Estadística Espacial

Juan Carlos Alvarez Forero
jk.alvarez@gmail.com

Universidad Distrital Francisco José de Caldas
Especialización en Sistemas de Información Geográfica

2014-II

AGENDA

1. El Curso
    - Aspectos Generales
    - Contenido Bibliografía
    - Evaluación
        - Talleres
        - Debates
        - Proyecto Final
2. Introducción a la estadística espacial
3. Probabilidad
    - Definiciones básicas
    - Eventos
    - Distribuciones de probabilidad
        - Caso Discreto
        - Caso continuo
        - Distribuciones continuas
    - Teorema del límite central

AGENDA

1 El Curso

2 Introducción a la estadística espacial

3 Probabilidad

Aspectos Generales

- Curso de 64 horas (Octubre 8 a Octubre 28)
- Lunes a Viernes de 18:00 a 22:00
- Sábado 08:00 a 12:00
- 4 horas Teórico-prácticas
- Receso 15min
- Reglas de juego

Temas Generales del curso

- Introducción
- Revisión
  - Probabilidad
  - Estadística descriptiva
  - Estadística inferencial
- Interpolacíon espacial

- Patrones puntuales
- Análisis de datos de áreas
- Datos de interacción espacial

Blibliografía

Isaaks, E. H., & Srivastava, R. M.(1989). **Applied geostatistics** (pp. xix-561). Oxford University Press

Bivand, R. S., Pebesma, E. J., & Rubio, V. G. (2008). **Applied spatial data: analysis with R**. Springer.

Hengl, T. (2007). **A practical guide to geostatistical mapping of environmental variables**. www.spatial.org.

Adicional sugerida en cada tema

Evaluación

- Dos pruebas escritas (30% de la nota final) (**15%** cada una) **Octubre 16 y Octubre 25**.
- Un reporte escrito con el resúmen de las actividades desarrolladas en los talleres (15% de la nota final) – **Octubre 27**.
- Debates (**30%** de la nota final) – **Octubre 15, Julio 1 y Julio 9**.
- Un proyecto final compuesto de reporte escrito y presentación (25% de la nota final, Envíos de avance semanal para revisión). **Octubre 28**.
- **Asistencia mínima del 87% para pasar el curso -> 14 clases completas**

Talleres

Clases con ejercicios prácticos en el laboratorio.

- Clase interactiva, Teórico-Práctica
- Se desarrollan diferentes guías prácticas
- Una entrega compilatoria al final del curso.

Debates

Clases donde los alumnos se reúnen en grupos para discutir un artículo relacionado con la aplicación de técnicas de estadística espacial y luego presentan un resúmen al resto de la clase.

- Tres durante el curso
- El mismo día se obtiene la nota

Proyecto Final

- Escogido por los estudiantes para trabajar y presentar en grupos de acuerdo a los contenidos de la materia.
- Presentación del proyecto (breve explicación del proyecto escogido) – **Octubre 11** (5 min máx)
- Informe escrito (Máximo 6 hojas) y presentación final (20 min max) – **Octubre 28.** (No teoría!, Metodología, análisis de resultados)
- Durante el desarrollo del curso hay tiempo específico para trabajar en el Proyecto
- Enviar avances semanalmente, preguntas o inquietudes.
- El software a usar es de libre elección: R (Paquetes específicos), ArcGIS (Extensiones de análisis geoestadístico y espacial), SAGA, SPRING, GSLib, GeoMS, más opciones en Fisher et al (2010) [2]

[2]Fischer, M. M., & Getis, A. (2010). Handbook of applied spatial analysis: software tools, methods and applications. Springer

AGENDA

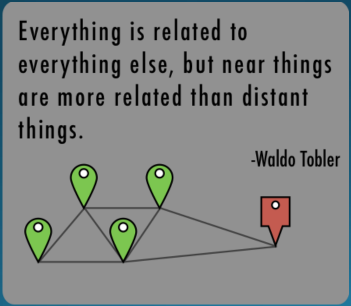
1. El Curso
2. Introducción a la estadística espacial
3. Probabilidad

