Resumen

Métodos cuantitativos para la decisión

Autor

León Pita Pedro

Este documento contiene apuntes teóricos, ejercicios y ejemplos además de un resumen de conceptos estadísticos como apoyo para el examen final del 13 dic 2017. El exámen consta de problemas de la primera parte de simulación/teoría de colas y preguntas test de ambas partes. **Las respuestas de los test no están verificadas!** Cualquier errata, comentario o añadido, enviar el PDF comentado a pleonpita@gmail.com, gracias.

Última actualización 12 de diciembre de 2017

Índice

Parte I

Simulation

1. Discrete Event Simulation Modelling

1.1. Definitions

- **System**: a collection of **entities** (people, messages, machines, servers, ...) that act and interact together toward some end.
- **State of a system**: collection of variables and their values necessary to describe the system at a given time.
- **Entity**: objects that compose a system.
- **Types of systems**
  - Discrete: state variables change at exact points of time.
  - Continuous: state variables change continuously.
- **Simulation models**: static vs. **dynamic**, deterministic vs. **stochastic**, continuous vs. **discrete**.

1.2. Example of a Simulation Model

Vistos los enunciados de parciales de años pasados, este ejemplo parece poco importante para el parcial.

$$IA(IntervalArrival) : 0,4; 1,2; 0,5; 1,7; 0,2; 1,6$$

$$ST(ServiceTime) : 2; 0,7; 0,2; 1,1; 3,7; 0,6$$

#	t_i llega	D_i espera	S_i servicio	l_i atendido	C_i se va
1	0,4	0	2	0,4	2,4
2	1,6	0,8	0,7	2,4	3,1
3	2,1	1	0,2	3,1	3,3
4	3,8	0	1,1	3,8	4,9
5	4	0,9	3,7	4,9	8,6
6	5,6	3	0,6	8,6	9,2

- t_2 = tiempo de llegada del anterior + IA de este

$$0,4 + 1,2 = 1,6$$

- D_2 = tiempo de salida del anterior - tiempo de llegada de este

$$2,4 - 1,6 = 0,8$$

- $t_2 = ServiceTime = 0,7$
- l_2 = max(tiempo de salida del anterior, tiempo de llegada de este)

$$1,6 + 0,8 = 2,4$$

- $C_2 = l_2 + S_2 = 2,4 + 0,7 = 3,1$

Reloj de simulación	# llega	status	cola
0			
0,4	#1	llega	0
1,6	#2	llega	1
2,1	#3	llega	2
2,4	#1	se va	1
3,1	#2	se va	0
3,3	#3	se va	0
3,8	#4	llega	0
4	#5	llega	1
4,9	...		

En esta tabla ordenamos los sucesos en su orden real en el tiempo. En 1,6 como #1 aún no se ha ido la cola aumenta uno. En 2,4 como #1 se va la cola disminuye uno. En 3,8 como no hay nadie cuando llega #4 la cola no aumenta.

1.3. Considerations on Simulation Techniques

- **Advantages**: flexible in modelling complex systems, long time frame, easy to compare alternatives, etc.
- **Drawbacks**: stochastic simulations produce only estimates, expensive to develop, large volume output, inadequate education and training, etc.

Taking a biased sample (*muestra sesgada*) from the population, when analysing statistically the results of a simulation, is an unforeseen drawback, common in statistical studies. This is called **inadequate education and training** and does not verify the condition of independency between variables. In 3.3 we will see the effect of warm-up period in this aspect.

1.4. Starting wiht Arena

Here, only the pseudocode is explained and is structured in two blocks: Definitions and Simulation code. It is strongly recommended to see examples of other years mid-terms as the problem is very likely to be about this (for the mid-term).

Definitions:

- **ENTITY** (entityName) % Individual entities are characterized by Attributes
  - attributeName1
  - attributeName2
  - ...
- **ATTRIBUTE** (attributeName)
  - number of rows
  - number of columns
  - initial value
- **VARIABLE** (variableName)
  - number of rows
  - number of columns
  - initial value

- **RESOURCE** (resourceName)
  - Fixed capacity or based on Schedule
  - Capacity value or ScheduleName
- **SCHEDULE** (scheduleName)
  - Capacity type or Arrival type

Simulation Code

- **CREATE** (entityName, InterArrDist, entPerArr, maxArrival, firstCreation)
  - *entityName*: Entity that is created by the module
  - *InterArrDist*: Interrarival time distribution with its parameters (expo(10))
  - *entPerArr*: Entities per arrival (a number or a distribution)
  - *maxArrival*: maximum number of arrivals
  - *firstCreation*: distribution for the first arrival time
- **PROCESS** (action, resource, quantity, delayDist)
  - *action*: Logic of the process (SDR,SD,DR,D)
  - *resource*: Resource type seized by the entity
  - *quantity*: Number of seized resources by the entity
  - *delayDist*: Delay time distribution with parameters
- **ASSIGN**
  - List of assignments for attributes or variables
- **READWRITE** (type, fileName)
  - *type*: Read or Write
  - *fileName*: File to read or write
  - List of assignments: inputs or outputs to read or write
- **DECIDE** (type) %description
  - *type*: 2-way or n-way by chance; 2-way or n-way by condition
  - List of Percentages or Conditions and destination labels

  % after a decide each destination label has its code specified #1 code,#2... until DISPOSE

- **RECORD**
  - *type*: Count, Expression, Entity Statistics, Time Interval, Time between
  - *value*: value or the expression to increment the counter
  - *attribute*: attribute to compute the Time Interval
  - *counterName*: output to record the value
- **BATCH**
  - *size*: number of entities to be batched
  - *rule*: any entity or by attribute
  - *save*: first, last, sum or product of attributes for the entity batched
- **SEPARATE**
  - *type*: duplicate or split batch
  - *number*: Number of duplicates
- **DISPOSE** %description

2. Repaso Estadística

¿Cómo enfocar un problema de inferencia?

Hipótesis - Muestreo - Contraste

2.1. Distribución de un estimador

Para la asignatura de métodos cuantitativos para la decisión se estudian dos estimadores: el de la media y el de la varianza ya que de una población se desconoce su media (μ) y su varianza(σ^2), por eso se busca estimarlas.

