Tema 1. Teoria del Mostreig Estadístic (Part I)

Matemàtiques II. Grau de Química

Francisco Guerrero

Dept. Matemàtica Aplicada, Universitat de València

Curs 2012-2013

Continguts

Continguts

Continguts de l'assignatura

- Estadística $\rightarrow$ Inferència estadística: Processos de decisió sobre una població, basats en l'anàlisi estadístic de dades provinents d'una o diverses mostres.
- Càlcul numèric
  - Aproximació de funcions per interpolació
  - Aproximació numèrica de processos d'integració: Integrals definides i equacions diferencials ordinàries.

Continguts

Estadística

Definició (Wikipedia)

Ciència, amb base matemàtica que tracta sobre la recol· lecció, l'anàlisi i la interpretació de dades, i que busca explicar i/o deduir condicions regulars en fenòmens de tipus aleatori.

- Estadística descriptiva: Tècniques per a presentar i resumir un conjunt de dades, per a facilitar la comprensió d'aquestes i/o per a generar hipòtesis.
- Inferència estadística: Tècniques que permeten deduir, a partir de les dades d'una mostra, alguna propietat de la població de la qual s'extraurà la mostra.
- Teoria de la probabilitat: Part de les matemàtiques que va nàixer per a explicar els jocs d'atzar i que constitueix la base de la inferència estadística.
  - Intervals de confiança, test d'hipòtesis (respostes a preguntes).
  - Estimació de paràmetres.
  - Correlació entre variables (regressió).
  - Anàlisi de variància (ANOVA).

Conceptes bàsics

- Població: Conjunt d'*individus* que es pretén estudiar.
- Mostra: Subconjunt de la població sobre el qual es disposen *dades*.
- Variable: Una **característica** de la població que pren un valor per a cada individu de la població. Les variables poden ser
  1. Categòriques (sexe, estat civil,. . .)
  2. Numèriques (temperatura, pes,. . .)
- Dades: Valors que pren una variable.

De vegades s'empren els termes *població* i *mostra* per a referir-se a les dades associades a un *individu*.

Conceptes bàsics

Experiment determinista

És aquell que repetit en igualtat de condicions dóna el mateix resultat.

Exemple

Medir la intensitat del corrent d'un circuit simple, amb una resistència i una pila: $I = E/R$.

Si el resultat del experiment està condicionat per una llei física, es tracta d'un experiment determinista.

Experiment aleatori

Aquells per als quals no es pot predir el resultat.

Exemple

Llançar una moneda a l'aire.

En un experiment aleatori, coneixem per endavant tots els possibles resultats però no podem predir el resultat concret de l'experiment.

Conceptes bàsics

Resum

Experiments deterministes: Les mateixes causes, en les mateixes condicions, produeixen els mateixos efectes.

Experiments aleatoris: Les mateixes causes, en les mateixes condicions, poden produir efectes diferents.

Continguts

Variables aleatòries

Exemple

- Llançar una moneda (o dos, o tres,. . .).
- Llançar un dau (o dos, o tres,. . .).
- Producció de peces mecàniques, defectuoses o no defectuoses.
- Avaries en una màquina en un interval de temps, 1,2,3.
- Mesura del diàmetre interior d'un rodament en una cadena de producció.

El resultat d'un experiment és

- Impredictible *a priori*, peró
- pertany a un conjunt que es pot descriure completament abans de dur a terme l'experiment.

Variable aleatòria

Variable que quantifica els resultats d'un experiment aleatori.

Espai mostral

Conjunt de possibles resultats (o successos) de l'experiment aleatori.

Espai mostral

L'espai mostral associat a un experiment aleatori és el conjunt de tots els possibles resultats de l'experiment.

- El denotarem per Ω.
- L'espai mostral és discret si els possibles resultats de l'experiment són una quantitat finita o numerable[1].

Exemple

- Tirar dues monedes a l'aire: $\Omega = \{CC, CX, XC, XX\}$ (discret).
- Número de partícules emeses per una font radioactiva $\Omega = \{0, 1, 2, 3, \ldots\} = \aleph$ (discret).

- L'espai mostral és continu si els possibles resultats formen un conjunt infinit i no numerable.

Exemple

- Mesura del diàmetre interior d'un rodament en una peça mecànica: $\Omega = [15.5, 18.2]$ (continu).

[1] numerable = si els seus elements es poden comptar

Successos aleatoris

Els elements d'Ω s'anomenen punts mostrals:

Succés

Un succés és un subconjunt de l'espai mostral Ω.

- Succés elemental Succés que conté un sol punt d'Ω.
- Succés compost Succés que conté més d'un punt d'Ω.
- Succés impossible Succés que no succeeix mai $A = \emptyset$.
- Succés segur Succés que sempre succeeix $A = \Omega$.

Es diu que ha ocorregut el succés A si el resultat de l'experiment és un punt mostral $\omega \in A$

Conclusió

TREBALLEM AMB CONJUNTS!

Tema 1. Teoria del Mostreig Estadístic

Francisco Guerrero

Introducció
Continguts de l'assignatura

Estadística
Definició i conceptes bàsics
Experiments aleatoris
Estadística descriptiva
Mesures de resum

Probabilitat
Introducció
Definició empírica
Definició de Laplace
Definició axiomàtica
Probabilitat condicionada
Teorema de factorització i probabilitat total

Relació entre conjunts

- $A \subset \Omega, \qquad A^c = \{\omega \in \Omega : \omega \notin A\}$
- $A, B \subset \Omega, \qquad A \cup B = \{\omega \in \Omega : \omega \in A \text{ o } \omega \in B\}$
- $A, B \subset \Omega, \qquad A \cap B = \{\omega \in \Omega : \omega \in A \text{ y } \omega \in B\}$
- $A, B \subset \Omega, \qquad A \subset B \text{ si } \forall \omega \in A \rightarrow \omega \in B$
- $A, B \subset \Omega, \qquad A = B \text{ si } A \subset B \text{ y } B \subset A$

Successos incompatibles

A i B són successos incompatibles si $A \cap B = \emptyset$

Lleis de Morgan

- $(A \cap B)^c = A^c \cup B^c$
- $(A \cup B)^c = A^c \cap B^c$

Continguts

Definició i conceptes

Definició

Conjunt de tècniques per a representar i resumir un conjunt de dades de naturalesa aleatòria, amb l'objectiu de facilitar la comprensió d'aquestes i/o generar hipòtesis.

