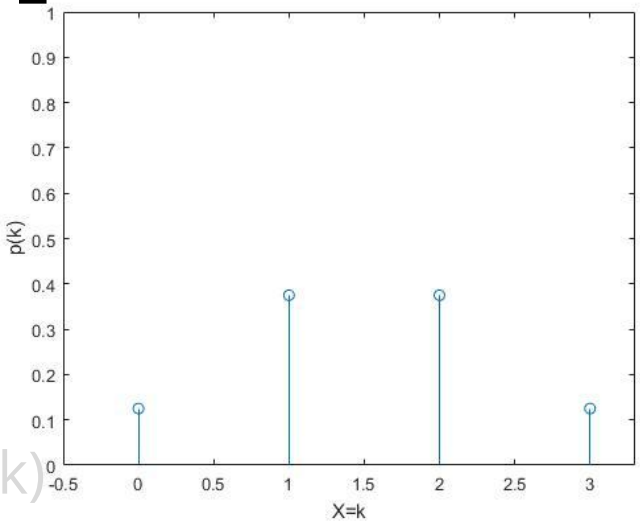
Variable Aleatoria (modelo matemático)

X= cantidad de caras en 3 tiradas de moneda. p(k)=P(X=k)?

Ω={ccc,ccs,csc,css,scc,scs,ssc,sss}, $\#\Omega=8=2^3$

p(0)=P(X=0)=1/8
p(1)=P(X=1)=3/8
p(2)=P(X=2)=3/8
p(3)=P(X=3)=1/8

Notar que la suma da 1, $\sum_k p(k)=1=\sum_k P(X=k)$



Variable binomial

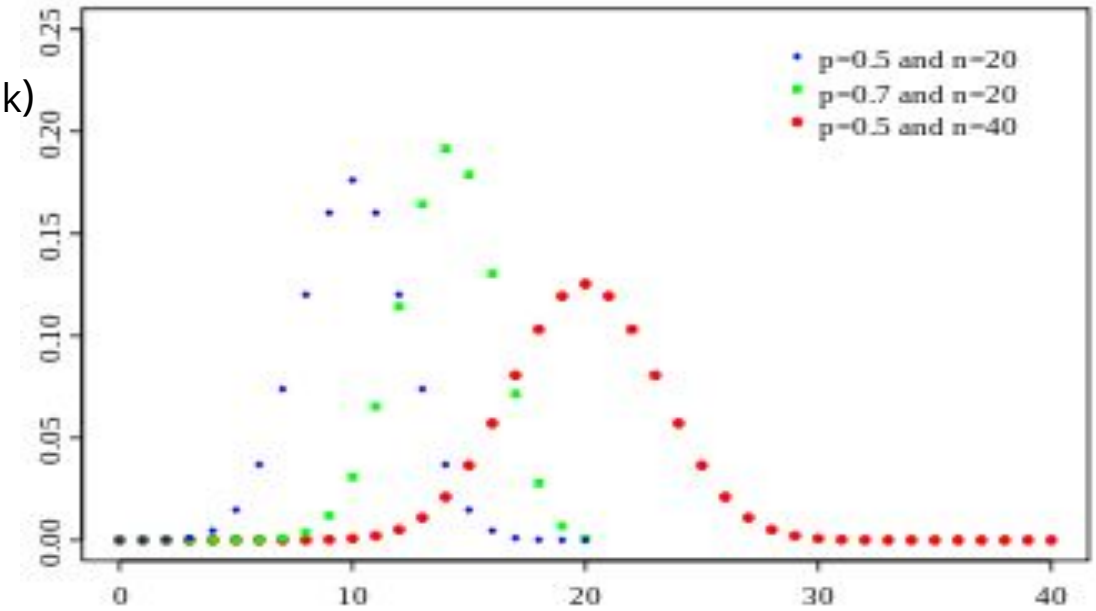
Sea X la v. a. discreta modela: cantidad de “éxitos” en una n-upla

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

k=0,1,...,n

p=probabilidad de “éxito”.