Estadística Espacial

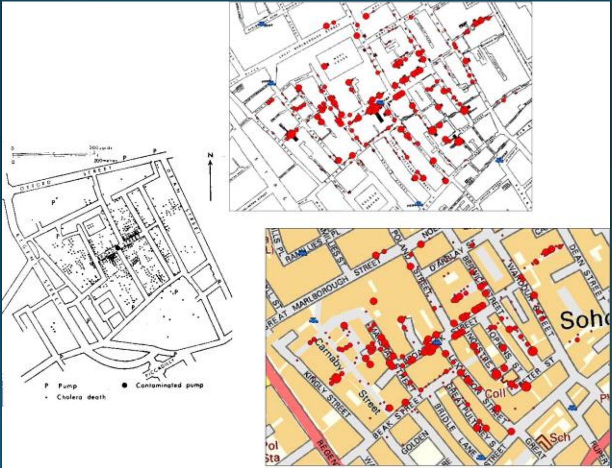
- Provee métodos estadísticos para analizar datos espaciales.
- Modela las relaciones en función de la posición
- Se basa en el supuesto de que los datos se correlacionan de acuerdo a su ubicación espacial

Estadística Espacial

- Provee métodos estadísticos para analizar datos espaciales.
- Modela las relaciones en función de la posición
- Se basa en el supuesto de que los datos se correlacionan de acuerdo a su ubicación espacial



Estadística Espacial



A map of cholera deaths in London, 1840s (John Snow)

Estadística Espacial

Se divide fundamentalmente en tres ramas:

- **Patrones puntuales:** Datos de un suceso puntual sobre el espacio.
- **Areas - Datos Lattice:** Datos agregados por polígonos (Discretos).
- **Geoestadística:** Datos espaciales con variación continua de los cuales solo se tiene una muestra.

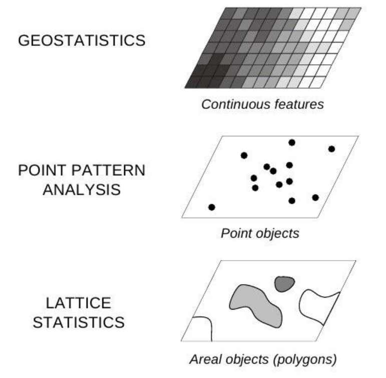


Figure: (Hengl,2009)

Patrones Puntuales

Es un proceso estocástico en el cual cada una de sus realizaciones contiene un número finito sobre un plano

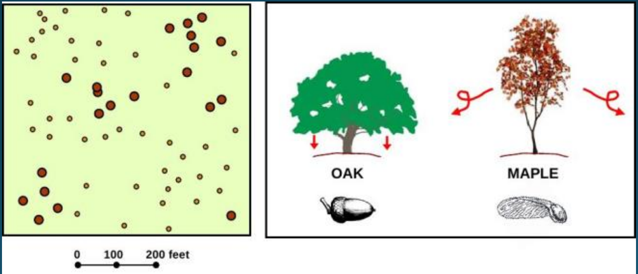


Figure 5.1. Section of Forest

Figure 5.2. Patterns of Seed Dispersal

Figure: http://www.seas.upenn.edu/ ese502/

Patrones Puntuales

El estudio se concentra básicamente en establecer si la posición espacial de ese proceso tiene algún patrón de comportamiento.

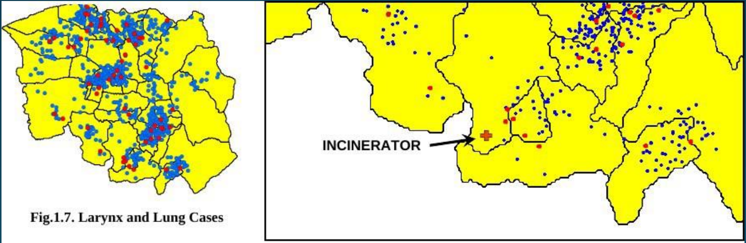


Fig.1.7. Larynx and Lung Cases

Fig.1.9. Incinerator Location

Figure: http://www.seas.upenn.edu/ ese502/

Patrones Puntuales

A partir de la modelación estadística, se establece si es significativamente más agrupado que aleatorio.