El pas previ per a la visualització de les dades són les taules de freqüències.

Freqüència absoluta d'un succés

Nombre de vegades que s'observa el succés.

Freqüència relativa d'un succés

Percentatge d'ocurrències pel que fa al total d'observacions.

Exemple

- Variable aleatòria $\rightarrow$ Pes d'un individu.
- Població $\rightarrow$ Xiquets entre 12 i 14 anys.
- Mostra $\rightarrow$ 30 xiquets triats a l'atzar en un col·legi de Burjassot.

Taules de freqüències

Exemple

S'obtenen els valors *en brut* següents:

42	47	54	53	50	51	51	43	80	62	62	65	46	47	42
52	54	53	62	60	51	65	46	47	54	46	47	51	63	63

Podem agrupar les dades calculant les freqüències absolutes de cada valor observat:

pes	42	43	46	47	50	51	52	53	54	60	62	63	65	80
frec	2	1	3	4	1	4	1	2	3	1	3	2	2	1

O bé agrupant-les per *classes*:

pes	freq. abs.	freq. rel.
40-45	3	3/30=0.1
46-50	8	8/30=0.227
51-55	10	10/30=0.333
56-60	1	1/30=0.0333
61-65	7	7/30=0.233
>66	1	1/30=0.0333

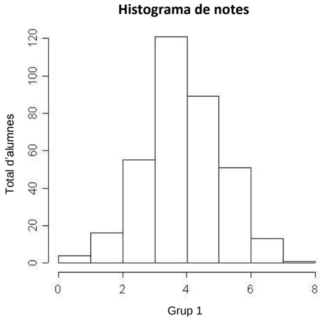
Visualització de dades

Histograma

Representació gràfica d'una taula de distribució de freqüències per classes.

Què es veu en l'histograma?

- Mínim-màxim
- Interval o intervals més freqüents
- Simetria o asimetria



Continguts

Mesures

- Mesures de posició o de tendència central
  - Mitja (solament per a variables quantitatives)
  - Mitjana
  - Moda
- Quartils i percentils
- Mesures de dispersió
  - Rang
  - Variància i desviació típica
  - Rang interquartil

La mitja

- Si les dades en brut recollides per a la variable X són $x_1, x_2, \ldots, x_n$, la mitja es defineix com:

$$\overline{X} := \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Si les dades estan recollides en una taula de freqüències absolutes:

x_1	x_2	...	x_k
n_1	n_2	...	n_k

$$\overline{X} := \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{n_1 + n_2 + \ldots + n_k} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i}$$

- Si les dades estan recollides en una taula de freqüències relatives:

x_1	x_2	...	x_k
f_1	f_2	...	f_k

$$\overline{X} := x_1 f_1 + x_2 f_2 + \ldots + x_k f_k = \sum_{i=1}^{k} x_i f_i$$

La mitjana i la moda

Mitjana

És el valor que està *enmig de les dades*, en el sentit que hi ha tantes dades menors que la mitjana com dades majors que ella.

Comparació entre mitja i mitjana

- Similars per a distribucions simètriques.
- Poden ser molt diferents per a distribucions asimètriques.
- La mitjana és molt menys sensible a valors atípics (outliers).

Moda

És el valor més repetit

Exercici

Calculeu la mitja, la mitjana i la moda de l'exemple anterior emprant les diferents fòrmules:

- Mitja = 56.633
- Mitjana = 51.5
- Moda =

Quartils i percentils

Una vegada ordenades, la mitjana deixa aproximadament la meitat de les dades *a l'esquerra* i la meitat, *a la dreta*.

- El Primer Quartil, Q1 és el valor que deixa *a l'esquerra* aproximadament la quarta part (1/4) dels valors ordenats, és a dir, la quarta part de les dades són menors que aquest valor i les 3/4 parts són majors.
- El Segon Quartil, Q2 és la mitjana: deixa *a l'esquerra* la meitat de les dades i l'altra meitat, a la dreta.
- El Tercer Quartil, Q3...
- El Percentil 10 (P_{10}) és el valor que deixa a *l'esquerra* aproximadament el 10% de les dades ordenades.

Exercici

- Què és el P_{90}?
- Calculeu els quartils de l'exemple anterior.

Mesures de dispersió

1. Rang: és la diferència entre el valor màxim observat i el mínim.
2. Variància:
$$s^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$
3. Desviació típica:
$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
4. Rang interquartil: Q3-Q1

Exercici (Calculeu aquestes quantitats per a les dades de l'exemple)

- Rang =
- s^2 =
- s =
- Q3 - Q1 =

Recomanacions

- Per a descriure variables la distribució de les quals s'assembla a la corba normal de Gauss, és usual utilitzar la mitjana i la desviació típica.
- Per a variables amb distribucions asimètriques és freqüent utilitzar la mitjana i els quartils. Açò és particularment cert en presència de valors atípics que desvirtuen la informació proporcionada per la mitja i la desviació típica.

Continguts

Introducció

Freqüència absoluta d'un succés

Nombre de vegades que ocorre el succés en dur a terme un nombre determinat de repeticions d'un experiment.

N = nombre de repeticions de l'experiment.
n_A = nombre de vegades que ocorre el succés A (freqüència absoluta d'A).

