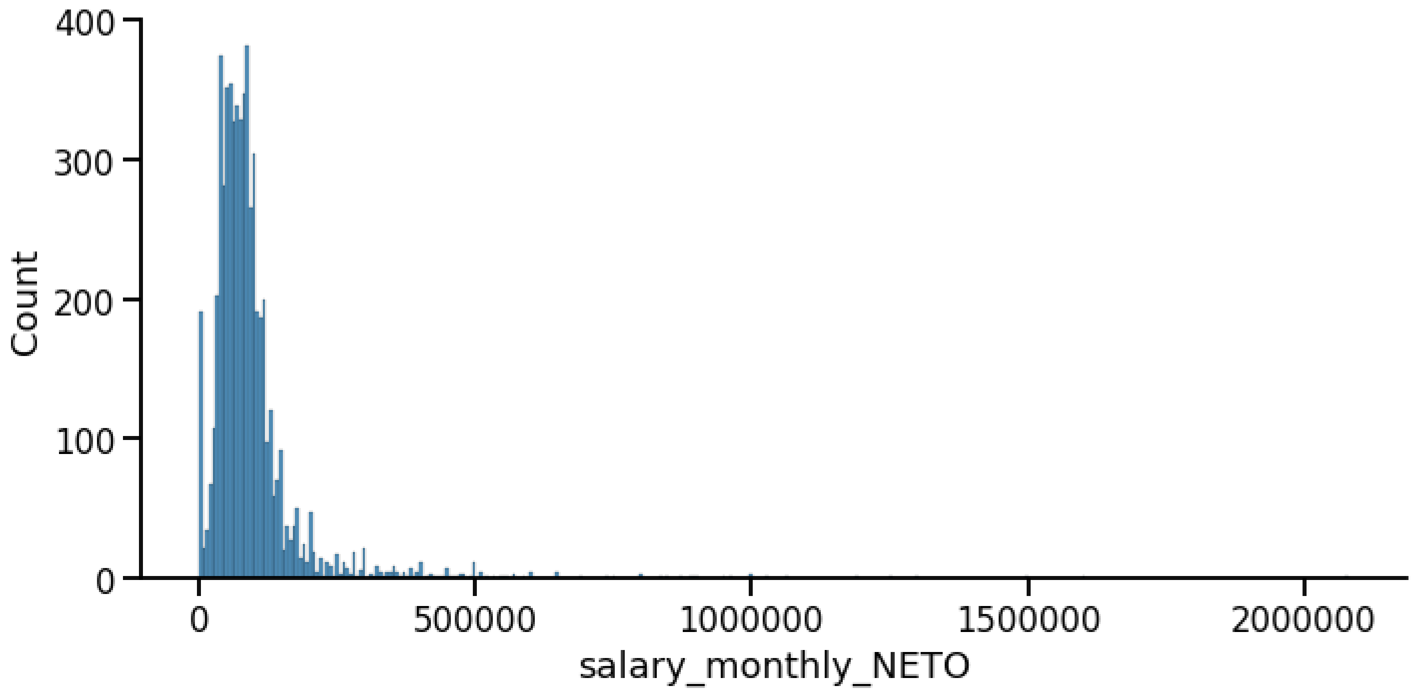
Procesamiento de Señales II
Ciencia de Datos II

Version 2022-2

Análisis y Visualización Datos y Modelos

Dr. José Ramón Iglesias
DSP-ASIC BUILDER GROUP
Director Semillero TRIAC
Ingenieria Electronica
Universidad Popular del Cesar



¿Más herramientas para describir el “comportamiento” de los datos de la columna salary_monthly_NETO?

Estadística Descriptiva

Estadística Descriptiva

Herramientas para resumir un conjunto de datos (modelados como realizaciones de una variable aleatoria), a través de ciertas medidas numéricas.

Representa la información de una manera distinta para facilitar su interpretación, pero no permite realizar predicciones o inferencias

Análisis de frecuencias

¿Cuánto ocurre cada uno de los valores de una v.a.?

Medidas de tendencia central

¿Cuál es el valor más representativo de una v.a.?

Medidas de dispersión

¿Qué tan alejados están los datos de la tendencia central?

Medidas de tendencia central

Dado un conjunto de
datos numéricos
(muestra)
$x = \{x_1, x_2, ..x_N\}$
se la piensa como
realizaciones
(independientes) de una
v.a. numérica

La **media muestral** (aritmética) o promedio se calcula como:

$$\bar{x} = \frac{1}{N} \sum_{i}^{N} x_i$$

Medidas de tendencia central

Dados datos numéricos

$x = \{x_1, x_2, ..x_N\}$

se los piensa como $x_i = X_i(\omega)$ para algún $\omega_i \in \Omega$

La **mediana** se calcula como:

1. Ordenar las realizaciones de menor a mayor
2. Si N es impar, la mediana es el valor central:

   $\text{mediana}=x_{(N+1)/2}$

1. Si N es par, la mediana es el promedio de los dos valores centrales:

   $\text{mediana}=(x_{N/2}+x_{N/2+1})/2$

Medidas de tendencia central

Dada X una v.a **categórica** y un conjunto de realizaciones

$x = \{x_1, x_2, \dots, x_N\}$

donde $x_i = X(\omega)$ para algún $\omega \in \Omega$, y $N = |x|$

La **moda** son los valores con mayor frecuencia, es decir, los que más se repite.