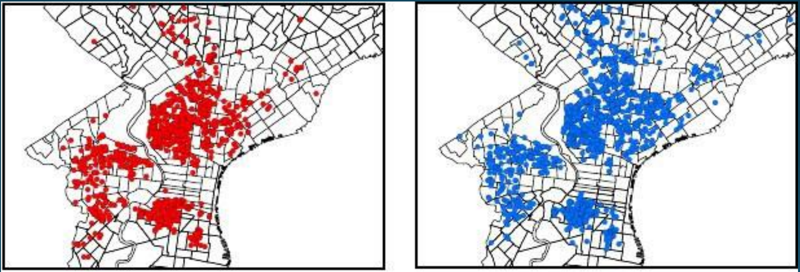


Fig.1.4. Off-Site Owners

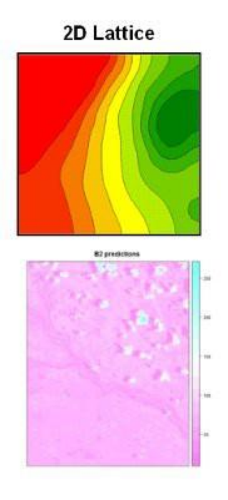
Fig.1.5. On-Site Owners

Figure: http://www.seas.upenn.edu/ ese502/

Datos Lattice

- Los datos son agregados de acuerdo a ciertos límites espaciales

- Se establecen matrices de pesos para representar las características de las relaciones de los elementos



Datos Lattice

Busca modelar la correlación espacial entre las variables de este tipo

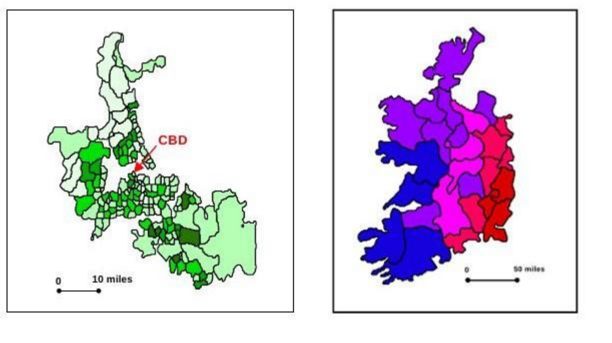


Figure 1.3. Raw Population Data   Figure 1.7. Blood Group A Percentages

Figure: http://www.seas.upenn.edu/ ese502/

Geoestadística

Busca modelar el comportamiento de datos continuos a partir de una muestra de puntos.

- Es la rama más conocida de la estadística espacial

- El método es "óptimo" cuando los datos se distribuyen normalmente y son estacionarios

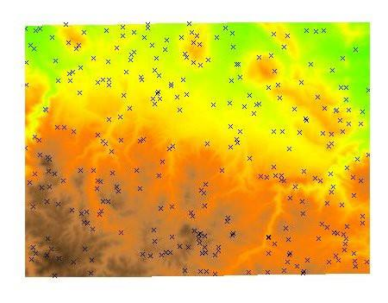


Figure: http://www.utexas.edu

Agenda

Definiciones Básicas

Aleatorio:
Que depende del azar o de la suerte

Experimento aleatorio:
Es un proceso del cual no se puede predecir su resultado. (Ej: Lanzar una moneda, un dado, jugar la loteria..etc) Se debe poder repetir y se deben poder conocer todos los posibles resultados.

Espacio muestral:
Conjunto formado por todos los posibles resultados de un experimento aleatorio

Aleatoriedad

Variable aleatoria:
Es una variable que puede tomar diferentes valores de acuerdo a sus probabilidades y determinados por el resultado de un experimento aleatorio. Es una función

Evento aleatorio:
Cada acontecimiento que ocurrirá o no dependiendo del azar. Es cada uno de los elementos o subconjuntos del espacio muestral de un experimento.

Probabilidad

Es la ciencia que permite cuantificar los posibles resultados de un experimento aleatorio.

Una ruleta tiene 4 divisiones de colores: azul, amarillo, rojo y verde. Cual es el chance de que pare en el color azul al girarla? Cual es el chance de que pare en rojo?

Probabilidad

Es la ciencia que permite cuantificar los posibles resultados de un experimento aleatorio.