Freqüència relativa d'un succés

Percentatge d'ocurrències del succés.

$$fr(A) = \frac{n_A}{N}$$

Propietats

- $0 \leq fr(A) \leq 1$
- $fr(\Omega) = 1$
- $fr(\emptyset) = 0$
- $fr(A \cup B) = fr(A) + fr(B) \qquad \text{si } A \cap B = \emptyset$

Continguts

Definició

Llei de regularitat estadística (fet empíric)

La freqüència relativa d'un succés s'estabilitza quan el nombre de repeticions de l'experiment creix indefinidament.

Exemple

A = nombre de vegades que ix cara en llançar una moneda.

$$fr(A) \overset{N\to\infty}{\longrightarrow} \frac{1}{2}$$

Donat un experiment aleatori, es defineix la probabilitat d'un succés com el límit de les freqüències relatives d'aquest succés quan el nombre de repeticions creix indefinidament.

$$p(A) = \lim_{N\to\infty} fr(A) = \lim_{N\to\infty} \frac{n_A}{N}$$

La definició empírica NO ÉS OPERATIVA en general!!

Exemple

Exemple

S'han dut a terme 1.000 llançaments d'una moneda.

N = nre. de llançaments	10	100	250	500	750	1000
n_A = nre. de cares	4	46	124	244	379	501
$fr(A)$	0.4	0.46	0.496	0.488	0.5053	0.501

Pot observar-se que com més gran és el nombre de llançaments, més s'aproxima la freqüència relativa al valor $\frac{1}{2}$, de manera que podríem pensar que la probabilitat de cara és igual que la probabilitat de creu i ambdues, iguals a $\frac{1}{2}$.

Açò només és una suposició, o una aproximació, ja que per a aplicar estrictament la definició empírica hauríem de continuar fins al infinit, la qual cosa resulta impossible.

Continguts

Definició

Fòrmula de Laplace

Si l'espai mostral està format per N resultats possibles i tots ells tenen la mateixa probabilitat (equiprobables), podríem dir que la probabilitat d'un esdeveniment A; $P[A]$; és

$$P[A] = \frac{N_A}{N} = \frac{\text{nombre de casos favorables}}{\text{nombre de casos possibles}}$$

on N_A es el nombre de resultats favorables de l'ocurrència de A.

Daus, monedes,. . . OK si no estan trucades!!

Exemple

Un joc consisteix a llançar un dau i observar el resultat. Es guanya si ix parell i es perd si ix senar.

$$\Omega = \{1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6\};$$

$$A = \text{ guanyar } = \text{ parell } = \{2 \quad 4 \quad 6\}$$

$$p(A) = \frac{\text{nombre de casos favorables}}{\text{nombre de casos possibles}} = \frac{3}{6}$$

Exemple

Exemple

Es llancen a l'aire dues monedes ben construïdes. De les afirmacions següents quina, si n'hi ha cap, és la solució correcta a la pregunta:

Quina és la probabilitat que apareguen dues cares?

- Ja que o *apareixen dues cares* o *no apareixen dues cares*, la probabilitat és $1/2$.
- El nombre de cares obtingut pot ser 0, 1 o 2. La probabilitat de dues cares és $1/3$.
- Encara que siguen monedes iguals, considerarem que podem etiquetar-les com "moneda1" i "moneda2". Tenint en compte aquest ordre, els possibles resultats són CC,CX,XC,XX. La probabilitat de dues cares és $1/4$.

Raoneu la resposta.

Limitacions

L'assignació de probabilitats en l'exemple del dau és bona si el dau està equilibrat, és a dir, el dau està ben construït i totes les cares tenen la mateixa probabilitat d'aparèixer.

Però en els daus trucats açò no és cert. Per exemple, suposem que en la pràctica s'observa que els nombres 2, 4, 5 i 6 solament apareixen un 10% de les vegades, mentre que tant l'1 com el 3 apareixen el 30% de les vegades.

En aquest cas, el dau està òbviament trucat! L'assignació de probabilitats coherent amb l'observació seria:

$$p(1) = 0.3, \quad p(2) = 0.1, \quad p(3) = 0.3, \quad p(4) = p(5) = p(6) = 0.1$$

Això, com es calcularia la probabilitat de $A = \{2, 4, 6\}$?

$$p(A) = ?$$

No podem recòrrer a la definició de Laplace!!

Continguts

Definició

(Segons Kolmogorov)

Si Ω és l'espai mostral d'un experiment aleatori,

es defineix una funció de probabilitat p com una aplicació que assigna a cada succés d'Ω (es a dir, cada subconjunt d'Ω) un nombre real $p(A)$ que compleix:

1. $A \subset \Omega \rightarrow 0 \leq p(A) \leq 1$
2. $p(\Omega) = 1$
3. Si $A_1, A_2, \ldots, A_n, \ldots$ successos mútuament excloents $(A_i \cap A_j = \emptyset) \rightarrow$

$$p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i)$$

Observacions i propietats

- Qualsevol funció que satisfaça les propietats anteriors és una funció de probabilitat.
- La definició de Laplace proporciona una funció de probabilitat que compleix les propietats anteriors.
- Les propietats d'una funció de probabilitat són paral·leles a les de la freqüència relativa!

Propietats

Totes les propietats es dedueixen de les tres propietats bàsiques

- La probabilitat de la unió de successos mútuament excloents és la suma de les probabilitats dels successos.

$$A_1, \ldots, A_n, \quad A_i \cap A_j = \emptyset, \quad i \neq j \quad \Rightarrow \quad P(A_1 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

- La probabilitat que no passe res és 0, $p(\emptyset) = 0$
- $P(A^c) = 1 - p(A)$, on $A^c = \Omega - A$
- $A \subset B \Rightarrow p(A) \leq p(B)$
- $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

Exemples

Exemple

Una empresa du a terme un control del 100% de la seua producció durant un període determinat de temps i observa que de cada lot de 50 peces en solen aparèixer 2 defectuoses, en la majoria dels casos. A = peça defectuosa, B = peça no defectuosa.

$$p(A) = \frac{2}{50} = 0.04, \qquad p(B) = \frac{48}{50} = 0.96 \qquad \Rightarrow \qquad P(A^c) = 1 - 0.04$$

Exemple

Tres cavalls, A, B, C intervenen en una carrera. El cavall A té el doble de probabilitat de guanyar que el cavall B i aquest, el doble que el cavall C. Quina és la probabilitat que té de guanyar cadascun dels tres cavalls?