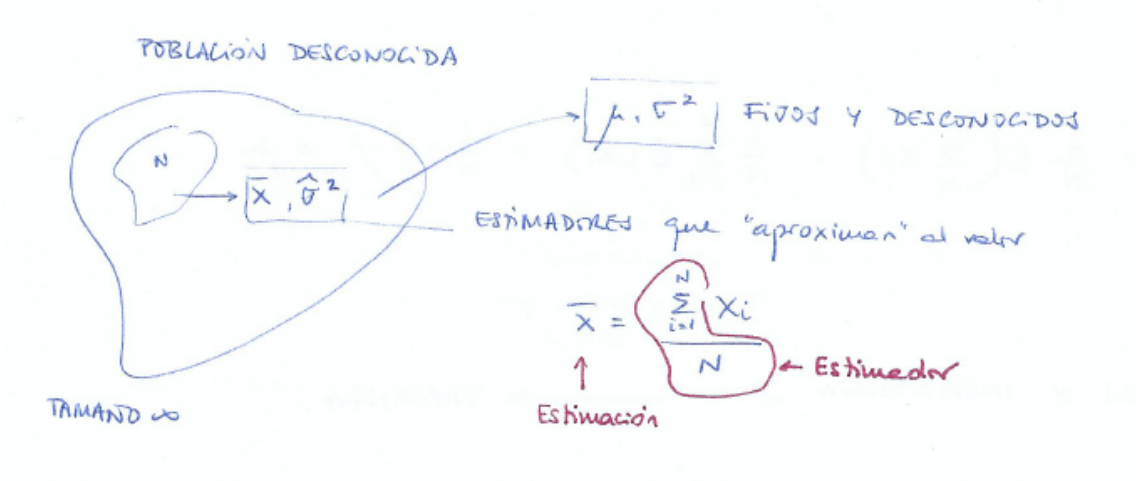


Figura 1: Diferencia población-muestra

Sin embargo, de una muestra finita se pueden calcular su media y su varianza. A estas se les denomina: media muestral ($\overline{X}$) y varianza muestral ($\hat{\sigma}^2$) y sus fórmulas son:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N}$$

- **Hipótesis 1**: $X_i \sim N(\mu, \sigma^2)$ La población de elementos sigue una distribución normal de media μ y varianza σ^2.

  Sesgo de un estimador: se dice que el estimador de la media es insesgado

si al hacer el promedio de todos los posibles estimadores se obtiene μ.

$$E(\overline{X}) = \mu$$

$$E\left(\frac{\sum_{i=1}^{N} X_i}{N}\right) = \frac{1}{N}.E\left(\sum_{i=1}^{N} X_i\right) = \frac{1}{N}.\sum_{i=1}^{N} E(X_i) = \frac{1}{N}.N.\mu = \mu$$

Sólo bajo independencia de muestras aleatorias y estimador insesgado

- **Hipótesis 2**: el muestreo es aleatorio (los elementos son independientes entre sí).

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$

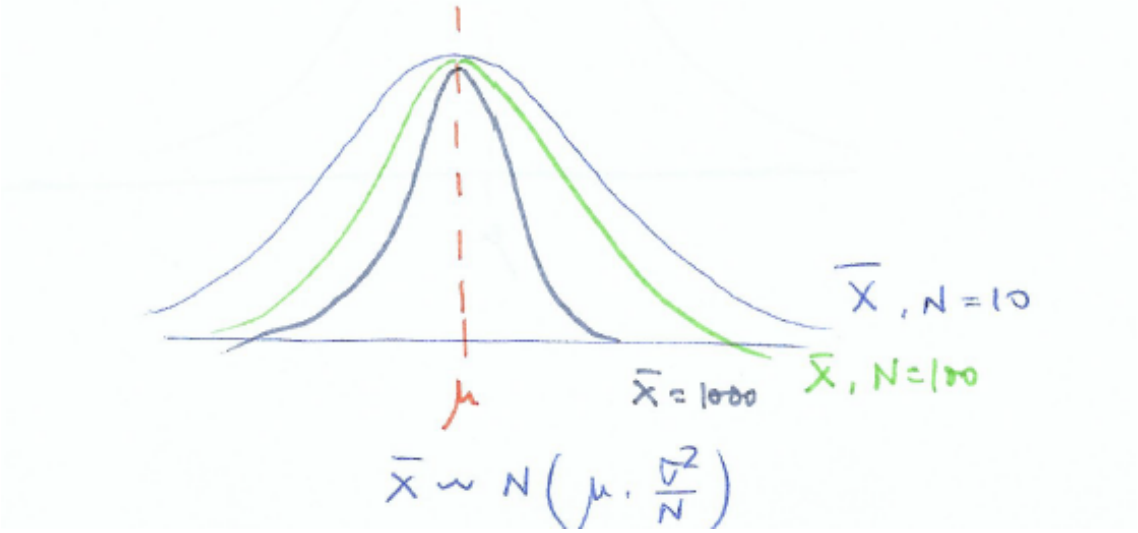


Figura 2: Distribución normal en función de n

Ejemplo: ¿Cómo cambiará el comportamiento del consumidor ante una variación del precio?

$\overline{X} = -0,83$ si aumentara el precio en 1 % ¿Bajaría un 0,83 % de la demanda?

$s = 0,5$, desviación típica ¿Cómo cambiará el valor de s?

$N = 30$

Hipótesis 2: $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$ tipificamos $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{N}}} \sim N(0,1)$

Como la varianza es desconocida, se estima con la varianza poblacional, trabajando con:

$$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{N}}} \sim t_{n-1}; n < 30$$

$$\frac{\overline{X}-\mu}{\frac{s}{\sqrt{N}}} \sim N(0,1); n > 30$$

Cálculo del intervalo de confianza (IC). Un intervalo de confianza está caracterizado por su **nivel de significación** α o su **nivel de confianza** $1-\alpha$.

$$IC_{1-\alpha}(\mu) = \left[\overline{X} \pm N(0,1).\frac{s}{\sqrt{N}}\right]$$

Rango posible del parámetro poblacional (n>30). Cuantifica la incertidumbre y ayuda a valorar la toma de decisiones ante diferentes hipótesis.