Una ruleta tiene 4 divisiones de colores: azul, amarillo, rojo y verde. Cual es el chance de que pare en el color azul al girarla? Cual es el chance de que pare en rojo?

Probabilidad

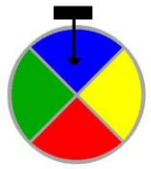
Experimento: girar la ruleta

Evento: un evento es que la ruleta pare en el color azul

Variable: el color que queda al detenerse la ruleta

Espacio muestral: El conjunto formado por azul, amarillo, rojo y verde.

Probabilidad: el chance de que salga uno de los colores. Es de 0.25



Probabilidad

Considerando un experimento aleatorio que tiene un espacio muestral finito:

- La probabilidad de ocurrencia de un evento es un número real en el rango comprendido de 0 a 1.
- Cuanto más se pueda presentar el evento, su probabilidad de será más cercana a 1, en el caso contrario, será más cercana a 0

Adapt. de Salinas H., Estadística y Probabilidad Clases, UDA, 2012

Probabilidad

- La suma de todas las probabilidades de cada uno de los elementos del espacio muestral es igual a 1
- La máxima probabilidad de ocurrencia de un evento es 1
- Para eventos que estén fuera del espacio muestral la probabilidad de ocurrencia es 0

Adapt. de Salinas H., Estadística y Probabilidad Clases, UDA, 2012

Probabilidad

Para calcular la probabilidad $P(A)$ de un evento A, se suman todas las probabilidades que se asignan a los elementos del espacio muestral que ocurran en el evento A.

$$P(A) = \frac{\#\ \text{de casos favorables}}{\#\ \text{de casos posibles}}$$

Si todos los elementos de espacio muestral tienen la misma probabilidad de ocurrir

Conteo

En probabilidad se requiere contar el número de formas que un evento puede ocurrir.

Ejemplo: un experimento que está compuesto por varias etapas en las cuales el número de resultados *m* en la n-ésima etapa es independiente a los resultados de las etapas previas.

El número *m* puede ser diferente para cada etapa y es necesario contar todas las opciones posibles que el experimento puede tener como resultado.

Combinaciones

La combinación corresponde al número de posibles arreglos de n elementos en grupos de r. En este caso no importa el orden de los elementos. Se denota nCr. Se calcula:

$$nCr = \frac{n!}{(n-r)!r!}$$

Ejemplo: De cuantas maneras se pueden sentar 10 personas si solamente hay cuatro sillas

Permutaciones

Una permutación corresponde al número de posibles arreglos de un conjunto de *n* elementos en subconjunto de *r* donde:

- Entran todos los elementos
- Hay dos tipos: con y sin repetición.
- Importa el orden de los elementos

Una permutación es una combinación ordenada.

Permutaciones

Se calcula:

$nPr = \frac{n!}{(n-r)!}$ cuando no se permite repetición

$nPr = n^r$ cuando se permite repetición

Ejemplos: El podium de ganadores en una carrera
Una clave bancaria

Relación entre eventos

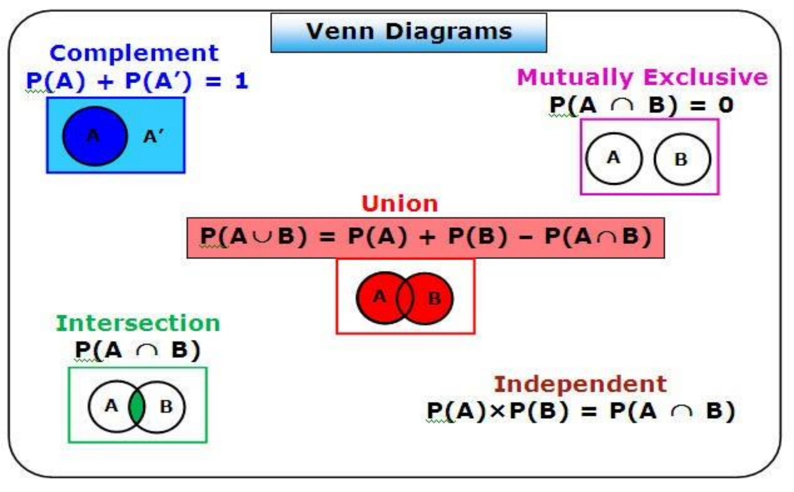


Figure: http://maths.nayland.school.nz

Eventos complementarios

Dos eventos son complementarios si la suma de la probabilidades $P(A)$ y $P(B)$ de un evento A y un evento B es igual a 1.