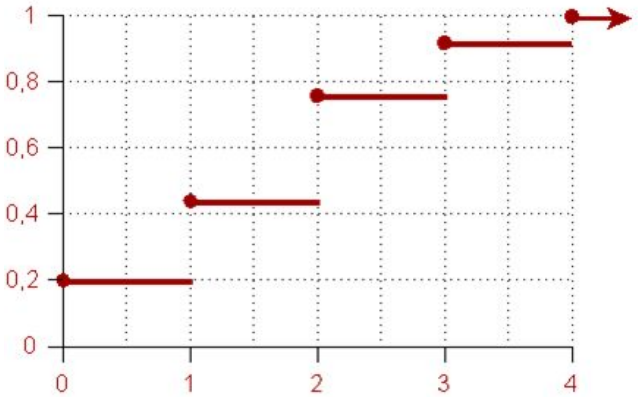
X~B(n,p)



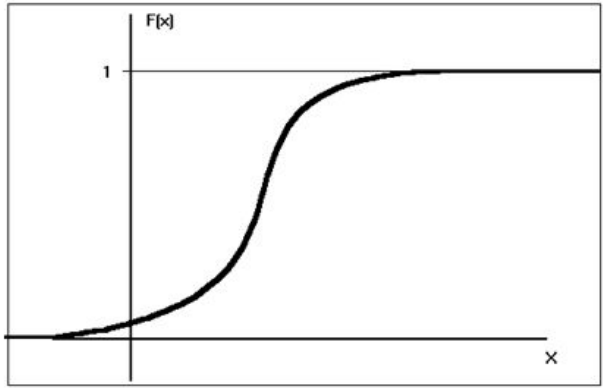
Función de Distribución Acumulada

La **Función de Distribución Acumulada** de la v.a. X, es la función F: $\mathbb{R} \rightarrow [0,1]$ definida por

$$F(t)=\mathbf{P}(X \leq t) = \mathbf{P}(\{\varpi / X(\varpi) \leq t\})$$



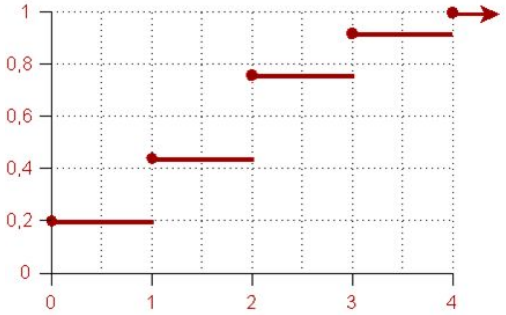
X discreta

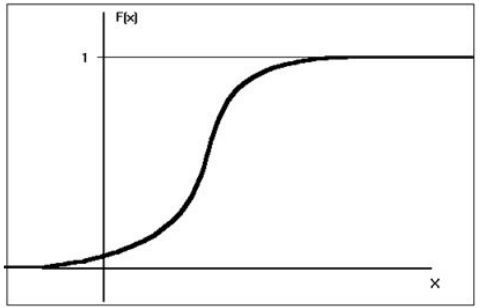


X continua

Función de densidad

FDA: F(t)=P(X≤t)





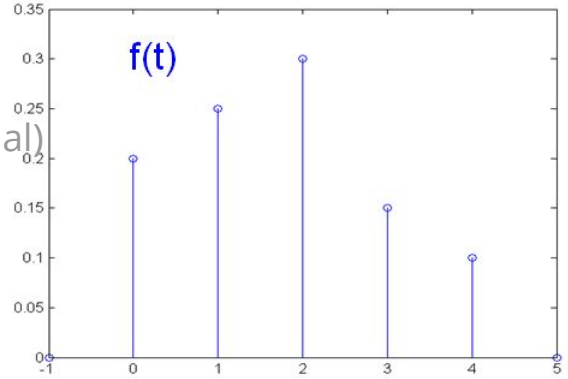
FDA: F(t)=P(X≤t)

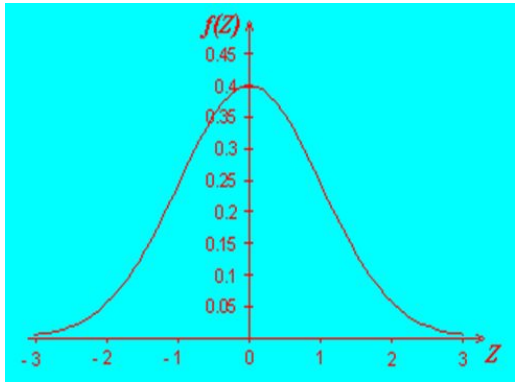
X discreta

X continua

densidad discreta
(probabilidad puntual)

f(t)=P(X=t)





densidad:

$f(t)=F'(t)$

$F(t)=\int^{t} f(x)\,dx$

Propiedades de función de densidad

1) $f(t) \geq 0$ para todo t

2) $\int f(t)\, dt = 1$ para variables continuas y (entre $-\infty$ y $+\infty$)