- $p(A)$ = probabilitat que guanye el cavall A
- $p(B)$ = probabilitat que guanye el cavall B
- $p(C)$ = probabilitat que guanye el cavall C

Succés segur $\Rightarrow A \cup B \cup C$

$$p(B) = 2P(A), \qquad p(C) = 2P(B) \qquad p(A \cup B \cup C) = 1$$

Exercicis

Exercici

Se sap que entre els 120 estudiants d'un col·legi major hi ha 60 que estudien Biològiques (B), 50 que estudien Farmàcia (F) i 20 que cursen ambdós estudis simultàniament. Determineu la probabilitat que un d'ells escollit a l'atzar estudie Biològiques o Farmàcia, i la probabilitat que no estudie ambdues simultàniament.

Exercici

En un cert espai de probabilitat hi ha dos successos A i B les probabilitats dels quals són:

- $P(A) = 0.4$,
- $P(A \cup B) = 0.7$,
- $P(B^c) = 0.55$.

Calculeu $P(A \cap B)$.

Continguts

Probabilitat condicionada

Exemple

- Suposem que en una rifa de 100 números, Joan ha comprat el número 35 i Pere, el 36. Quina probabilitat té Joan d'encertar? I Pere?
- Suposem que sabem que el resultat de la rifa acaba en 5. Quina probabilitat té Joan d'encertar? I Pere?

Definició

Siga Ω l'espai mostral d'un experiment aleatori. Siguen A, B dos successos de Ω amb $p(B) > 0$.

$P(A/B)$ = probabilitat que ocòrrega el succés A si ha ocorregut el succés B.

$$P(A/B) := \frac{p(A \cap B)}{p(B)} \qquad \equiv \qquad p(A \cap B) = P(A/B)p(B)$$

En aquesta definició, P és una funció de probabilitat, és a dir, compleix totes les propietats d'una funció de probabilitat establertes per Kolmogorov.

Successos independents

Definició

A i B són successos independents si $p(A) > 0$, $p(B) > 0$ i

$$P(A/B) = p(A) \quad \Leftrightarrow \quad P(A \cap B) = p(A)p(B) \quad \Leftrightarrow \quad P(B/A) = p(B)$$

La probabilitat condicionada satisfà els requeriments de la definició axiomàtica de probabilitat. En particular:

- $P(A^c/B) = 1 - P(A/B)$
- $P((A_1 \cup A_2)/B) = P(A_1/B) + P(A_2/B) - P((A_1 \cap A_2)/B)$
- ...

Exercicis

Exercici

- Considerem l'experiment aleatori de llançar dues monedes a l'aire. Anomenem A el succés obtenir cara en llançar la primera moneda i B, el succés obtenir cara en llançar la segona moneda. Quina és la probabilitat d'obtenir dues cares després d'efectuar el primer llançament?
- Repetim el procediment anterior amb dues monedes trucades de manera que per a cada moneda $p(C) = 0.8$ i $p(X) = 0.2$. Quina probabilitat tenim ara d'obtenir 2 cares després d'efectuar el llançament?
- Una urna conté 5 boles blanques i 4 de negres. Si s'extrau una bola a l'atzar, quina és la probabilitat que siga blanca?, i que siga negra?
- Si s'extrauen dues boles i es retorna a la urna la bola que s'extrau cada vegada, quina és la probabilitat que les dues siguen blanques?
- Si s'extrauen dues boles sense devolució (sense reemplaçar), quina és la probabilitat que les dues siguen blanques?

Continguts

Teorema de factorització

$$p(A \cap B) = P(A/B)p(B)$$

$$p(A \cap B \cap C) = P(A/(B \cap C))P(B \cap C) = P(A/(B \cap C))P(B/C)p(C)$$

i així en general.

Exercici

- Una urna que conté 5 boles blanques i 4 negres. Si s'extrauen tres boles amb devolució, quina és la probabilitat de traure B1N2N3?
- Si s'extrauen tres boles sense reemplaçar, quina és la probabilitat de traure B1N2N3?

Teorema de la probabilitat total

Definició

Els successos $A_1, A_2, \ldots, A_n$ constitueixen una partició d'Ω si són mútuament excloents, amb una probabilitat no nula i $\Omega = A_1 \cup A_2 \cup \ldots \cup A_n$.

Teorema

Si els successos $A_1, A_2, \ldots, A_n$ formen una partició d'Ω i $B \in \Omega$, aleshores

$$B = \bigcup_{i=1}^{n} (B \cap A_i)$$

i els conjunts $(B \cap A_1), \ldots, (B \cap A_n)$ són mútuament excloents. Així,

$$p(B) = \sum_{i=1}^{n} p(B \cap A_i) \quad \Rightarrow \quad p(B) = \sum_{i=1}^{n} p(B/A_i)p(A_i)$$

Example

Exemple

Suposem que tenim 4 caixes amb components electrònics dins. La caixa 1 conté 2000 components, amb un 5% de defectuoses; la caixa 2 conté 500 components, amb un 40% de defectuoses; les caixes 3 i 4 contenen 1000 components, amb un 10% de defectuoses. Quin és la probabilitat d'escollir a l'atzar un component defectuós?