Sólo hay más de una moda cuando el conteo de dos valores es igual.

Se puede definir de forma análoga la **moda o intervalo modal** para datos numéricos pero no es único (como la visualización del histograma).

Medidas de posición

Dados datos numéricos

$x = \{x_1, x_2, ..x_N\}$

(pensados como realizaciones de una v.a)

El **percentil-k** es el valor x_i tal que el k% de los valores de la muestra son menores a x_i.

No hay una única fórmula para calcular los percentiles, pero en general:

1. Ordenar las realizaciones tal que $x_j \leq x_{j+1}$
2. Seleccionar el elemento de la serie en la posición: menor entero mayor o igual a k*N/100.

Medidas de dispersión

Dados datos numéricos

$x = \{x_1, x_2, ..x_N\}$

(pensados como realizaciones de una v.a)

La **varianza muestral** mide la variación de los datos a través de la distancia cuadrada a la media muestral.

$$v = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

La **desviación estándar** es la raíz cuadrada de la varianza. Está en la misma unidad que los datos.

El **coeficiente de variación** es la desviación estándar dividida la media muestral. Es comparable entre distintas v.a.

Medidas de dispersión

Dados datos numéricos

$$x = \{x_1, x_2, ..x_N\}$$

El rango y el rango intercuartílico miden en qué intervalo se encuentran un cierto porcentaje de los datos.

Rango:

percentil-100 - percentil-0

Rango intercuartílico:

percentil-75 - percentil-25

Q3 - Q1

Usos de los percentiles y rangos

- En el caso de la mediana (percentil-50), medir la tendencia central
- Contextualizar el valor de un dato con respecto a otros

  Una persona de sexo femenino de 6 años mide 95cm

  Está en el 10% de personas con menor estatura del mismo grupo. [Curva]

- Identificación y eliminación de valores extremos

Demo con Notebook

CIED2 Datos y Modelos.ipynb

En **una frase**:
¿Cuánto cobran l@s programadores en un país como Argentina?

¿Respuestas?
¿Qué pregunta respondimos en realidad?

¿Cuánto cobran l@s programadores experimentados en un país como Argentina?

¿Afecta el nivel de estudios en el salario de l@s programador@s en un país como Argentina? ¿Cómo?

Ejercicio

Seguir el proceso de análisis propuesto:

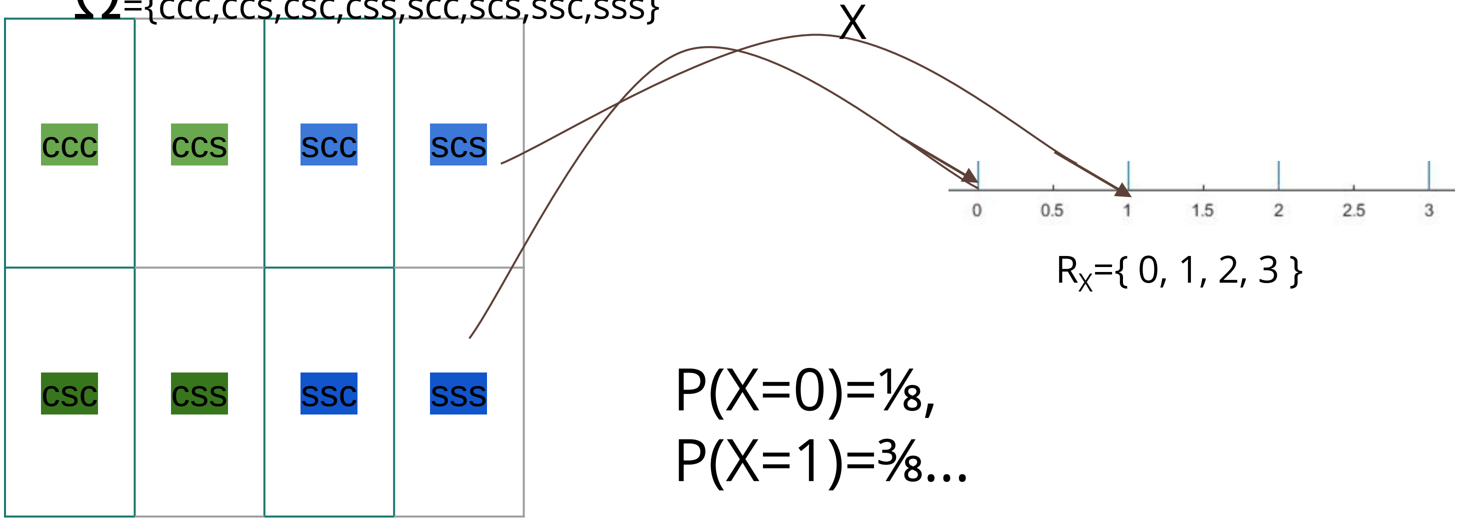
1. Hipótesis
2. Análisis de v.a.
3. Experimento

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

Variable Aleatoria (discreta numérica)

X= cantidad de caras en 3 tiradas de moneda.

Ω={ccc,ccs,csc,css,scc,scs,ssc,sss}



R_X={ 0, 1, 2, 3 }

P(X=0)=⅛,
P(X=1)=⅜...