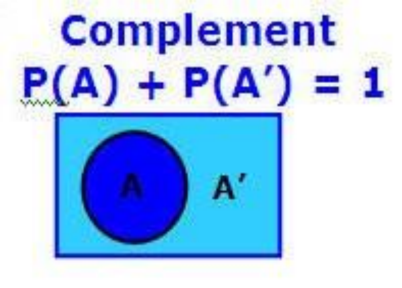


Figure: http://maths.nayland.school.nz

Intersección de eventos

La intersección de dos eventos $P(A \cap B)$ es el subconjunto de los resultados en común.

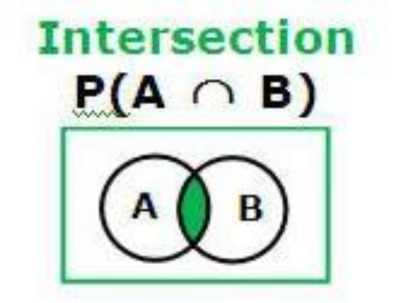


Figure: http://maths.nayland.school.nz

Eventos mutuamente excluyentes

Dos eventos son mutuamente excluyentes si no tienen resultados en común en el espacio muestral, i.e su intersección es igual a 0.

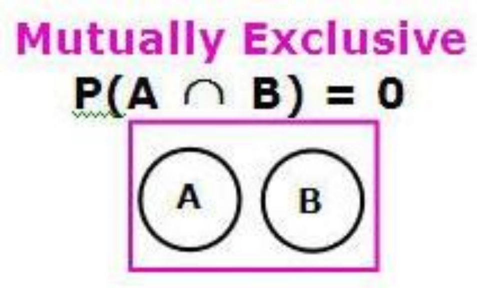


Figure: http://maths.nayland.school.nz

Unión de eventos

La probabilidad de ocurrencia de los eventos A o B es igual a la suma de las probabilidades individuales de los dos procesos menos la probabilidad de intersección de los mismos

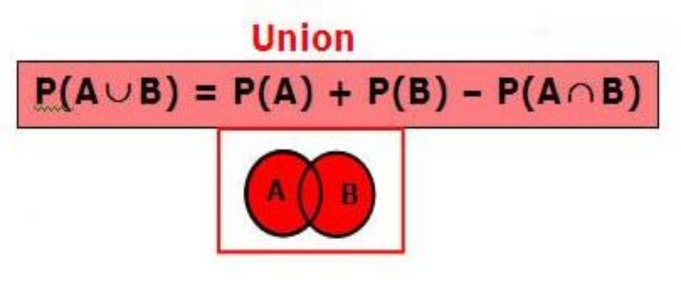


Figure: http://maths.nayland.school.nz

Probabilidad condicional

La probabilidad condicional, es la probabilidad de que ocurra un evento B dado que un evento A ya ha ocurrido.

Se calcula:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \text{donde: } P(A) > 0$$

Probabilidad condicional

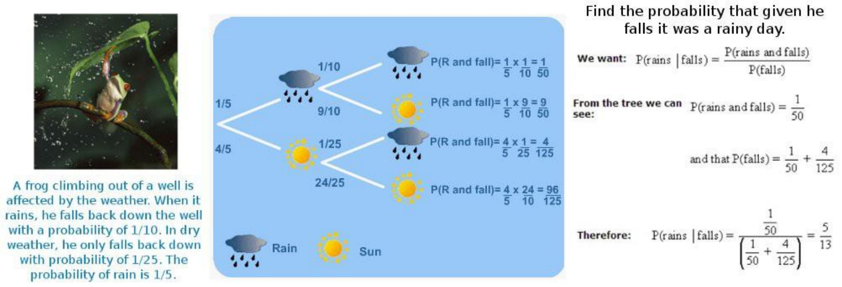


Figure: Tomado de: http://www.s-cool.co.uk

Probabilidad condicional

Find the probability that given he falls it was a rainy day.