Si anomemem D: component defectuós i C_i component de la caixa i-èsima. Aleshores, tenim que:

$$P[C_1] = \frac{2000}{2000+500+1000+1000} = \frac{4}{9} \quad P[C_2] = \frac{500}{2000+500+1000+1000} = \frac{1}{9}$$

$$P[C_3] = \frac{1000}{2000+500+1000+1000} = \frac{2}{9} \quad P[C_4] = \frac{1000}{2000+500+1000+1000} = \frac{2}{9}$$

A més, $P(D/C_1) = 0.05$, $P(D/C_2) = 0.4$, $P(D/C_3) = 0.1$ i $P(D/C_4) = 0.1$. Si emprem el teorema de la probabilitat total:

$$P(D) = P(D/C_1)P(C_1) + P(D/C_2)P(C_2) + P(D/C_3)P(C_3) + P(D/C_4)P(C_4) =$$

$$= 0.05\frac{4}{9} + 0.4\frac{1}{9} + 0.1\frac{2}{9} + 0.1\frac{2}{9} = 0.11111$$

Tema 1. Teoria del Mostreig Estadístic (Part I)

Matemàtiques II. Grau de Química

Francisco Guerrero

Dept. Matemàtica Aplicada, Universitat de València

Curs 2012-2013

Tema 1. Teoria del Mostreig Estadístic (Part II)

Matemàtiques II. Grau de Química

Francisco Guerrero Cortina

Dept. Matemàtica Aplicada, Universitat de València

Curs 2012-2013

Continguts

Continguts

Tema 1. Teoria del Mostreig Estadístic

Francisco Guerrero Cortina

Variables aleatòries: El cas discret
Variable aleatòria
Models probabilístics discrets

Variables aleatòries: El cas continu
Variable aleatòria
Models probabilístics continus

La normal i la seua importància
La distribució normal
Sumes de variables aleatòries
Teorema central del límit

Conceptes bàsics

El concepte de variable aleatòria sorgeix de la necessitat de fer mes manejables matemàticament els resultats dels experiments aleatoris, que en molts casos són qualitatius i segueixen patrons molt similars, encara que la naturalesa de l'experiment no ho siga.

Exemple

- Experiment 1: tirar una moneda defectuosa tal que $p(C) = 0.9$.
  $\Omega = \{C, X\} \quad \equiv \quad \{0, 1\}, p(0) = P(C) = 0.9$
- Experiment 2: Observar una peça fabricada en un procés que produeix un 1% de peces defectuoses.
  $\Omega = \{B, D\} \quad \equiv \quad \{0, 1\}, p(B) = p(0) = 0.9$

Aquests dos experiments aleatoris són 'essencialment' el mateix!

El maneig de variables aleatòries permet 'classificar' els experiments aleatoris, de manera que el càlcul de probabilitats és molt més senzill.

Tema 1. Teoria del Mostreig Estadístic

Francisco Guerrero Cortina

Variables aleatòries: El cas discret
Variable aleatòria
Models probabilístics discrets

Variables aleatòries: El cas continu
Variable aleatòria
Models probabilístics continus

La normal i la seua importància
La distribució normal
Sumes de variables aleatòries
Teorema central del límit

Definició

Definició 1

Siga Ω l'espai mostral associat a un experiment aleatori. Una variable aleatòria és una aplicació:

$$X : \Omega \to \mathbb{R}$$

Definició 2

Donada una variable aleatòria X, anomenarem suport de X, S_X o imatge de X, Im_X el conjunt de possibles valors de X en $\mathbb{R}$.

El suport d'una variable aleatòria pot ser discret (finit o numerable) o continu (un interval de $\mathbb{R}$, o tot $\mathbb{R}$).

Variable aleatòria discreta

El seu suport és discret:

$$S_X = \{x_1, x_2, \ldots, x_n, \ldots\}$$

Variable aleatòria contínua

El seu suport és un interval de $\mathbb{R}$.

Llei de probabilitat d'una v. a.

Una variable aleatòria (associada a un determinat experiment aleatori) trasllada la informació probabilística rellevant des d'espai mostral Ω a $\mathbb{R}$, mitjançant una probabilitat induïda que es defineix sobre els conjunts de la recta real $\mathbb{R}$.

Exemple

Considerem l'experiment de llançar dos daus i la v. a.

$$X = \text{ suma de les cares}, \quad S_X = \{2, 3, 4, \dots, 12\} \qquad P_X(0) = 0$$

La taula següent mostra els possibles resultats de l'experiment (tots són equiprobables per a daus equilibrats) i la suma dels valors de les cares:

(1,1) 2	(2,1) 3	(3,1) 4	(4,1) 5	(5,1) 6	(6,1) 7
(1,2) 3	(2,2) 4	(3,2) 5	(4,2) 6	(5,2) 7	(6,2) 8
(1,3) 4	(2,3) 5	(3,3) 6	(4,3) 7	(5,3) 8	(6,3) 9
(1,4) 5	(2,4) 6	(3,4) 7	(4,4) 8	(5,4) 9	(6,4) 10
(1,5) 6	(2,5) 7	(3,5) 8	(4,5) 9	(5,5) 10	(6,5) 11
(1,6) 7	(2,6) 8	(3,6) 9	(4,6) 10	(5,6) 11	(6,6) 12

Observeu que $P_X(4) = \frac{3}{36}$, ja que s'obté quatre com a suma dels possibles resultats $\{\{1,3\},\{2,2\},\{3,1\}\}$.

Funció de quantia

La distribució de probabilitat de X, P_X, ens proporciona la informació necessària per a conèixer el comportament probabilístic de X.