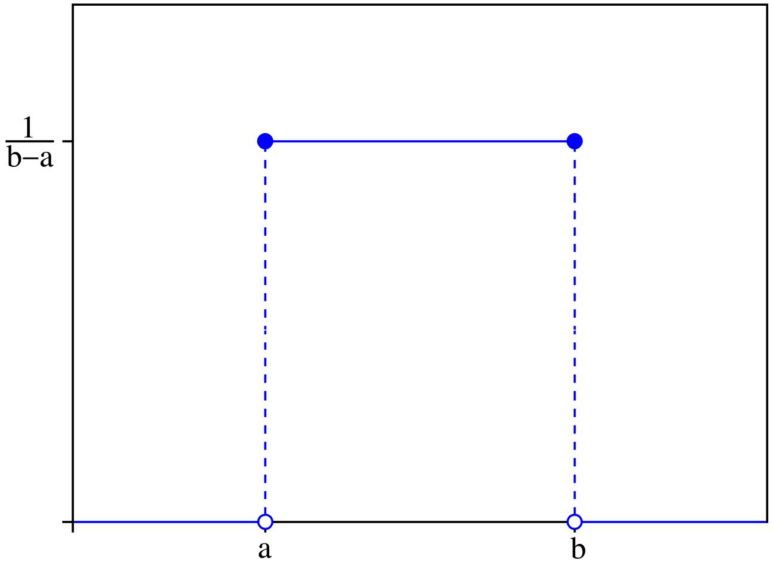
2) $\sum f(t) = 1$ para variables discretas (para todos los valores)

cualquier función que cumple con 1 y 2 es una función de densidad de alguna v. a.

Distribución Uniforme

X v.a. tiene **distribución uniforme** si su función **densidad** es

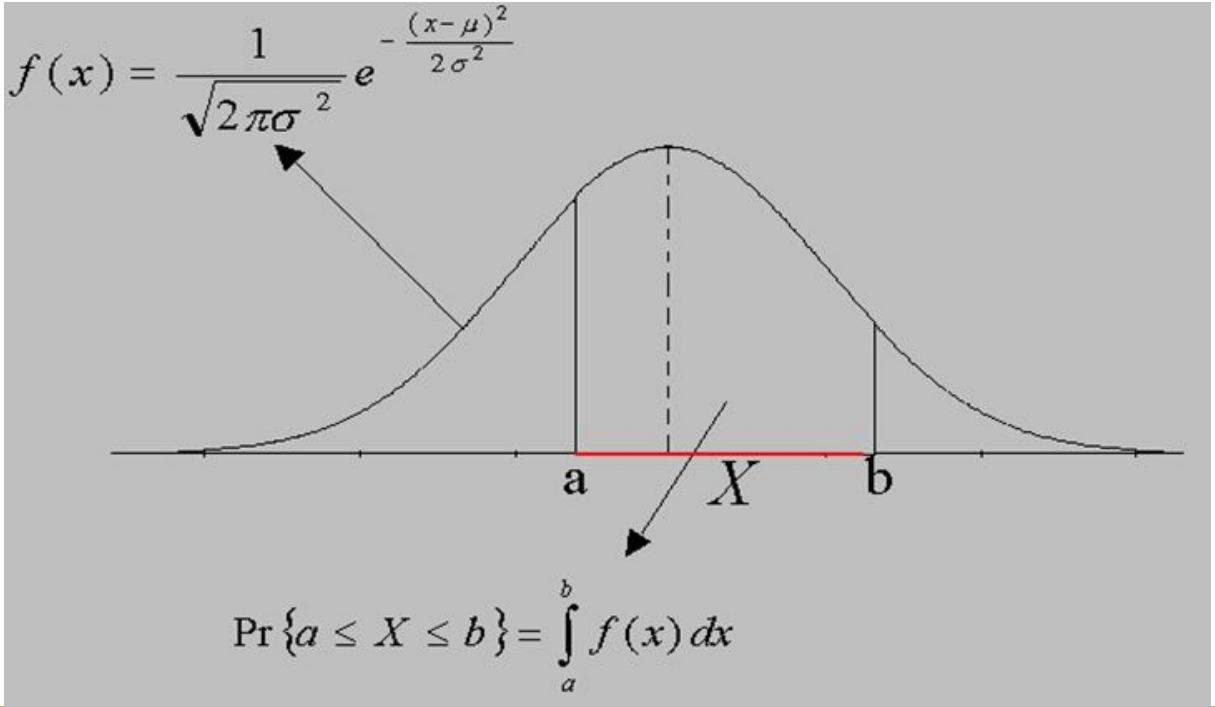
$f(t)=1/(b-a)$ si $a \leq t \leq b$, 0 c.c.