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)]
```

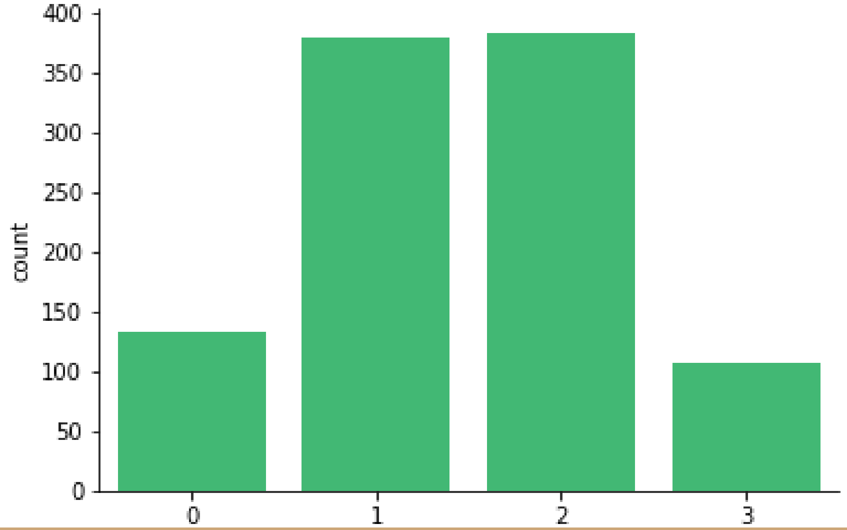
```
[(0, 0.132), (1, 0.379), (2, 0.383), (3, 0.106)]
```

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)]
```

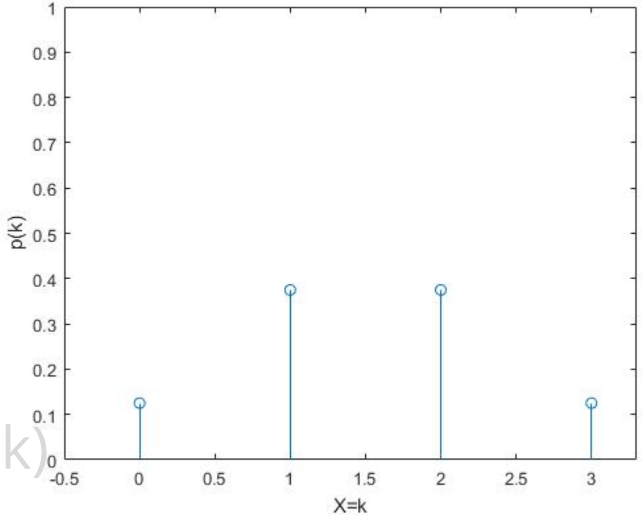
la Proporción de la muestra tal que X=k, estima la probabilidad P(X=k), p/ k=0,1,2,3



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

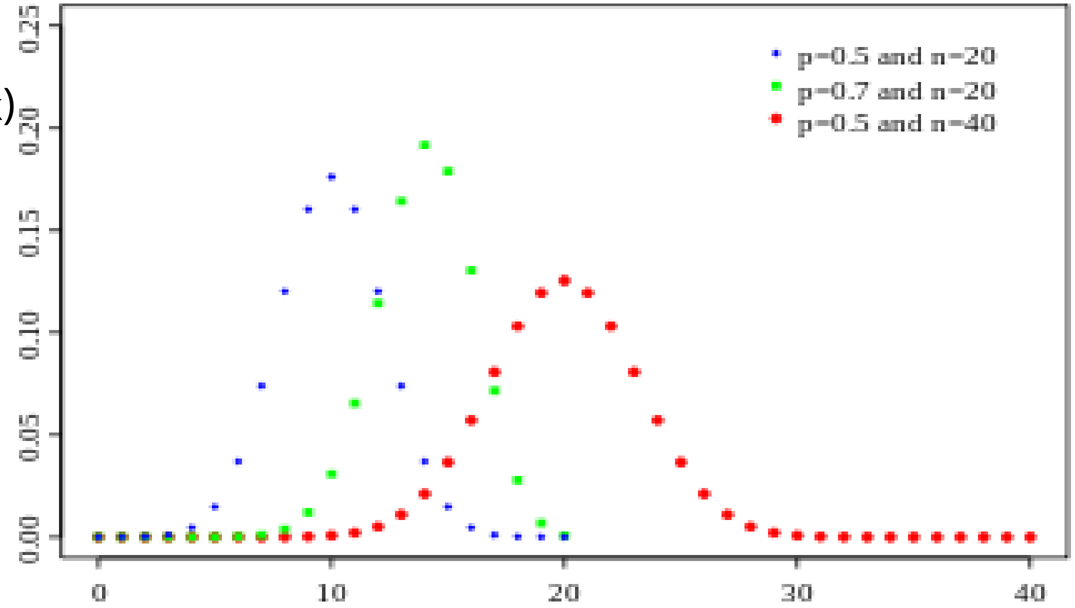
Sea X la v. a. discreta modela: cantidad de “éxitos” en una n-upla

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

k=0,1,...,n

p=probabilidad de “éxito”.