We want: $P(\text{rains} \mid \text{falls}) = \dfrac{P(\text{rains and falls})}{P(\text{falls})}$

From the tree we can see: $P(\text{rains and falls}) = \dfrac{1}{50}$

and that $P(\text{falls}) = \dfrac{1}{50} + \dfrac{4}{125}$

Therefore: $P(\text{rains} \mid \text{falls}) = \dfrac{\frac{1}{50}}{\left(\frac{1}{50} + \frac{4}{125}\right)} = \dfrac{5}{13}$

Figure: Tomado de: http://www.s-cool.co.uk

Probabilidad condicional

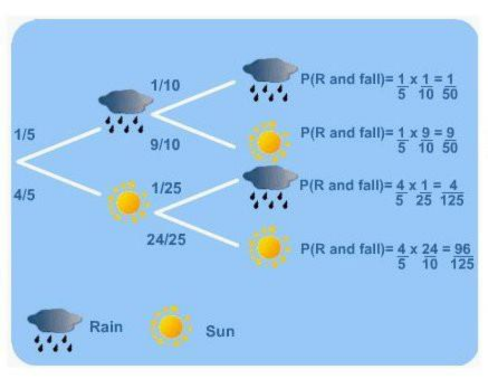


Figure: Tomado de: http://www.s-cool.co.uk

Eventos Independientes

Dos eventos A y B se consideran independientes si **la ocurrencia de uno NO afecta la del otro**. Para este caso se cumple:

$$P(A \cap B) = P(A) \cdot P(B)$$

$$P(B|A) = P(B)$$

$$P(A|B) = P(A)$$

Reglas

Si los eventos $A, B, C, \cdots, Z$ son mutuamente excluyentes:

- $P(A \cup B \cup C \ldots \cup Z) = P(A) + P(B) + P(C) + \ldots + P(Z)$

Si los eventos $A, B, C, \cdots, Z$ son independientes:

- $P(A \cap B \cap C \ldots \cup Z) = P(A) \cdot P(B) \cdot P(C) \cdot \ldots \cdot P(Z)$

Si $A_1, A_2, A_n, \ldots, A_n$ es una partición de un espacio muestral:

- $P(A_1) + P(A_2) + P(A_n) + \ldots + P(A_n) = P(\Omega) = 1$

Reglas

Teorema de probabilidad total:

Si los eventos $B_1, B_2, ..., B_k$ constituyen una partición de un espacio muestral, tal que $P(B_i) \neq 0$ para $i = 1, 2, ..., k$ entonces para cualquier evento A del espacio muestral se tiene:

$$P(A) = \sum_{i=1}^{k} P(B_i \cup A) = \sum_{i=1}^{k} P(B_i) P(A|B_i)$$

Reglas

Teorema de Bayes:

Si los eventos Si $B_1, B_2, \ldots, B_k$ constituyen una partición de un espacio muestral, tal que $P(B_i) \neq 0$ para $i = 1, 2, \ldots, k$ entonces para cualquier evento A en el espacio muestral tal que $P(A) \neq 0$ se tiene:

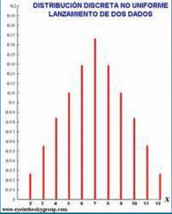
$$P(B_r|A) = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}$$

Variables aleatorias

Discreta

Toman valores enteros (Asociadas a experimentos en el que se mide el número de veces que sucede algo)

Función de probabilidad

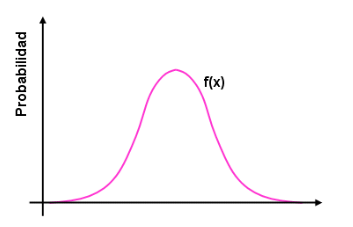


Variables aleatorias

Continua

Pueden tomar todos los valores de un intervalo (El resultado de medir)

Función de densidad
de probabilidad



Distribución de probabilidad

- Una distribución de probabilidad muestra los posibles valores de una variable aleatoria y sus probabilidades asociadas (que valores toma y con que frecuencia)
- Puede estar dada por una tabla, una gráfica o una expresión matemática (fórmula)