Notación X~U(a,b), a<b parámetros

Distribución Normal o Gaussiana

X v.a. continua tiene distribución normal (Gaussiana) si su función de densidad es la siguiente:



Con $\mu \in R$ y $\sigma^2 \in (0,\infty)$

parámetros

Notación $X \sim N(\mu,\sigma^2)$

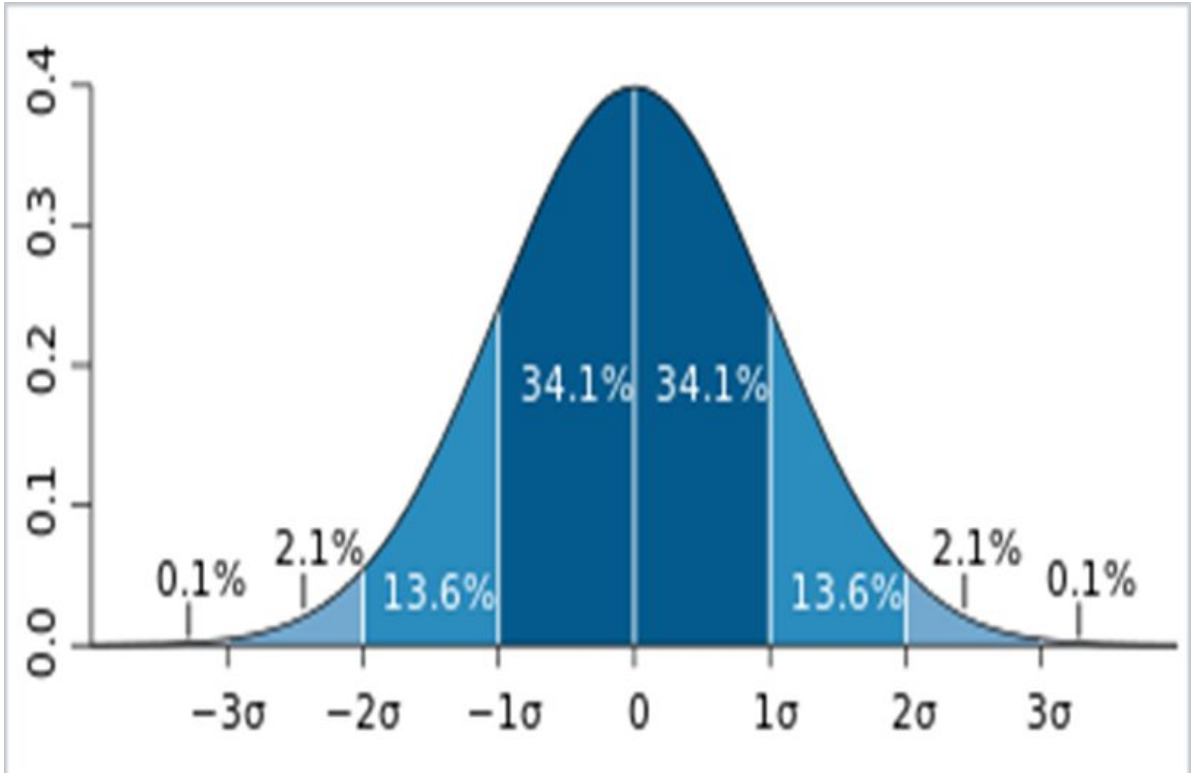
Distribución Normal o Gaussiana

$X \sim N(0,\sigma^2)$

si además $\sigma^2=1$

X~N(0,1), se dice

Normal Estándar

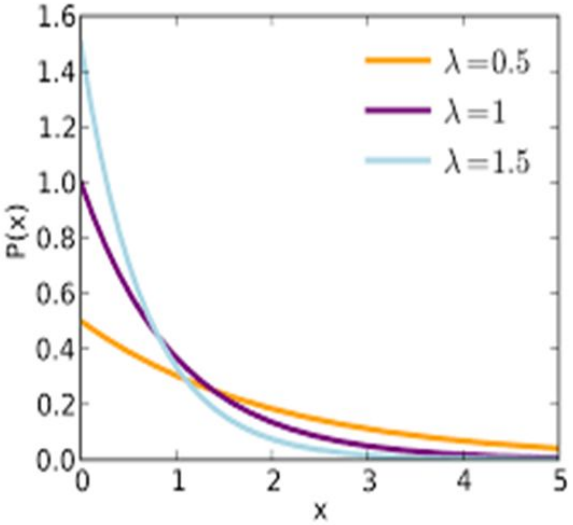


Distribución Exponencial (caso especial de Gamma)