Figura 3: Zonas de rechazo y NO rechazo

- **Error de tipo I**: Probabilidad de rechazar cuando NO debería.
- **Error de tipo II**: Probabilidad de NO rechazar cuando debería.

Volviendo al ejercicio, el test de hipótesis:

$$H_0 : \mu = -0,5$$

$$H_1 : \mu \neq -0,5$$

Estadístico de contraste:

$$T : \frac{\overline{X}-\mu}{\frac{s}{\sqrt{N}}} \sim N(0,1)$$

El IC a $1-\alpha = 95\,\%$ para una distribucuón N(0,1) es $[-1,96; 1,96]$ (simétrico). Por tanto, para un valor absoluto de T elevado, se rechaza la hipótesis nula.

$$T : \frac{-0,83-(-0,5)}{\frac{0,5}{\sqrt{30}}} = -3,61 \notin [-1,96; 1,96]$$

Rechazo H_0

Cambiando la hipótesis inicial por una nueva:

$$H_0 : \mu = -0,8$$

$$H_1 : \mu \neq -0,8$$

El nuevo estadístico de contraste es:

$$T : \frac{-0,83 - (-0,8)}{\frac{0,5}{\sqrt{30}}} = -0,32 \in [-1,96; 1,96]$$

NO rechazo H_0

Cálculo del p_valor (o *valor de significación frontera*):

$$p\_valor = 2 * P(N(0,1) < |-0,32|)$$

$$p\_valor = 2 * (1 - Prob(0,32))$$

$$p\_valor = 2 * (1 - 0,6255)$$

$$p\_valor = 0,755 = 75\,\%$$

Por tanto, para un nivel de significación del 70 %, o lo que es lo mismo, un nivel de confianza del 30 % (demasiado bajo), se obtiene IC: $[-0{,}38, 0{,}38]$ y -0.32 no está contenido en la zona de Rechazo. Sin embargo, para niveles de confianza aceptables (90 % o 95 %) siempre rechazaríamos la hipótesis.

Rechazo H_0

3. Analizing Simulation Output

3.1. Types of Statistical Variables

- **Observation-based data**: sequence of equally weighted data values that do not persist over time (ex. average waiting time for a queue).
- **Time-based data**: is associated with the duration or interval of time that an object is in a particular state (ex. average number of people).

It is important to differentiate statistical variables **within replication**, for example, the average of arrivals of the replication number 16, and statistical variables **across replication**, for example, the total average of arrivals of all replications. It is said *an average of averages.*

3.2. Finite-Horizon Simulations

A well defined ending time o condition can be specified. They are focused on **transient** period. In these kind of problem, precision depends on the size of the sample, n. As the populational variance is unknown, the sample variance is used, s.

$$h = t_{\alpha/2,n-1}.\frac{s}{\sqrt{n}} \leq E$$

For $n \geq 30$ thanks to TCL, the statistic is no longer distributed as a t-Student of n-1 freedom degrees but as a N(0,1).

$$h = q_{\alpha/2}.\frac{s}{\sqrt{n}} \leq E$$

For a given half-width h_0 for a sample size n_0, if we want know the sample size n to achieve a better half-width h we apply:

$$n \approx n_0.\left(\frac{h_0}{h}\right)^2$$

3.3. Infinite-Horizon Simulations

These kinds of simulations are focused on the **steady-state** period. Here the transient is named **warm-up** period and it needs to be deleted using the **method of Replication-Deletion**.

- Make R replications.
- Y_{ir} is the *ith* observation within replication r.
- Compute the average across replications.
- Plot $\overline{Y}_{.i}$ for each i.
- Apply smoothing techniques to this plot.
- **Visually assess** where the plots start to converge.
- The run length of the simulation should be **at least 10 times the warm-up period**.

Basically this method consists in deleting the warm-up period *a ojo*. Some consequences are:

- Sample variance increases as observations are deleted.
- **Bias decrease** as the effect of warm-up period is deleted.
- Throwing away data increases the computational time on collecting usable data.
- Assessing the warm-up period may requires significant data storage.

Specifying a warm-up period in Arena causes to schedule a warm-up event for time T_w. At that time, **all the accumulated statistical counters are cleared** so that the net effect is that statistics are only collected over the period from T_w to T_e.

3.4. Comparing system configurations

In this problem there are **2** populations, so μ_1 and μ_2. The null hypothesis (*hipótesis nula*) is $\mu_1 = \mu_2$. This is the same as creating a new statistic $\hat{D} = \overline{X}_1 - \overline{X}_2$, so the null hypothesis $H_0 : \hat{D} = 0$ and $H_1 : \hat{D} \neq 0$.
In order to compare both configurations, their variances are supposed to be equal to an average variance named s_p^2.

$$T = \frac{\overline{X}_1 - \overline{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} - \frac{1}{n_2}}} = \frac{\hat{D}}{s_p \cdot \sqrt{\frac{1}{n_1} - \frac{1}{n_2}}} \sim t^{\alpha/2}{}_v$$

$$v = n_| + n_2 - 2$$

$$s_p = \frac{(n_1 - 1).s_1 + (n_2 - 1).s2}{n_1 + n_2 - 2}$$

Once the value of the statistic T is obtained, its **p_value** is calculated by looking into the t-Student table with $n_1 + n_2 - 2$ freedom degrees (*grados de libertad*). The p_value is the α that would give the value of T calculated, it is a border value. For $\alpha > p\_value$ H_0 must be rejected, and for $\alpha < p\_value$ accepted.

In the case that both configurations have different variances $s_1 \neq s_2$ then the statisitic resulting is:

$$T = \frac{\hat{D}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim {t^{\alpha/2}}_v$$

$$v = \left[\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2+1}} - 2 \right]$$

4. Queueing Theory

4.1. Procesos de Poisson - Exponencial

Una distribución de Poisson es una distribución de probabilidad discreta que modeliza [número de llegadas al sistema]. Las llegadas, por hipótesis, son **independientes** unas de otras.