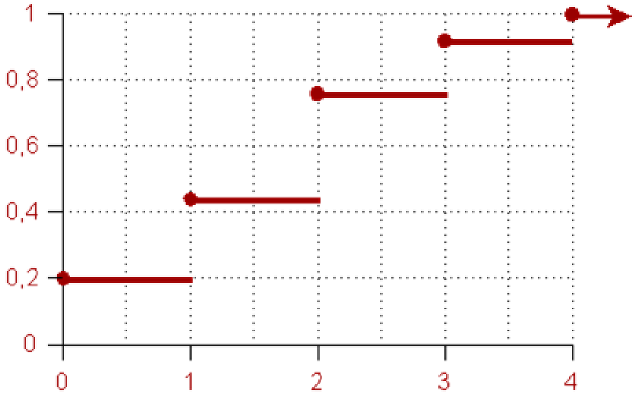
X~B(n,p), ejemplos?

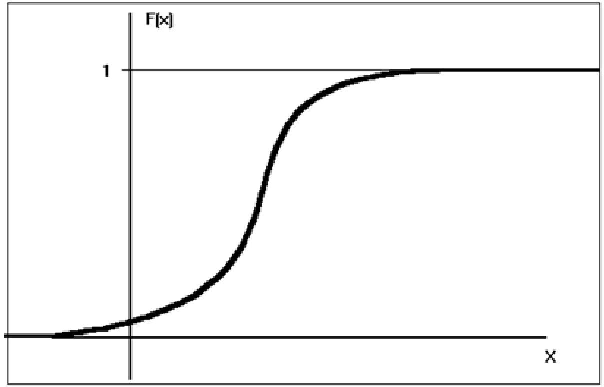


Función de Distribución Acumulada

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

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



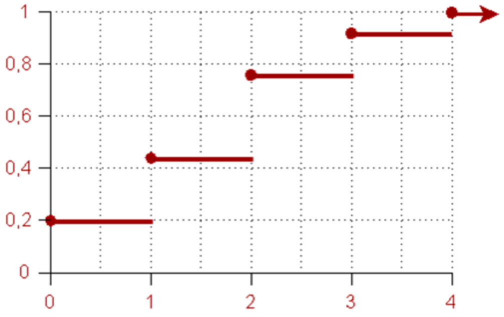


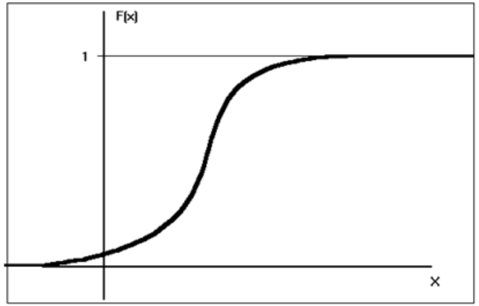
X discreta

X continua

Función de densidad

FDA: F(t)=P(X≤t)





FDA: F(t)=P(X≤t)

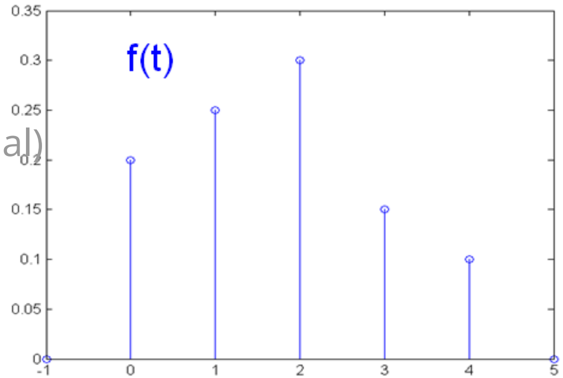
X discreta

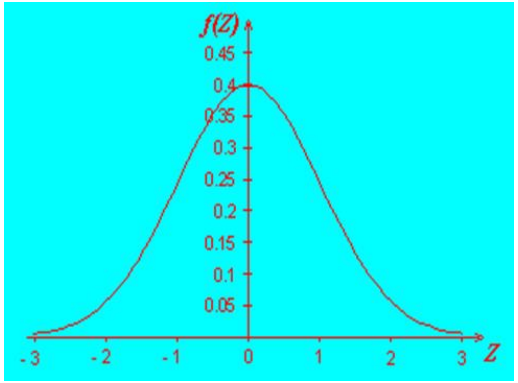
X continua

densidad discreta

(probabilidad puntual)

f(t)=P(X=t)





densidad:

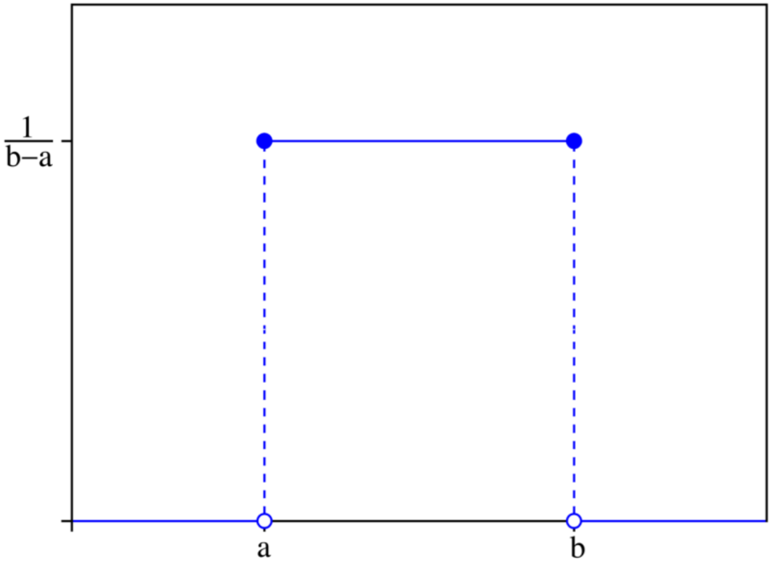
f(t)=F'(t)

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

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.