Exemple

Per a la v. a. anterior (la suma de les cares obtingudes en dues tirades d'un dau), la funció P_X sobre les dades del suport, S_X, és la següent:

x_i	2	3	4	5	6	7	8	9	10	11	12
$P_X(X = x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Per a una variable aleatòria discreta, la *llei de probabilitat* o *distribució de probabilitat* de X queda determinada pels valors $P_X(X = x_i)$, $i = 1, 2, \ldots$

Definició

A partir d'aquests valors, podem definir la funció de quantia o funció de densitat de probabilitat o funció de massa de la variable aleatòria com segueix:

$$f_X(x) = \begin{cases} P_X(X = x_i), & x = x_i \\ 0, & x \neq x_i \end{cases}$$

Funció de distribució

En ocasions resulta més convenient recòrrer a la funció de distribució de probabilitat, definida a partir de la probabilitat induïda:

$$F_X : \mathbb{R} \rightarrow [0,1]$$
$$F_X(x) = P(X \leq x)$$

La funció de distribució és aleshores:

$$F_X(x) = \sum_{x_i \leq x} P_X(X = x_i) = \sum_{x_i \leq x} f_X(x_i)$$

i, per tant, $F_X(x)$ és una funció escalonada, els salts de la qual es produeixen en els punts de S_X.

Exercici

Dibuixar la funció de quantia i la funció de distribució associada a la variable aleatòria X= suma de les cares de dos daus equilibrats.

Exemple. Variable uniforme

Variable aleatòria discreta uniforme. $X \approx U(n)$

- $S_X = \{x_1, x_2, \ldots, x_n\}$
- $P(X = x_i) = \dfrac{1}{n}$

Exemple

Experiment típic: Es du a terme una prova que té un nombre finit de resultats equiprobables.

- monedes $n = 2$
- daus $n = 6$
- baralles $n = 40$
- ...

monedes, daus, etc. *equilibrats*.

Exemple. Variable de Bernoulli

Variable aleatòria de Bernoulli $X \approx B(p)$

- $S_X = \{0, 1\}$
- $P(X = 1) = p; \qquad p(X = 0) = 1 - p$

Un experiment aleatori s'anomena de Bernoulli si verifica:

- L'experiment consisteix a observar elements d'una població i classificar-los en dues categories: èxit (E) i Fracàs (F).
- La probabilitat d'èxit és p, i la de fracàs, $q = 1 - p$ en cada realització de l'experiment.
- Les observacions són indepedents.

Esperança i variància

Esperança

Esperança

Esperança d'una variable aleatòria discreta, X, amb suport S_X:

$$E(X) = \sum_{x \in S_X} xP(X = x)$$

Exemple

X= nombre de cares en llançar dues vegades una moneda equilibrada.

$S_X = \{0, 1, 2\}$

x	0	1	2
$p(x)$	1/4	1/2	1/4

$E(X) = \sum_{i=0}^{2} iP(X = i) = 0 \cdot (1/4) + 1 \cdot (1/2) + 2 \cdot (1/4) = 1$

Exemple

X= nombre d'uns en llançar dues vegades un dau equilibrat.

$S_X = \{0, 1, 2\}$

x	0	1	2
$p(x)$	25/36	10/36	1/36

$E(X) = \sum_{i=0}^{2} iP(X = i) = 0 \cdot (25/36) + 1 \cdot (10/36) + 2 \cdot (1/36) = 1/3$

Esperança i variància

Esperança

- En general, l'esperança és una mitjana ponderada dels valors que pren la variable, la qual cosa dóna més pes als valors més probables.
- $E(X)$ no té per què coincidir amb un dels valors de S_X.

Exemple

- Càlcul de l'esperan c ca d'una v. a. uniforme:

$$E(X) = \sum_{i=1}^{n} x_i P(X = x_i) = \sum_{i=1}^{n} x_i \frac{1}{n} = \bar{x}$$

- Càlcul de l'esperança d'una v. a. de Bernoulli

$$E(X) = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = 0 \cdot (1 - p) + 1 \cdot p = p$$

Esperança i variància

Variància

Variància

Variància d'una variable aleatòria discreta, X, amb suport S_X:

$$Var(X) = \sum_{x \in S_X} (x - E(X))^2 P(X = x)$$

Exemple

X= nombre de cares en llançar dues vegades una moneda equilibrada.

$S_X = \{0, 1, 2\}$

x	0	1	2
$p(x)$	1/4	1/2	1/4

$E(X) = 1$; $Var(X) = (0-1)^2 \frac{1}{4} + (1-1)^2 \frac{1}{2} + (2-1)^2 \frac{1}{4} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

La variància és una mesura de dispersió

Exemple

X= nombre d'uns en llançar dues vegades un dau equilibrat.

$S_X = \{0, 1, 2\}$

x	0	1	2
$p(x)$	25/36	10/36	1/36

$E(X) = 1/3$; $Var(X) = (0-1/3)^2 \frac{25}{36} + (1-1/3)^2 \frac{10}{36} + (2-1/3)^2 \frac{1}{36} = 0.2778$

Continguts

Distribució uniforme discreta

Distribució uniforme discreta

$$X \approx U(n)$$

$$S_X = \{x_1, x_2, \ldots, x_n\}, \qquad P(X = x_i) = \frac{1}{n}$$

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad Var(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

Exemple

Experiment típic: Es du a terme una prova que té un nombre finit de resultats equiprobables (monedes, daus, etc. *equilibrats*).

- monedes $n = 2$
- daus $n = 6$
- baralles $n = 40$
- ...

Distribució de Bernoulli

Distribució de Bernoulli

$$X \approx B(p)$$

$$S_X = \{0, 1\}, \qquad P(X = 1) = p; \qquad p(X = 0) = 1 - p$$

$$E(X) = p, \qquad Var(X) = pq = p(1 - p)$$

Un experiment aleatori es denomina de Bernoulli si verifica:

- L'experiment consisteix a observar elements d'una població i classificar-los en dues categories: èxit (E) i Fracàs (F).
- La probabilitat d'èxit és p, i la de fracàs, $q = 1 - p$ en cada realització de l'experiment.
- Les observacions són independents.