Parámetros de entrada:

- Promedio por unidad de tiempo $\lambda.t$
- Variable número de llegadas $N(t)$
- Valor que toma N, k

$$P\left([N(t+h) - N(t)] = k\right) = \frac{e^{-\lambda.t}.(\lambda.t)^k}{k!}$$

Ejemplo: los clientes llegan a una tienda a razón de 10 por hora, cada entrada tiene probabilidad de 1/2 de ser hombre o mujer. Asumen que en la primera hora han llegado 10 mujeres. Calcula la probabilidad de que entren 10 hombres en la siguiente hora.

Como son sucesos independientes, no importa que hayan llegado 10 mujeres en la primera hora.

$$\lambda^{hombres} = 10.\frac{1}{2} = 5$$

$$P(N(1) = 10) = \frac{e^{-5*1}.(5*1)^{10}}{10!}$$

¿Y que entren más de 10?

$$P(N(1) > 10) = \sum_{i=11}^{inf} \frac{e^{-\lambda}.(\lambda)^i}{i!} = 1 - \sum_{i=0}^{10} \frac{e^{-\lambda}.\lambda^i}{i!}$$

Intervalos de tiempo infinitesimales $h \to 0$

$$P(N(h) = 0) = e^{-\lambda.h} \simeq 1 - \lambda.h$$

$$P(N(h) = 1) = e^{-\lambda.h}.\lambda.h \simeq (1 - \lambda.h).\lambda.h \simeq \lambda.h$$

$$P(N(h) > 1) = 1 - \lambda.h - (1 - \lambda.h) \simeq 0$$

Para una Poisson la probabilidad de que llegue más de una persona al mismo tiempo es casi nula.

Variable aleatoria, tiempo de espera del sistema: WT

$P(WT > t) = e^{-\lambda.t}$ = probabilidad de que para un tiempo t no lleguen clientes.

$$P(N(t) = 0) = \frac{e^{-\lambda.t}.(\lambda.t)^0}{0!} = e^{-\lambda.t}$$

$P(WT \leq t) = 1 - e^{-\lambda.t}$ = función de distribución de la exponencial.

Ejemplo: los trabajos llegan cada 15 segundos a una media de 4 por minuto. ¿Cuál es la probabilidad de estar esperando un trabajo menos de 30 segundos?

$$P(WT < 0,5) = 1 - e^{-4*0,5}$$

4.2. Elementos de una cola de espera

Notación de Kendall: M / M / 2 / inf

- M, Interarrival time pattern = M:Exponencial, D:Constante,Ek:Erlang, G: General
- M, Service time pattern
- 2, Number of servers
- inf, System capacity

Símbolos:

- λ ratio medio de llegada
- μ ratio medio de servicio
- t tiempo
- c número de servidores del sistema
- k capacidad del sistema
- $\frac{1}{\lambda}$ tiempo medio entre llegadas consecutivas
- $\frac{1}{\mu}$ tiempo medio de servicio
- ρ ratio de uso $\rho = \frac{\lambda}{c \cdot \mu}$

Rendimiento de la línea de espera

- n estado del sistema, número de clientes (en cola o siendo servidos)
- P_n probabilidad de que haya n clientes en el sistema
- L número medio de clientes en el sistema, $L = E[n]$
- n_q número de clientes en cola
- L_q número medio de clientes en cola, $L_q = E[n_q]$
- t tiempo total en el sistema
- W media del tiempo total en el sistema, $W = E[t]$
- t_q tiempo de espera en cola
- W_q media del tiempo de espera en cola, $W_q = E[t_q]$

4.3. Birth-Death process

Conceptos básicos:

- Una variable aleatoria N(t) toma valores n, gente en cola.
- Las distintas VA son **independientes**
- Gracias a la distribución exponencial podemos obviar los casos en los que nacen/mueren 2 o más personas.

Diagrama de transición:

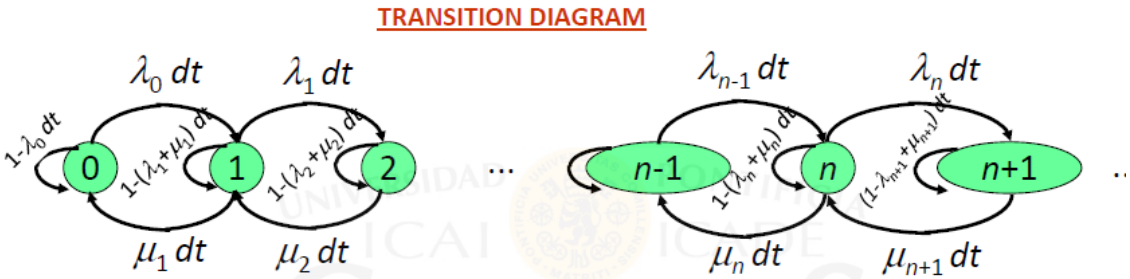
En un diagrama de transición se representan los posibles estados como círculos y dentro de ellos se escribe el valor de la variable de estado (o las variables de estado). Los cambios de estado se rigen por dos factores:

- λ, ratio medio de llegada. Para pasar de n=0 a n=1 se necesita que llegue un cliente (ratio medio de llegada).
- μ, ratio medio de servicio. Para pasar de n=1 a n=0 se necesita que un cliente de la cola pase a servicio (ratio medio de servicio).

Por último, para pasar de n=0 a n=0, quedarse en un estado, se dice que tiene un factor de 1 menos la suma de todos los factores que salen del estado n=0. Lo mismo se puede aplicar para cualquier n.

Matriz de transición:

En la diagonal tenemos la suma de factores de seguir en un estado (flechas de llegada del diagrama). Los elementos ij son los ratios de transición de un estado a otro. Por ejemplo, el elemento de fila=0 y columna=1 es el ratio de pasar de n=0 a n=1, o lo que es lo mismo, flechas con salida del estado 0 y llegada al estado 1 en el diagrama.