Adaptado de: http://colposfesz.galeon.com

Distribución de probabilidad

- En el caso discreto puede ser descrita para cada resultado (Función de masa de probabilidad)
- En el caso continuo solo para intervalos finitos (Función de probabilidad de densidad)

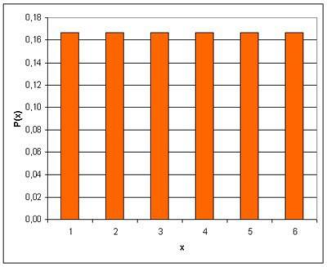
Adaptado de: http://colposfesz.galeon.com

Uniforme Discreta

Un número finito de resultados con igual probabilidad.

$$p(x = X) =$$

$$f(x) = \begin{cases} 1/k\,, & x = x_1, x_2, \dots x_k \\ 0, & en\ otro\ caso \end{cases}$$



Ej: La probabilidad de que el resultado del lanzamiento de un dado sea 6

Bernoulli

Un único experimento con dos posibles resultados: éxito o fracaso.

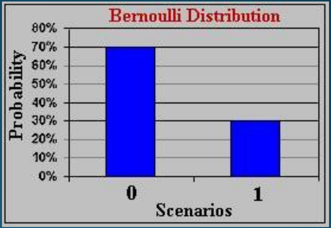
$$p(x) = \begin{cases} \rho^x \rho(1-\rho)^{1-x}, & x = 0, 1 \\ 0, & en\ otro\ caso \end{cases}$$

donde:

p es la probabilidad de éxito en una sola prueba.

x es el número de éxitos en la prueba.

El parámetro es p, $0 \leq p \leq 1$.



Ej: La probabilidad de obtener cara al lanzar una moneda

Binomial

La probabilidad del número de éxitos en *n* ensayos Bernoulli independientes con una probabilidad fija de ocurrencia del éxito entre los ensayos.

$$p(x = X) =$$

$$f(x) = \begin{cases} \binom{\eta}{x} \rho^{x} (1 - \rho)^{x}, & x = 0,1,2 \ldots \eta \\ 0, & en\ otro\ caso \end{cases}$$

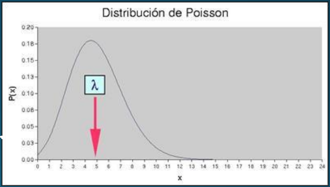
Ej: La probabilidad de obtener 5 aciertos en un examen de diez preguntas verdadero / falso que se contestan al azar

Poisson

Determinar la probabilidad de que ocurran x éxitos por unidad de tiempo, área o producto.

$$p(x = X) =$$

$$f(x) = \begin{cases} \frac{e^{\lambda}\lambda^{x}}{x!}, & x = 0,1,2\,... \\ 0, & en\ otro\ punto\ o\ valor \end{cases}$$



Ej: En promedio hay 50 incendios por localidad anualmente. Cual es la probabilidad de que no haya ningún incendio manãna?

Función de distribución acumulada

Describe la probabilidad de que el valor de una variable aleatoria sea menor o menor o igual a un valor dado:

$$F(x) = P(X \leq x)$$

Función de distribución acumulada - Discreta

En el caso discreto, corresponde a la suma de las probabilidades de los valores menores o iguales a x:

$$F(x) = \sum_{x_i} f(x_i)$$

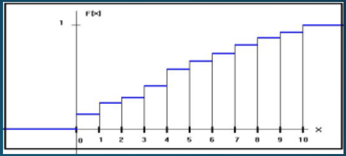


Figure: Función de distribución de v.a discreta

Función de distribución acumulada - Continua

Para las variables aleatorias continuas, las funciones de probabilidad cumplen las siguientes propiedades:

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$P[a \leq X \leq b] = \int_{a}^{b} f(x)dx$$

Función de distribución acumulada - Continua

Las distribuciones teóricas más conocidas para variables aleatorias continuas son:

- Normal
- Chi-cuadrado (X^2)
- Gamma
- t
- F

Distribución Normal

Una variable aleatoria X tiene distribución normal si y solo si su función de densidad es:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{1}{2\sigma^2}(x-\theta)^2\right]}$$

donde $\infty < x < \infty$, $\sigma > 0$ y $-\infty < \theta < \infty$

Distribución Normal

$\theta = E(X)$ y $\sigma^2 = Var(X)$ corresponden a la media y a la varianza de la variable aleatoria. **Si** $\theta = 0$ **y** $\sigma^2 = 1$**, se dice que X tiene una "distribución normal estándar"**