Distribució binomial

Distribució binomial

$$X \approx \mathcal{B}(n, p)$$

$X =$ nombre d'èxits en n proves de Bernoulli

$$S_X = \{0, 1, 2, \ldots, n\}, \qquad P(X = k) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$E(X) = np, \qquad Var(X) = npq$$

- L'experiment aleatori consisteix a repetir n vegades una prova (n fix).
- En cada prova s'observa si apareix o no un esdeveniment I.
- La probabilitat d'èxit val p i és la mateixa en totes les proves.
- El resultat d'una prova no influeix sobre les altres.

Es du a terme un nombre fix, n, de proves de Bernoulli

Distribució geomètrica

Distribució geomètrica

$$X \approx \mathcal{G}(p)$$

$X =$ nombre de vegades que cal repetir una prova de Bernoulli per obtenir el primer èxit:

$$S_X = \{1, 2, 3, 4, \ldots ...\}, \qquad P(X = k) = p(1-p)^{k-1}$$
$$E(X) = \frac{1}{p}, \qquad Var(X) = \frac{1-p}{p^2}$$

Exemple

$X =$ nombre de vegades que s'ha de llançar una moneda perquè isca cara.
$X = \mathcal{G}(p), \quad p = .5$

$$P(X = 2) = .5(.5) = .25, \qquad P(X = 5) = .5(.5)^4 = 0.0313$$

Exercici

Calculeu la probabilitat que perquè isca un 1 en un dau equilibrat hàgem de fer 5 llançaments. Calculeu la probabilitat de s'hagen de fer com a molt 5 llançaments.

Distribució binomial negativa

Distribució binomial negativa

$$X \approx \mathcal{BN}(r, p)$$

$X =$ nombre d'assajos necessaris per obtenir un nombre d'èxits fix, r, en proves de Bernoulli.

$$S_X = \{r, r+1, r+2, \ldots\}; \qquad P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

$$E(X) = \frac{r}{p}, \qquad Var(X) = \frac{r(1-p)}{p^2}$$

Observació

El quadre següent assenyala les diferències entre les variables binomial i binomial negativa.

	Nre. assajos	Nre. èxits
Binomial	fix	variable
Binomial negativa	variable	fix

Distribució hipergeomètrica

Processos d'extracció sense reemplaçament:

Suposem que en una urna hi ha N boles de les quals N_1 són blanques i $N_2 = N - N_1$ són negres. Extraiem m boles sense reemplaçament.

Distribució hipergeomètrica

$X =$ nombre de boles blanques $X \approx \mathcal{H}(N, m, N_1)$

$$P(X = k) = \frac{\binom{N_1}{k}\binom{N_2}{m-k}}{\binom{N}{m}}$$

$$E(X) = m\frac{N_1}{N}, \qquad Var(X) = m\frac{N_1}{N}(1 - \frac{N_1}{N})\frac{N-m}{N-1}$$

Distribució de Poisson

Procés de Poisson

Procés de Poisson

Experiment en el qual s'observa l'aparició de successos puntuals sobre un suport continu (interval de temps, de longitud, de superfície, etc.), de manera que:

1. El nombre de resultats esperats en un interval és independent del nombre de resultats en un altre interval disjunt.
2. La probabilitat que un succés ocòrrega en un interval de longitud molt petit és proporcional a la mesura de l'interval.
3. La probabilitat que ocòrrega mes d'un resultat en un interval curt és menyspreable.

Exemple

El nombre mitjà de partícules radioactives que passen a través d'un comptador Geiger durant 1ms en un experiment de laboratori és de 4. Quina és la probabilitat que entren 6 partícules al comptador en 1ms determinat?

$X =$ nombre de partícules que entren al comptador en 1ms

Calculeu $P(X = 6)$

Distribució de Poisson

Definició

Distribució de Poisson

$$X \approx \mathcal{P}(\lambda)$$

$X =$ nombre de successos en una unitat de suport, en un procés de Poisson.
$\lambda =$ mitjana esperada en una unitat de mesura.

$$S_X = \{0, 1, \ldots, \}, \qquad P_\lambda(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$E(X) = \lambda, \qquad Var(X) = \lambda$$

Exemple

El nombre mitjà de partícules radioactives que passen a través d'un comptador Geiger durant 1ms en un experiment de laboratori és de 4. Quina és la probabilitat que entren 6 partícules al comptador en 1ms determinat?

$X =$ nombre de partícules que entren al comptador en 1ms $\approx \mathcal{P}(4)$.

$$P_4(X = 6) = \frac{4^6 e^{-4}}{6!} = .1042$$

Distribució de Poisson

Exemples

Exemple

La distribució de Poisson serveix com a model aproximat per a les funcions de quantia de variables aleatòries del tipus següent:

- Nombre de polsos que explica un mesurador de radiació en un interval de longitud fixada (1ms, 2 minuts,...).
- Nombre d'accidents que ocorren en una ruta de carretera en un cert període de temps (1 setmana, 1 any,...).
- Nombre de trucades telefòniques que arriben a una centraleta en un cert període de temps (15 minuts, 1 dia,...).
- Nombre de paràsits que explica un tècnic de laboratori en un volum (fix) de sang (1ml, 1l). Es refereix a repetides mostres de la sang del mateix pacient.

Distribució de Poisson

Propietats

La distribució de Poisson té la propietat següent:

$$X \approx \mathcal{P}(\lambda)$$

$X =$ nombre de resultats de l'esdeveniment en un conjunt de mesura unitat,
$\lambda =$ el nombre mitjà de resultats d'un esdeveniment en un experiment en una unitat de mesura,
si $Y =$ nombre de resultats de l'esdeveniment en un conjunt que representa s unitats de mesura

$$Y \approx \mathcal{P}(\alpha) \text{ con } \alpha = \lambda s$$

A més, si

- $X_n = \mathcal{B}(n, p_n)$, $n = 1, 2, \ldots$ i
- $n \cdot p_n = \lambda \; \forall n$,

aleshores:

$$X_n \overset{n\to\infty}{\approx} \mathcal{P}(\lambda)$$

Distribució de Poisson

Exemple

Exemple

Suposem que en exploracions astronòmiques dutes a terme, amb una certa tècnica, es troben una mitjana de 3.2 galàxies per grau quadrat. Trobar l'àrea que hem d'explorar per tenir la seguretat en un 95% de trobar més de 100 galàxies.