TRANSITION MATRIX

State at $t + dt$

State at t	0	1	2	3	4
0	1-λ0 *dt*	λ0 *dt*	0	0	0
1	μ1 *dt*	1-(λ1+μ1) *dt*	λ1 *dt*	0	0
2	0	μ2 *dt*	1-(λ2+μ2) *dt*	λ2 *dt*	0
3	0	0	μ3 *dt*	1-(λ3+μ3) *dt*	λ3 *dt*
4	0	0	0	μ4 *dt*	1-(λ4+μ4) *dt*

Figura 4: Diagrama y matriz de transición

Las variables de estado se establecen de modo que todos los estados queden perfectamente definidos.

- **Ejemplo 1**: un cajero automático. El número de servidores, c, es constante e igual a 1. En cada instante sólo se necesita saber el número de clientes que esperan.
- **Ejemplo 2**: en este caso el banco a partir de 4 clientes esperando (n=4), abre un segundo cajero (c=2). Siguiendo con la solución anterior, para el estado en el que haya 3 clientes (n=3) hay dos estados posibles:
  - Que sólo haya un cajero abierto disponible y la cola aún no haya llegado a tener 4 clientes.
  - El segundo caso, que la cola haya llegado a tener 4 clientes, el segundo cajero haya abierto y, por tanto, se esté atendiendo a dos clientes en ese

instante.

Entonces se dice que la variable de estado es compuesta: el número de clientes en espera y el número de cajeros abiertos.

Cálculo general del estado estacionario: Para calcular el estado estacionario se busca que las probabilidades de encontrarse en los n estado posibles sean invariables:

$$\frac{d}{dt}p_0 = 0; \frac{d}{dt}p_1 = 0; ...; \frac{d}{dt}p_n = 0$$

Partiendo del primer estado n=0:

$$\frac{d}{dt}p_0 = 0 \rightarrow aplicandoLaDefinicionDeDerivada \rightarrow \frac{p_0(t+dt) - p_0(t)}{dt} = 0$$

La probabilidad de estar en el estado n=0 un instante después es igual a la suma de probabilidades de llegadas al estado 0 (factor x probabilidad de encontrarse en el estado de salida en el momento t).

$$p_0(t+dt) = (1 - \lambda_0.dt) \cdot p_0(t) + \mu_1 \cdot dt \cdot p_1(t)$$

$$\frac{[(1 - \lambda_0.dt) \cdot p_0(t) + \mu_1 \cdot dt \cdot p_1(t)] - p_0(t)}{dt} =$$

$$\frac{p_0(t) - \lambda_0 \cdot dt \cdot p_0(t) + \mu_1 \cdot dt \cdot p_1(t) - p_0(t)}{dt} =$$

$$\frac{-\lambda_0 \cdot dt \cdot p_0(t) + \mu_1 \cdot dt \cdot p_1(t)}{dt} = -\lambda_0 \cdot p_0(t) + \mu_1 \cdot p_1(t) = 0;$$

$$\lambda_0 \cdot p_0(t) = \mu_1 \cdot p_1(t)$$

Para el siguiente caso, n=1, el procedimiento es igual, sólo cambia la probabilidad de seguir en el estado 1 en t+dt:

$$p_1(t+dt) = [1 - (\mu_1 + \lambda_1).dt].p_1(t) + \lambda_0 \cdot dt \cdot p_0(t) + \mu_2 \cdot dt \cdot p_2(t)$$

Sustituyendo en esta expresión el valor de $p_0(t)$ en función de $p_1(t)$ obtenido como resultado del primer estado y calculando la derivada de $p_1(t)$ igual a 0 se obtiene:

$$\lambda_1 \cdot p_1(t) = \mu_2 \cdot p_2(t)$$

Realizando este cálculo para todos los estados el resultado es el mismo, por tanto, se concluye:

$$p_n = \frac{\lambda_0 \cdot \lambda_1 ... \lambda_{n-1}}{\mu_1 \cdot \mu_2 ... \mu_n} \cdot p_0 = C_n \cdot p_0$$

Si a esta condición se le añade que el sumatorio de las probabilidades de todos los estados debe ser 1 se obtiene:

$$\sum_{i=0}^{inf} p_i = 1; p_0 + C_1 \cdot p_0 + ... + C_n \cdot p_n = 1$$

$$p_0 = [1 + C_1 + ...C_n]^{-1}$$

Volviendo al contexto de qué estamos hablando, λ es la tasa media de llegada y μ la tasa media de ocupación. Los valores λ_i y μ_i suelen ser constantes para todo el sistema. Es decir, el ratio medio de que llegue alguien será igual si hay 0, 2 o 30 clientes. Así también con el ratio medio de servicio, poco importa que haya 3, 4 o 30 personas, el cajero va a tardar lo mismo de media. Por tanto, se establece el uso medio del servidor[1]

$$\rho = \frac{\lambda}{c \cdot \mu}$$

En los casos que c=1[2] para el estado estacionario se obtiene: $P_n = \rho^n \cdot P_0$ Para no confundir entre el uso medio del servidor (ρ) y la probabilidades de cada estado, se escribirá p mayúscula.

Medidas de rendimiento:

Estas son las mismas que las detalladas en la subsección 4.2 Elementos de una cola de espera, aquí se especifican sus fórmulas generales. Para cada sistema el cálculo de estas magnitudes es diferente, pudiendo llegar a ser muy complejo. En las slides del tema[3] se pueden encontrar sus valores para los sistemas:

- M/M/1
- M/M/1/k
- M/M/c
- M/M/c/k

Para cada uno diferenciando el caso de infinitos clientes y el caso de tener número finito de clientes.

- Número medio de clientes en el sistema: $L = E[n] = \sum_{n=0}^{inf} n \cdot P_n$

[1] Subsección 4.2 Elementos de una cola de espera, Símbolos

[2] Númer de servidores del sistema

[3] s_QueueingTheory.pdf

- Número medio de clientes en espera: $L_q = E[n_q] = \sum_{n=c+1}^{inf} (n-c) \cdot P_n$
- Ratio medio de llegada: $\overline{\lambda} = \sum_{n=0}^{inf} \lambda_n \cdot P_n$
- Tiempo total medio en el sistema: $W = E[t] = \frac{L}{\overline{\lambda}}$
- Tiempo de espera medio: $W_q = E[t_q] = \frac{L_q}{\overline{\lambda}}$

Parte II

Data Analysis

5. Multivariate Data Analysis

5.1. One-way model

Problema: ¿Son las medias de mi población distintas?