X v.a. tiene distribución

exponencial si su densidad es:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & si \quad x \geq 0 \\ \\ 0 & si \quad x < 0 \end{cases}$$

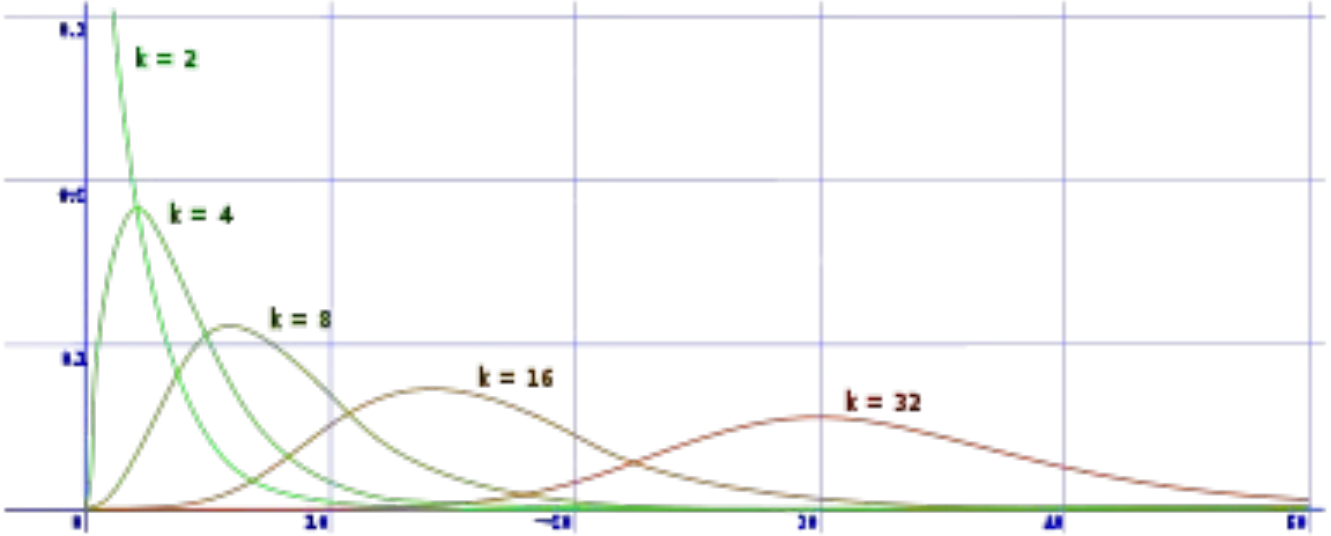


Notación X~Exp(λ), λ>0 parámetro

suele utilizarse para modelar tiempo de espera

Distribución Chi Cuadrado

Diremos la v.a. X tiene distribución Chi- cuadrado con k grados de libertad. Notación X~ χ_k^2 si su función de densidad está dada por:



Medidas estadísticas de una v.a. o de una densidad

X v.a. numérica con densidad f

- **Media o Esperanza** de X (Medida de posición):

$\mu=E(X)=\int t\, f(t)\, dt$ ó $\mu=E(X)=\sum t\, f(t)$,promedio ponderado por la densidad ($\mu \in R$)

- **Varianza** (Medidas de dispersión):

$\sigma^2 =Var(X)= E((X-\mu)^2)=\int(t-\mu)^2\, f(t)\, dt$ ó $\sigma^2=E((X-\mu)^2)=\sum(t-\mu)^2\, f(t)$ ($\sigma^2 \in R^+$)

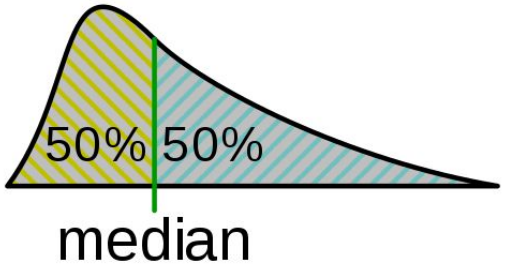
En una va con densidad normal coinciden con los parámetros μ y σ^2 respectivamente

Mediana

Se ordena la muestra de menor a mayor: $x_{(1)},\ldots x_{(n)}$ y se calcula...