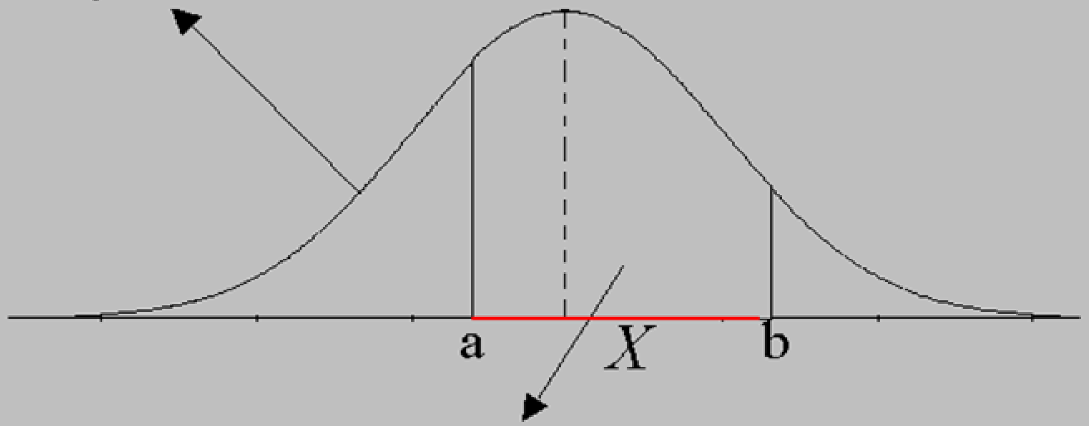


Notación $X \sim U(a,b)$, $a<b$ parámetros

Distribución Normal o Gaussiana

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

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$\Pr\{a \leq X \leq b\} = \int_a^b f(x)\,dx$$