Ejemplos: ¿Depende la demanda eléctrica de la hora del día? ¿Depende la nota académica del grupo de clase?

Variable dependiente: demanda eléctrica y nota académica.

Factor: hora del día y grupo de clase.

$$Y(X = x_i) = \mu_i + E_i = \mu + \alpha_i + E_i$$

- μ_i: media del grupo i
- E_i: error aleatorio
- μ: media global
- α_i: desviación para cada grupo i

Hipótesis

1. Residuos $E_{ij} \sim N(0, const)$ **Homocedasticidad**: $Var(E_i) = constante$
2. Observaciones independientes

Test de hipótesis

$H_0 : \mu_1 = ... = \mu_m$

$H_1 : algun\ \mu_i \neq \mu_j$

$$F = \frac{VarianzaExplicada}{VarianzaNoExplicada} \sim F_{m-1,N-m}$$

$$VarianzaExplciada = \frac{\sum_{i=1}^{m}\sum_{j=1}^{N_i}(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2}{m-1}$$

$$VarianzaNoExplicada = \frac{\sum_{i=1}^{m}\sum_{j=1}^{N_i}(Y_{ij} - \overline{Y}_{i\bullet})^2}{N-m}$$

- **m**: númer de grupos o valores que puede tomar el factor (grupo de clase A, B y C).
- **N**: tamañode los grupos. Todos los grupos tienen el mismo tamaño.
- $\overline{Y}_{i\bullet}$: media del grupo i. $\overline{Y}_{\bullet\bullet}$: media de la muestra.
- Y_{ij}: valor de la variable dependiente de una observación j en el grupo i.

5.2. Two-way model

$$Y(X = x_i) = \mu + \alpha_i + \beta_i + \alpha\beta_{ij} + E_{ij}$$

- μ: media global
- α_i: desviación para cada grupo i
- β_i: desviación para cada grupo j
- $\alpha\beta_{ij}$: desviación interacción ij
- E_i: residuos

Test de hipótesis

$H_0 : \mu_i$ *iguales* , μ_j *iguales*

$H_1 :$ *algun* μ_i *distinto* , *algun* μ_j *distinto*

Número de réplicas: observaciones con valores iguales de ambos factores i, j. El número de réplicas debe ser igual para todos los valores. En el ejemplo de las palomitas, necesitamos tener el mismo nÃºmero de observaciones de ventas para una marca (Popitas) y un tipo de palomitas (saladas). Si tenemos 20 datos de ventas de Popitas saladas también habrá 20 datos de ventas de Popitas mantequilla, y lo mismo para el resto de marcas.

Para este modelo encontramos 3 estadísticos y para cada uno obtendremos un p_valor.

- Relevancia del factor 1
- Relevancia del factor 2
- Relevancia del efecto cruzado

Se puede dar que un factor no sea relevante pero sí el efecto cruzado. Esto significa que el otro factor afecta la relevancia del primero.

Para p-valores pequeños **rechazo** la hipótesis nula de que los μ_i son iguales. Entonces, el factor es **relevante**.

Para p-valores grandes, **acepto** la hipótesis nula de que los μ_i son iguales. Entonces, el factor **no es relevante**.

En un test de hipótesis hablamos de nivel de confianza $1-\alpha$ y nivel de significación α. Además, el p-valor también se le llama nivel de significación umbral[4]. Para niveles de significación mayores que que el p-valor el estadístico F estará en la región de rechazo y para niveles de significación menores que el p-valor el estadśtico F estará en la región de aceptación.

5.3. Componentes Principales

Problema: reducir la dimensión de variables explicativas.

Para reducir el número de variables explicativas se va a rotar la matriz de diseño X en una matriz Z. La matriz de rotación A es la matriz de autovectores de la matriz de covarianzas S. A la matriz $Z = X \cdot A$ obtenida se le llama matriz de componentes principales.

- **Matriz de diseño**: X de n filas (observaciones) y p columnas (variables).
- **Matriz de covarianzas**: $S = Cov(X_i, X_j) = \frac{\sum_{k=1}^{n}(X_{i,k}-\overline{X}_i)(X_{j,k}-\overline{X}_j)}{n-1}$
- **Autovalores** λ: $det(S - \lambda \cdot I) = 0$ varianza explicada por cada componente principal.
- **Autovectores** a_i: $(S - \lambda_i \cdot I) \cdot a_i$ componentes principales.
- **Matriz de rotación**: A = $[a_o \ a_1 \ ... \ a_p]$ en orden decreciente de varianza explicada. Matriz ortogonal $A^t A = I$
- **Matriz de componentes principales**: n filas (observaciones) y p columnas (componentes principales).

[4] 2

6. Clustering and Classification

6.1. Distancias

- **Entre observaciones**: $D(x_u, x_v) = \sqrt[L]{\sum_{j=1}^{p} \left| d_{x_j}(x_{ju}, x_{jv}) \right|^L}$ L=2 euclídea, L=1 Manhattan.
- **Entre variables**: $D(x_u, x_v)_{LS} = \frac{Cov(x_u, x_v)}{\sqrt{Var(x_u) \cdot Var(x_v)}}$ donde LS = Learning Set.

6.2. Dendogramas

Método gráfico. Empezando con las n variables desagrupadas unimos las variables con distancia mínima entre ellas. Una vez tenemos un conjunto hecho, se toma la distancia mínima entre variables teniendo en cuenta el nuevo grupo de variables. Este método se llama Single Link.