Mediana Muestral vs

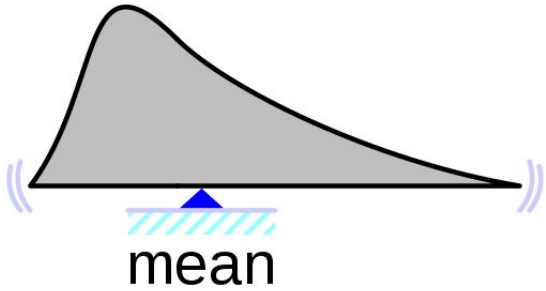
Mediana de una v.a. X, o de su densidad es x_e tal que $P(X\leq x_e)=P(X\geq x_e)$



Media

Media Muestral $\sum_{i=1}^{n} x_i /n$, (promedio) vs

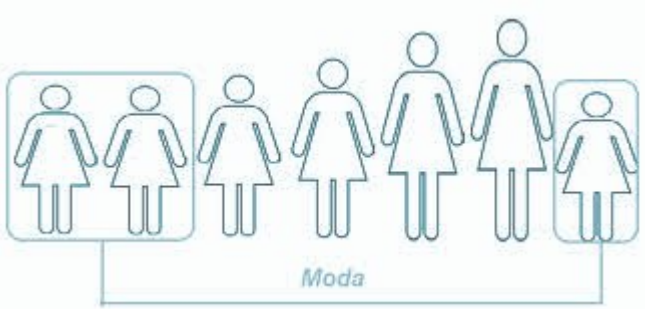
Media o Esperanza de una v.a. X, $\mu=E(X)= \int t\, f(t)\, dt$ ó $\mu=E(X)=\sum t\, f(t)$

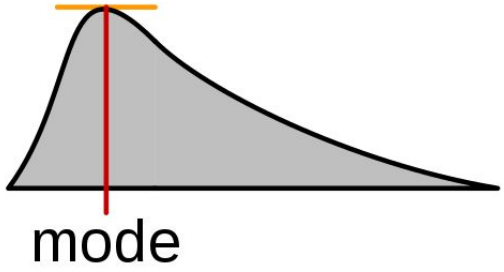


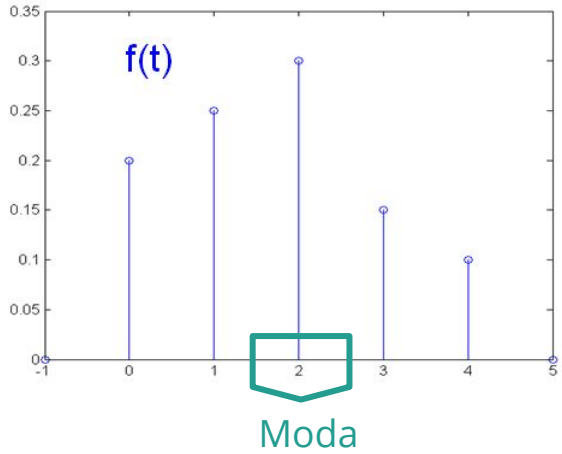
Moda

Resultado (o intervalo) con mayor frecuencia en la **muestra**. vs

Valor con **mayor probabilidad** o **densidad** x_o tal que $f(x_o) \geq f(x)$, p/ todo x

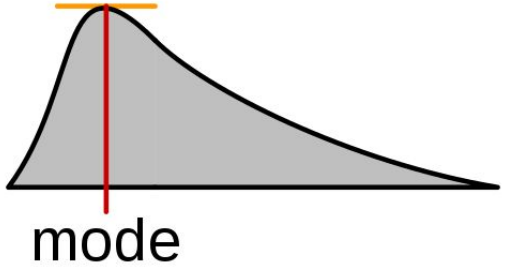




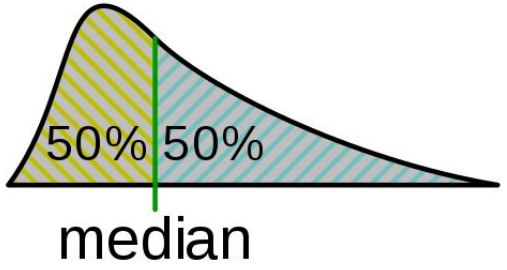


Comparación de Medidas

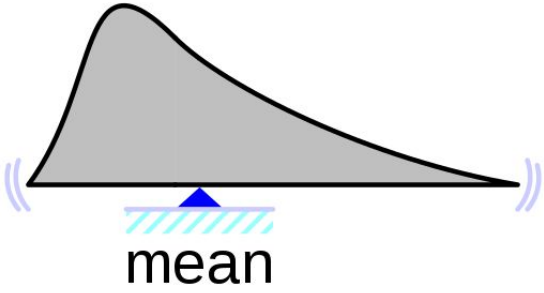
Moda:



Mediana:



Media:

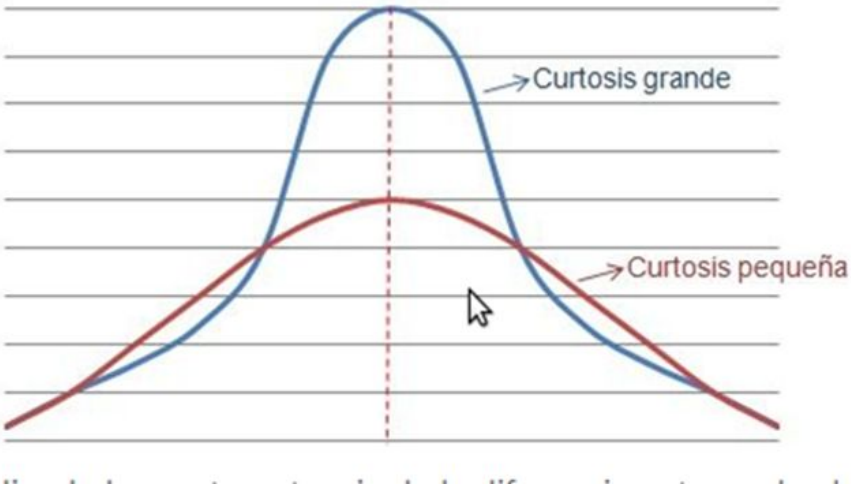


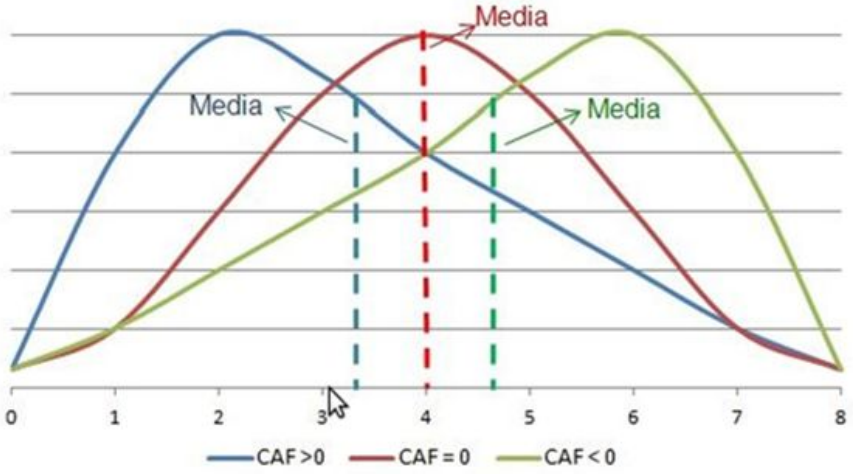
https://upload.wikimedia.org/wikipedia/commons/thumb/3/33/Visualisation_mode_median_mean.svg/512px-Visualisation_mode_median_mean.svg.png

Otras Medidas, del modelo (de una v.a.)

Dada una **función de densidad f** (de una v.a. X) se define:

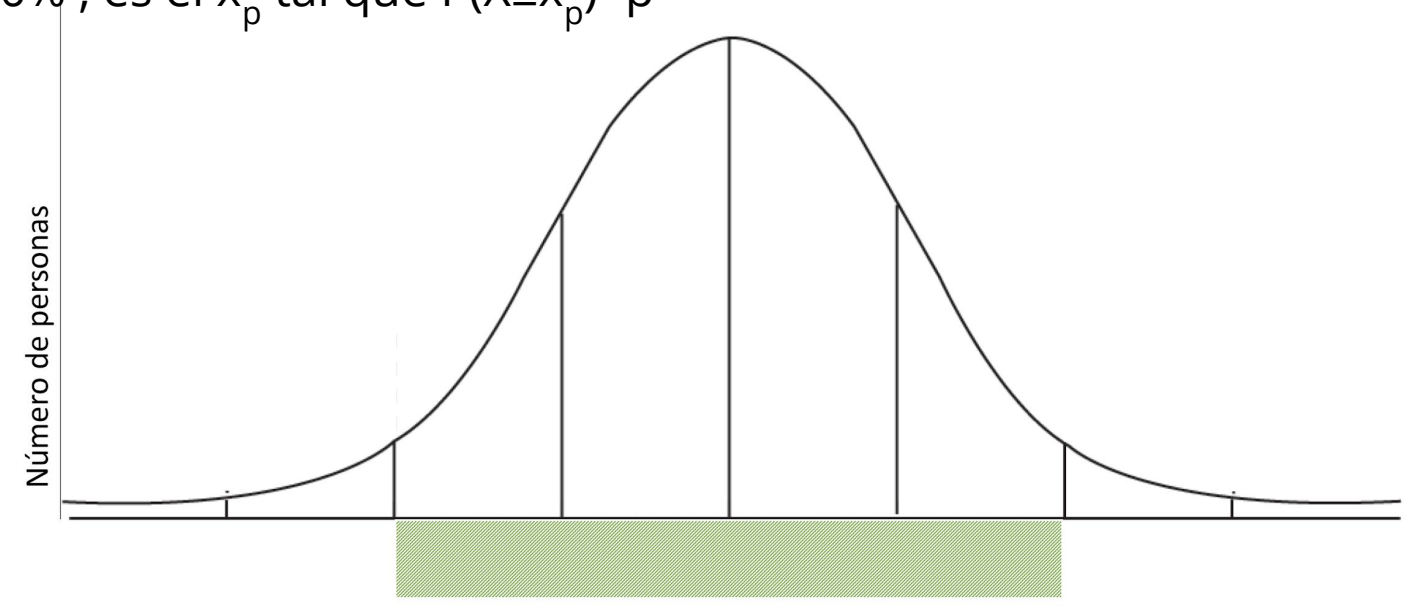
Desvío: $\sigma = (\sigma^2)^{1/2} = (Var(X))^{1/2}$ -**Kurtosis**: $E((X-\mu)^4)/\sigma^4$ -**Sesgo/Asimetría**: $E(X-\mu)^3/\sigma^3$





Percentiles

El percentil es una medida de posición. El p-ésimo percentil o percentil p$_x$100% , es el x_p tal que $P(X \leq x_p)=p$



link1

Algunas propiedades de v.a. y su distribución

- Si $X \sim N(\mu,\sigma^2)$ y $Z=(X-\mu)/\sigma$, entonces $Z \sim N(0,1)$

- Si $Z \sim N(0,1)$, entonces $Z^2 \sim \chi_1^2$ Chi cuadrado con 1 gl

Población y muestra

Cuando recogemos los datos muchas veces es imposible relevar la característica de interés de todo el grupo entero (población) o universo, se examina una pequeña parte del grupo, llamada muestra.

Se denotan los n datos de una muestra: $x_1,...x_n$ (observaciones/repeticiones de la v.a. X)

Medidas a partir de datos↔Medidas <u>muestrales</u>

Sean los n datos de una muestra: $x_1,...x_n$ (observaciones de la v.a.)

Media muestral (promedio): $x_M=\sum_{i=1}^{n} x_i /n= \overline{X}$

Varianza muestral : $\sum_{i=1}^{n} (x_i -x_M)^2/n$

Asimetría muestral $$CA_F = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^3}{N \cdot S_x^3}$$

Curtosis muestral $$Curtosis = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^4}{N \cdot S_x^4} - 3$$

siendo $\overline{x}$ la media y S_x la desviación típica

Tendencia

La tendencia habitual si se tiene una variable descrita en los términos de la Media±Desviación estándar es a hacer aquellas típicas inferencias que **sólo son ciertas si la variable se ajusta bien a la distribución normal**:

- Media±1DE supone el 68.5% aproximadamente de la población,
- Media±2DE supone el 95% aproximadamente de la población
- Media±3DE supone el 99.5% aproximadamente de la población

Bondad de ajuste

Resumen la discrepancia entre los valores observados y los valores esperados en el modelo de estudio.

- Gráficos QQ (Quantil modelo vs Quantil muestral)

Dentro de los test más usados para normalidad:

- Test de Kolmogorov-Smirnov (Test KS)

(En próxima semana veremos Test de Hipótesis)

Notebook

03_Distribuciones.ipynb