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

parámetros

Notación $X \sim N(\mu,\sigma^2)$

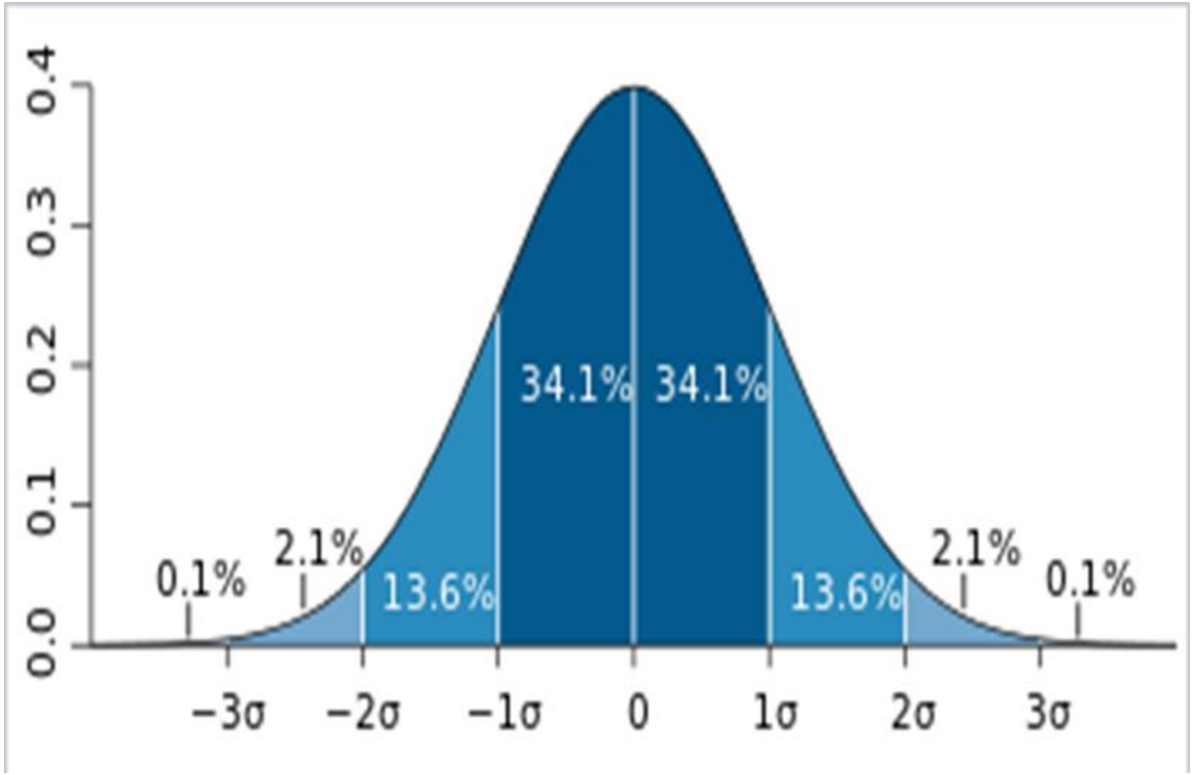
Distribución Normal o Gaussiana

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

si además $\sigma^2=1$

$X \sim N(0,1)$, se dice

Normal Estándar

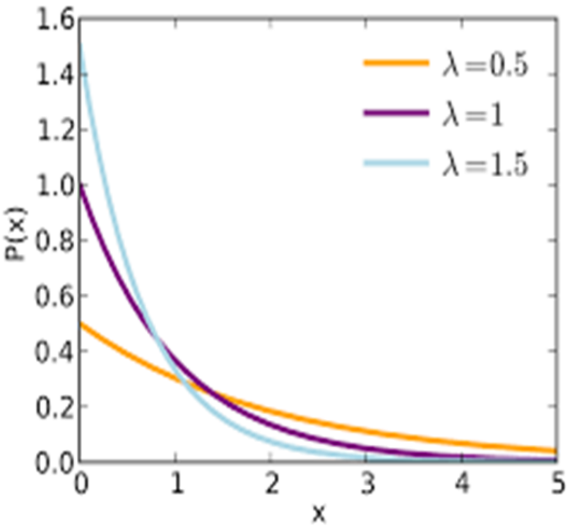


Distribución Exponencial (caso especial de Gamma)

X v.a. tiene distribución

exponencial si su densidad es:

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

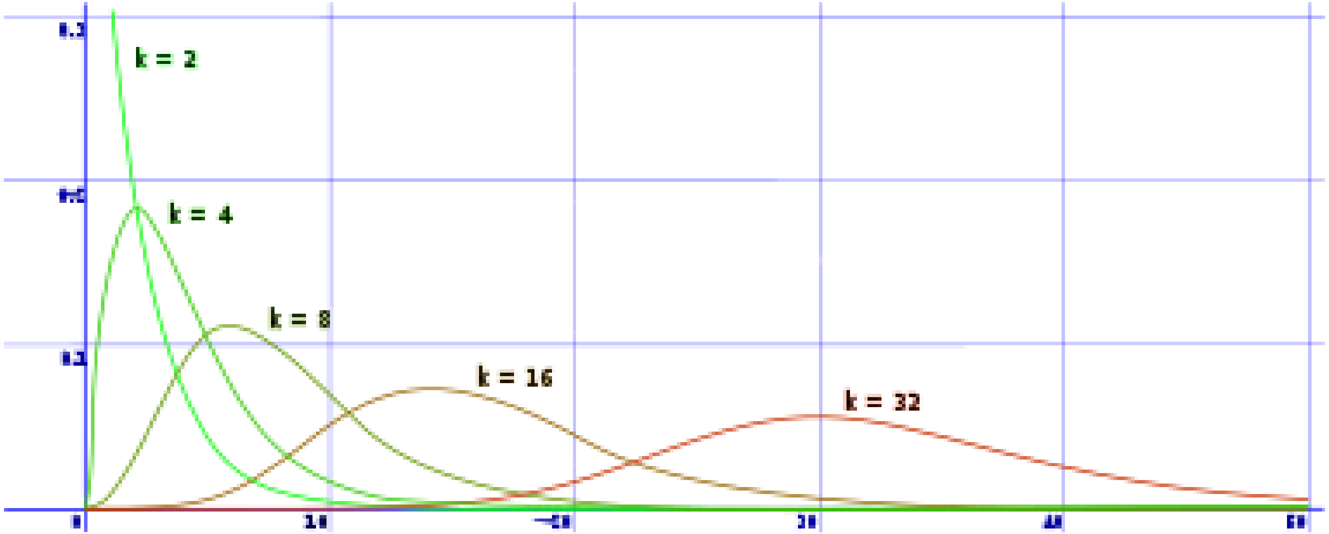


Notación X~Exp(λ), λ>0 parámetro

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):

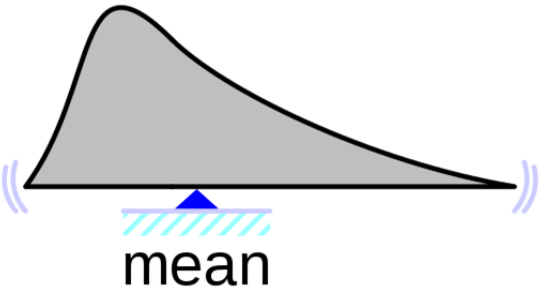
$\sigma^2=Var(X)=E((X-\mu)^2)=\int (t-\mu)^2 f(t)\, dt$ ó $\sigma^2=E((X-\mu)^2)=\sum (t-\mu)^2 f(t)$ ($\sigma^2\in R^+$)

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

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 $\mu=E(X)=\sum t\, f(t)$

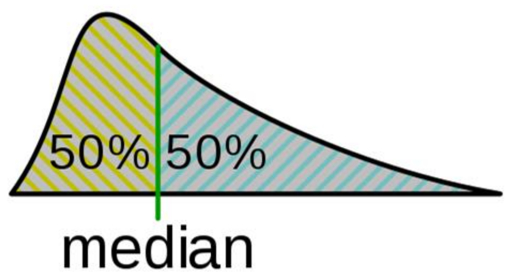


Mediana

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

Mediana Muestral vs

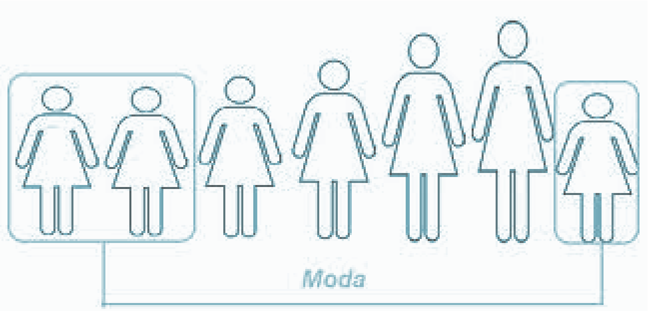
Mediana de una v.a. X, o de su densidad es x_e tal que $P(X \leq x_e) = P(X \geq x_e)$