6.3. K-means

Toma de decisión entre minimizar el error de quantización y minimizar el número de clusters.

$$QE = \sum_{i=1}^{K} QE_i = \sum_{i=1}^{K} [\sum_{e \in LS; e \in Cluster_i} \|x_e - c_i\|^2]$$

- QE: error de cuantización
- K: número de clusters
- x_e: observación e
- c_i: centro del cluster i $c_i = argmin \|x_i - m_k\|^2$
- $\|\|$: norma aplicada para medir la distancia.

6.4. Árboles de clasificación

Función de entropía

$$H(LS(n)) = -\sum_{i=1}^{N_c} (p(n, c_i) \cdot log[p(n, c_i)])$$

Donde $p(n, c_i)$ es la proporción de puntos pertenecientes a la clase i y N_c el número total de clases.

- **Criterio de separación**: se escoge el criterio que separe el Learning Set en un conjunto de clases N_c que minimice la entropía H(LS(n)).
- **Criterio para dejar de separar**: se harán tantas separaciones como sean necesarias hasta alcanzar un nivel de entropía mínimo $H(LS(n)) < H^{min}$.

7. Regression

7.1. Regresión Lineal

Problema: Estimar la recta, plano o hiperplano que minimice la proyección del punto en la dirección del eje.

$$y_i = \alpha + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + ... + \beta_p \cdot x_{ip} + \epsilon_i$$

- y_i: variable explicada
- x_{ij}: j variables explicativas
- α: intercept
- ϵ_i: residuos

Hipótesis

1. x_{ij} exógenas: constantes en muestras repetidas.
2. $\epsilon_i \sim N(0, \sigma_i)$ el valor de x no afecta al error.
3. Estimación de $\alpha, \beta_1, ..., \beta_p$ tal que mín $\hat{\epsilon_i}^t \cdot \hat{\epsilon_i}$

Test de significatividad individual

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0$

$$T = \frac{\hat{\beta_j} - \beta_{H_0}}{SE(\hat{\beta_j}} = \frac{Estimacion}{ErrorEstandar}$$

Hablamos de error estándar como la desviación típica de un parámetro estimado. Para obtener $\hat{\alpha}$ y $\hat{\beta_j}$ se resuelve el sistema lineal obtenido de derivar la suma total

de los residuos respecto a α y β_j.[5]

$$\hat{\beta}_j = \frac{Cov(x_j, y)}{Var(x_j)}$$

$$SE(\hat{\beta}_j) = \sqrt{\frac{Var(\epsilon_i)}{n \cdot Var(x_j)}}$$

$$T = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim t_{n-2}$$

Para valores de T grandes (o muy negativos, porque la T de Student tiene dos colas) **rechazamos** la hipótesis nula $\beta_j = 0$ y entonces la variables es **relevante** en el modelo. Sin embargo, si sustituimos los valores de $\hat{\beta}_j$ y $SE(\hat{\beta}_j)$ en el estadístico se pueden observar otros factores más allá de la relevancia de la variable que afectan al tamaño de T.

$$T = \frac{\sqrt{n} \cdot Cov(x_j, y)}{\sqrt{Var(\epsilon_i) \cdot Var(x_j)}}$$

- Cuanto mayor sea la muestra ($n >>$) menor será el error estándar y por tanto, $T >>$.
- Si la variable x_j y la variable y están muy relacionadas $Cov(x_j, y) >>$ entonces $T >>$.
- Si la varianza del residuo es demasiado grande, el test nos dará valores del estadístico T pequeños entonces aceptamos H_0 por tanto, x_j no es relevante. Pero puede que para valores menores de varianza del error rechacemos H_0.
- (**importante**) Si la variable explicativa x_j no tiene una varianza considerable ($Var(x_j) <<$) entonces obtenemos valores de T grandes y por tanto, rechazamos H_0. Cuidado con variables de pequeñas varianzas!

Coeficiente de determinación: no tiene unidades de medida y toma valores entre 0 (muy disperso) y 1 (cerca de la línea).

$$R^2 = \frac{S^2_{REG}}{S^2_Y} = 1 - \frac{S^2_{RES}}{S^2_Y}$$

Test de significación global

$H_0 : \beta_1 = ... = \beta_p = 0$

[5] Ver la slide 18 de sRegression.pdf

$H_1 : algun\ \beta_j \neq 0$

$$F = \frac{S^2_{REG}}{S^2_{RES}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \sim F_{p,n-p-1}$$

Para valores de F grandes, p-valor pequeño, entonces rechazamos H_0 y por tanto, alguna variables explicativa es relevante.

7.2. Modelos aditivos, regresión no-lineal

La diferencia entre los modelos lineales y no-lineales es que para los modelos no-lineales, los parámetros β_j se pueden encontrar dentro de las funciones base $B_j(x, \beta_j)$. En el modelo lineal se puede dar que la dependencia con las variables explicativas x_j no sea lineal pero la dependencia con los parámetros β_j siempre debe ser lineal.

$$Y = f(x_1 ... x_m) + E$$

$$f(x) = \beta_0 + \sum_{j=1}^{M} \beta_j \cdot B_j(x)$$

Sin embargo, para el modelo no lineal tenemos:

$$f(x) = g(\beta_j, x)$$

Funciones Base:

- **Polinomios**: $B(x) = x^j / j = 0, ..., m$
- **Polinomios de Hermite**: $P_j(x) = 2 \cdot x \cdot P_{j-1}(x) - 2 \cdot (j-1) \cdot P_{j-2}(x) / j = 2, ..., m$ Esta es una base de funciones ortogonales.
- **Funciones sigmoides**
- **Funciones radiales o gausianas**

7.3. Redes Neuronales y Perceptron Multicapa

Usan modelos aditivos, no lineales, debido a la función de activación. Diferenciamos entre dos familias de redes neuronales. Aquellas que usan funciones base radiales **RBFN** y aquellas que usan sigmoides como funciones base **MLP**.

$$Sigmoide : f(\nu) = \frac{1}{1 + exp(-\nu)}$$

Las funciones sigmoides empiezan valiendo 0 y tras una rampa acaban valiendo 1.

Mean Squared Error:

$$MSE = \frac{1}{N}\sum_{e=1}^{N}\sum_{i=1}^{q}[y_{ie} - MLP_i(x_e)]^2$$

Al igual que para la regresión encontramos el factor de determinación R^2, para las redes neuronales encontramos el MSE.