Importancia de la Distribución Normal

- La mayoría de los datos se comportan de acuerdo a ésta (Una concentración de datos hacia el promedio)
- Solamente usa dos parámetros: la media y la varianza
- Muchas de las pruebas y métodos estadísticos asumen que los datos siguen esta distribución
- Si no la tienen se pueden transformar para hacer inferencia sobre estos

Importancia de la Distribución Normal

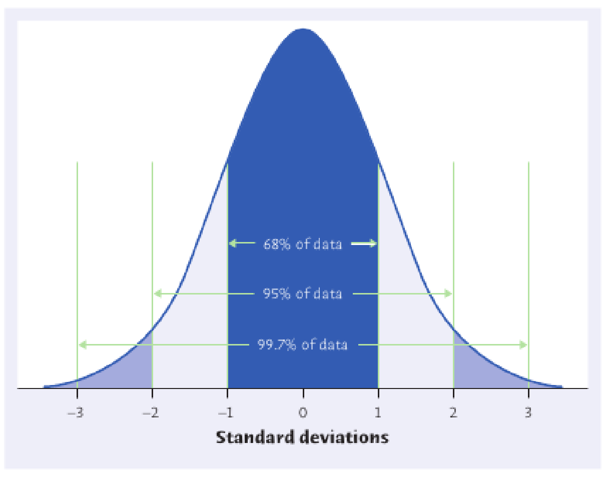


Figure: Moore et al. The basic practice of statistics, 2013

Otras distribuciones continuas

Distribución	Descripción
Uniforme	Corresponde al experimento ideal de elegir un número al azar en el intervalo (a, b)
Exponencial	Se utiliza para representar el tiempo que transcurre hasta la primera ocurrencia del proceso de Poisson
t-student	Se utiliza para estimar la media de una distribución normal cuando el tamaño de la muestra es pequeño
Ji-cuadrado	Sigue la suma de los cuadrados de n variables independientes e idénticamente distribuidas según una distribución normal

Función de distribución acumulada

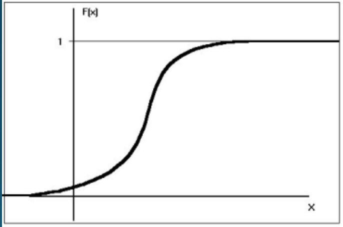
Describe la probabilidad de que el valor de una variable aleatoria sea menor o menor o igual a un valor dado:

$$F(x) = P(X \leq x)$$

Función de distribución acumulada

Para una variable continua, da el resultado del área bajo la curva de la función de densidad:

$$F(x) = \int_{-\infty}^{x} f(x)dx$$



(a) Función de distribución de v.a continua

Teorema del limite central

Si tenemos un **grupo de variables independientes** y todas siguen el mismo modelo de distribución la suma de ellas se distribuye según **la distribución normal**

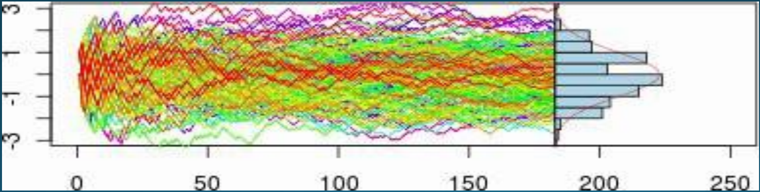


Figure: http://www.datanalytics.com/2011/01/26/la-ley-de-los-grandes-numeros-y-el-teorema-central-del-limite-en-dos-animaciones/

Teorema del limite central

- La media muestral es similar a la muestra poblacional
- De acuerdo con este teorema, la desviación estándar de la distribución de las medias es igual a la desviación estándar de la población dividida por la raíz cuadrada del tamaño de la muestra.

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

- A esta medida se le llama el error estándar de la media
- Este teorema nos permite hacer estimaciones sobre muestras de tamaño grande sin conocer su distribución.
- Es cierto sea cual sea la distribución de la población siempre y cuando su desviación estandar sea finita