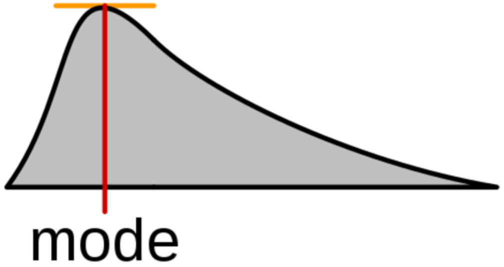


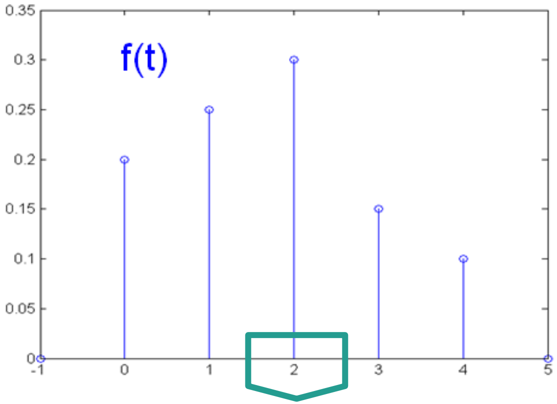
Moda

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







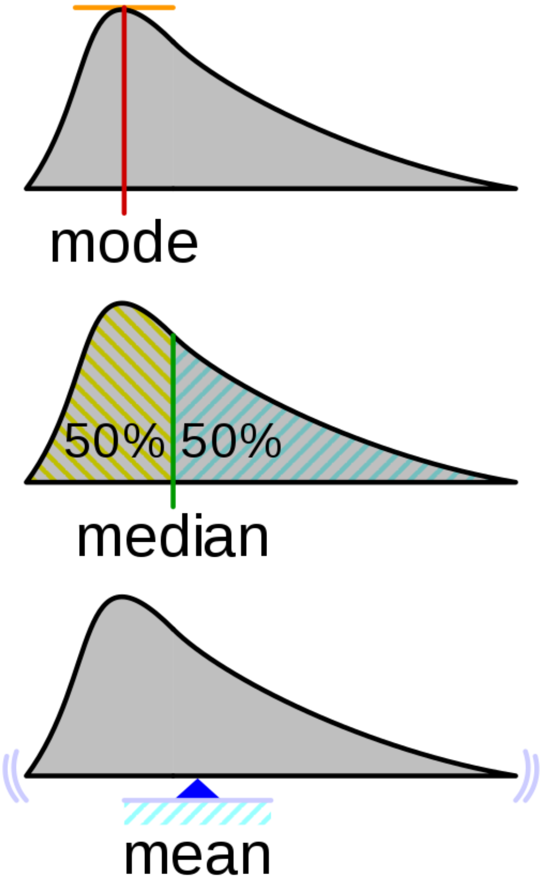
Moda

Comparación de Medidas

Moda:

Mediana:

Media:

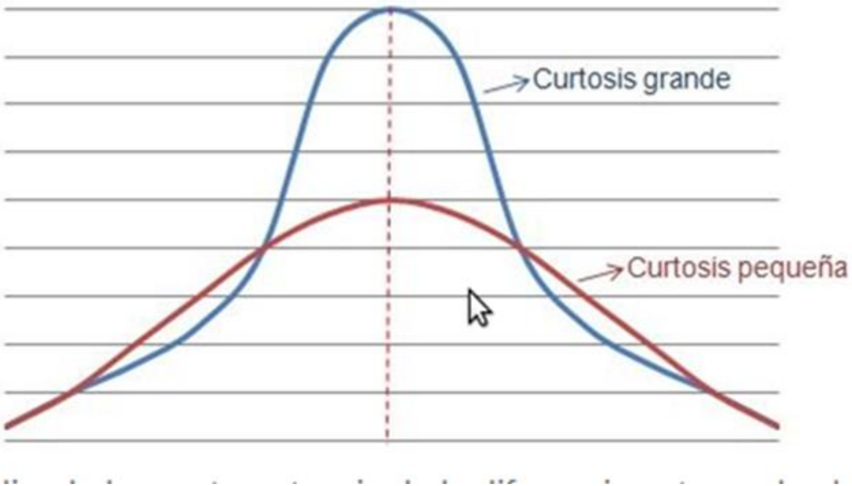


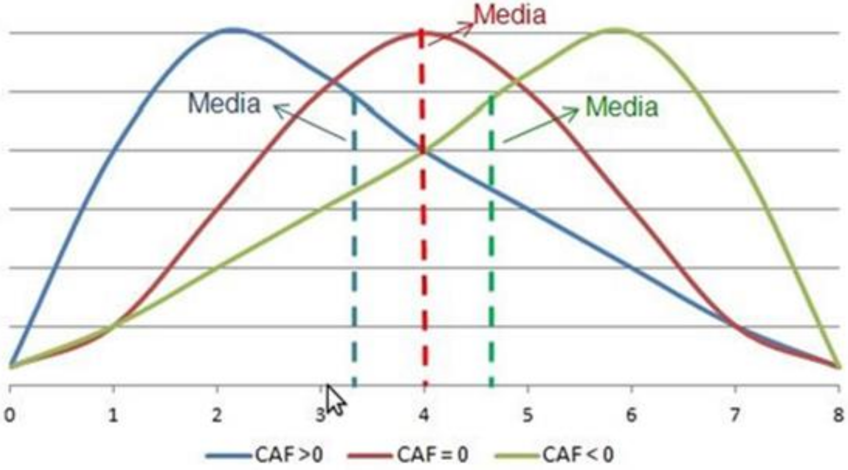
https://upload.wikimedia.org/wikipedia/commons/thumb/3/33/Visualisation_mode_median_mean.svg/512px-Visualisation_mode_median_mean.svg.png