Cálculo de parámetros:

Al igual que en la regresión derivamos los residuos totales respecto a los parámetros e igualamos a 0, en redes neuronales aplicamos el mismo procedimiento, derivando MSE parcialmente para obtener los parámetros β_{ij} que ahora llamamos pesos y los parámetros α_{ij} de las funciones base. Este método se llama el **Método del gradiente**[6].

[6] Ver la slide 77 d sRegression.pdf

Parte III

Test Data Analysis

8. Test Análisis Multivariacional

Las respuestas no están verificadas!

1. c) **Nothing can be said without an analysis of variance.** a y d son falsas porque no se ha realizado ningún análisis estadístico. b es falsa porque n=20 es una muestra suficiente.

2. c) **Differences among the grades of the groups are not statistically significant with a confidence level of 98.4 %.** a es falsa, el nivel de confianza es $1 - p\_valor$. b es falsa porque no menciona el nivel de confianza. d es falsa, no habla de diferencias estadísticas.

3. a) **The hypothesis of equality among the means of the three groups is accepted with a confidence level of 98.4 %.** b falsa, nivel de significación en lugar de confianza. c falsa, sería verdadera en caso de decir que al menos un grupo presenta diferencias estadísticas para $\alpha > 1{,}6$. d falsa, mismo error que en b.

4. c) **Grades for groups A and B are significantly different and B and C are significantly similar.** El test de hipótesis tiene como nula $H_0 : \mu_1 = ... = \mu_n$ por tanto si aceptamos, son significativamente parecidas. a y d son falsas porque A-B no son significativamente similares. b falsa, B-C son significativamente similares y no "no significativamente distintos".

5. a) **Greater than the residual variance, approximately twice.**

$$F = \frac{VarianzaExplicada}{VarianzaResiduos} = 2{,}26 \sim 2$$

6. *...incentive mechanism that recognizes relevant differences among the sale performance of the agencies...* Sea una variable Y_i que mida el rendimiento de una agencia, aplicamos un ANOVA a esta variable.

7. a) 1 componente principal

   b) $Z = X \cdot A$ y la matriz $A = \begin{pmatrix} \epsilon & \sim 1 \\ \sim 1 & \epsilon \end{pmatrix}$ para $\begin{pmatrix} x_1 y_1 \\ x_2 y_2 \\ ..... \end{pmatrix}$

   c) Los autovalores son la varianza explicada por cada componente. (99,1)

   d) Primera opción (horizontal). El eje x sería la primera componente principal que es la que explica la mayor parte de la varianza.

8. a) Las componentes principales son otrogonales, por tanto, no pueden estar relacionadas.

   b) Explican principalmente grupos complementarios. 1: física, gráfica, química. 2: cálculo, álgebra, computación.

   c) 3 componentes para explicar más del 90 % de la varianza.

   d)
   $$79{,}65 = \frac{2{,}1 + 1{,}46}{\sum_i latent_i}$$

   e) Sí, misma respuesta que en a.

9. No entra análisis factorial.

9. Test Clustering and Classification

Las respuestas no están verificadas!

1.

2. a) **The distance metric (Euclidean, Manhattan, correlation coefficient, etc.) used.** QE depende directamente de la norma/distancia escogida ($\|\|$).
   $$QE = \sum_{e \in LS} \|x_e - c\|^2$$

3. a) **Can impact the number of clusters obtained.**

4. Análisis discriminante no entra.

5. b) **There is no clear trade-off between number of clusters and quantization error, but 7 seems to be a good approximation.** La toma de

decisión del número de clusters busca un punto medio entre minimizar QE y el número de clusters. c falsa, no confundir K-Means y PCA.

6. 1. Displacement, Weight, HorsePower por un lado y Acceleration/MillesPerGallon por otro.
   2. 4 clusters
   3. b) Falsa, son los más ecológicos +desplazamiento, -comsumo
   4. Análisis discriminante no entra.
   5. Análisis discriminante no entra.
   6. Tiene error 0, fijádonos en el gráfico con la división de clusters es fácil comprobar la división. Un árbol más simple sería con una única división de Displ<21,5 teniendo muy poco error.
   7. KSOM no entra.

10. Test Regression

Las respuestas no están verificadas!

1. b) **Depend on the observations available.** a falsa proque puede ser positivo o negativo. c falsa, sistema de ecuaciones lineal.
2. b) **Is uncertain and this uncertainty depends on the standard error.** a es falsa, la pendiente β se estima con $T = \frac{\hat{\beta}}{SE(\hat{\beta})}$. c falsa, la incertidumbre disminuye con n.
3. b) **Is adimensional, i.e., is expressed in per unit from 0 to 1.** a falsa, $R^2 = \frac{\hat{S}^2_{REG}}{\hat{S}^2_Y} = 1 - \frac{\hat{S}^2_{RES}}{\hat{S}^2_Y} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$ c falsa.
4. c) **At least one variable is explanatory.** El test de hipótesis tiene como nula $H_0 : \hat{\beta}_1 = ... = \hat{\beta}_p$
5. b) **Is detected by inconsistency between tests on the coefficients of the variables and test on the overall fit.**
6. a) **The variable must be eliminated from the regression model.** $p\_value >>$, aceptamos la nula $H_0 : \hat{\beta}_j = 0$, variable no relevante.

7. c) **Will have a normal distribution if the regression model fits the observations very well.** $\epsilon_i \sim N(0, \sigma_i^2)$ es una hipótesis del problema que se cumplirá en caso de que el modelo se ajuste a las observaciones.
8. b) **Can have basis functions with adjusting parameters in it.** Falsa, porque las funciones base pueden ser no lineales, y el modelo debe ser lineal respecto a los parámetros a estimar.
9. b) **Only uses polynomials a basis functions.** Falsa, las funciones base pueden ser también radiales, de Hermite o sigmoidales.
10. c) **Are the basis functions used in a multiple linear regression model.** Falsa. a verdadera, son un tipo de funciones base. b verdadera, su media y su varianza (slide 38).