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

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

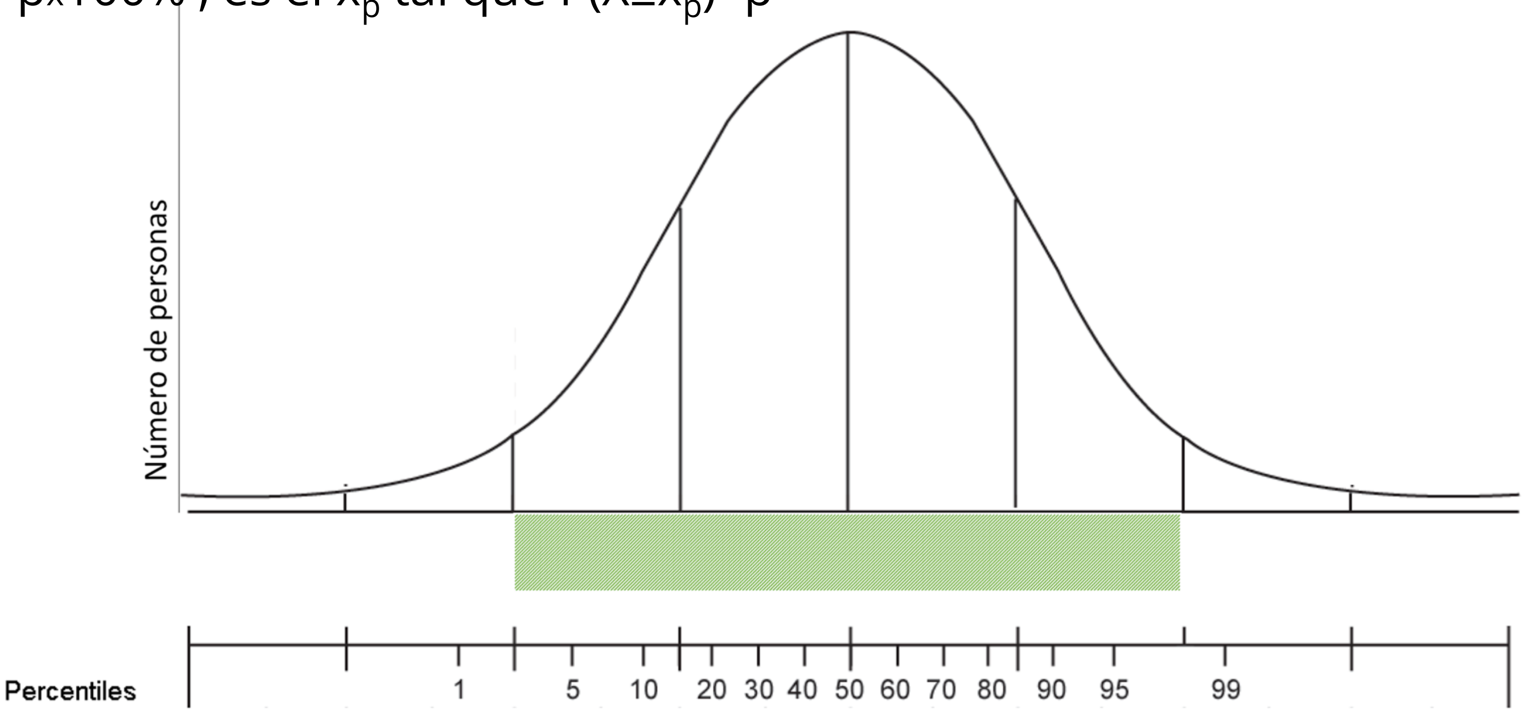
Desvío: $\sigma = (\sigma^2)^{1/2} = (Var(X))^{1/2}$ -**Kurtosis**: $E((X-\mu)^4)/\sigma^4$ -**Sesgo/Asimetría**: $E(X-\mu)^3/\sigma^3$





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: $x_1,\ldots x_n$ (observaciones/realizaciones de la v.a. X)

Medidas a partir de datos ↔ Medidas muestrales

Sean los n datos de una muestra: $x_1, \ldots x_n$ (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 los datos están descritos en los términos de $\overline{X}$ y S_X (desvío), es hacer aquellas típicas inferencias que **sólo son ciertas si la distribución de los datos se ajusta bien a la distribución normal**:

- $\overline{X} \pm S_X$ supone el 68.5% aproximadamente de la población,
- $\overline{X} \pm 2S_X$ supone el 95% aproximadamente de la población
- $\overline{X} \pm 3S_X$ 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 muestral vs Quantil modelo)

Dentro de los test más usados para normalidad:

- Test de Kolmogorov-Smirnov (Test KS)

(En próxima semana veremos Test de Hipótesis)