$X =$ nombre de galàxies en l'àrea de mesura 1 grau quadrat $\approx \mathcal{P}(\lambda)$, $\lambda = 3.2$.

$Y =$ nombre de galàxies en l'àrea de mesura s graus quadrats $\approx \mathcal{P}(s\lambda)$, $\lambda = 3.2$.

$P(Y > 100) = 0.95, \quad \equiv \quad P(Y \leq 100) = .05$

Utilitzar la funció de distribució de la variable aleatòria.

Distribució de Poisson

Exemple

Exemple

Suposem que en exploracions astronòmiques dutes a terme, amb una certa tècnica, es troben una mitjana de 3.2 galàxies per grau quadrat. Trobar l'àrea que hem d'explorar per tenir la seguretat en un 95% de trobar més de 100 galàxies.

$X =$ nombre de galàxies en l'àrea de mesura 1 grau quadrat $\approx \mathcal{P}(\lambda)$, $\lambda = 3.2$.

$Y =$ nombre de galàxies en l'àrea de mesura s graus quadrats $\approx \mathcal{P}(s\lambda)$, $\lambda = 3.2$.

$P(Y > 100) = 0.95, \quad \equiv \quad P(Y \leq 100) = .05$

Utilitzar la funció de distribució de la variable aleatòria.

Com que $X \approx \mathcal{P}(\alpha s)$, es tracta de calcular λ perquè $P(X \leq 100) = .05$ en $X \approx \mathcal{P}(\lambda)$.

$\lambda = 118$, $s = 37$ graus quadrats.

Continguts

Conceptes bàsics

El concepte de variable aleatòria sorgeix de la necessitat de fer mes manejables matemàticament els resultats dels experiments aleatoris, que en molts casos són qualitatius i segueixen patrons molt similars, encara que la naturalesa de l'experiment no ho siga.

Exemple

- Experiment 1: Tirar una moneda defectuosa tal que $p(C) = .9$.
  $\Omega = \{C, X\} \quad \equiv \quad \{0, 1\}, p(0) = P(C) = .9$
- Experiment 2: Observar una peça fabricada en un procés que produeix un 1% de peces defectuoses.
  $\Omega = \{B, D\} \quad \equiv \quad \{0, 1\}, p(B) = p(0) = .9$

Aquests dos experiments aleatoris són *essencialment* el mateix!

El maneig de variables aleatòries permet *classificar* els experiments aleatoris, de manera que el càlcul de probabilitats és molt mes senzill.

Definició

V. a. discretes: models per a dades que tenen una quantitat numerable de resultats amb probabilitat no nul·la.

V. a. contínues: models per a registres de dades contínues.

Exemple

- contingut de glucosa en una solució
- mesura del diàmetre interior d'un rodament
- altura d'un individu...

Per a una variable aleatòria contínua,

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

on $F_X(x)$ és la funció de distribució de la variable aleatòria.

Per a una v. a. contínua, $F_X(x)$ és una funció contínua:

No té salts $\rightarrow P(X = a) = 0, \forall a$.

No hi ha funció de quantia per a v. a. contínues.

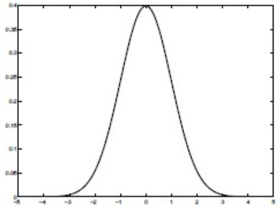
Exemple: la normal

Per a una v. a. normal(0,1), la funció de distribució és:

$$P(X \leq x) = F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2} dx$$

Observeu que el costat dret defineix una àrea sota la corba que defineix la funció

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



Aquesta és la situació habitual per a v. a. contínues.

Funció de densitat

Definició

Definició

Una variable aleatòria és (absolutament) contínua si hi ha una funció

$$f : \mathbb{R} \quad \rightarrow \quad \mathbb{R}^+$$

de manera que la seua funció de distribució $F_X(x) = P(X \leq x)$ compleix

$$F_X(x) = \int_{-\infty}^{x} f(t)dt$$

Definició

La funció $f(t)$ és la funció de densitat (funció de densitat de probabilitat) de la v. a. X.

$$P(X \leq a) = F_X(a) = \int_{-\infty}^{a} f(t)dt$$

$$P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_{a}^{b} f(t)dt, \qquad \forall a \leq b$$

En una v. a. contínua, la probabilitat d'un interval de $\mathbb{R}$ es calcula mesurant l'àrea que tanca la funció de densitat de probabilitat sobre l'interval.

Funció de densitat

Propietats

Les condicions que ha de complir $f : \mathbb{R} \rightarrow \mathbb{R}^+$ per a ser una funció de densitat de probabilitat són:

- $f(x) \geq 0, \quad \forall x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f(t)dt = 1$

Exemple

$$f(x) = \alpha \mathbb{I}_{[0,1]}(x) = \alpha \begin{cases} 1 & x \in [0,1] \\ 0 & \text{si no} \end{cases}$$

- Què ha de valdre α perquè $f(x)$ siga una funció de densitat de probabilitat?
- Calculeu $F(x)$.
- Què val $P(X > 0.5)$?

Exemple

$$f(x) = e^{-x} \mathbb{I}_{[0,\infty]}(x)$$

Comproveu que $f(x)$ *pot actuar* com a funció de densitat de probabilitat d'una v. a. contínua. Calculeu $F(x)$.

Esperança i variància

Si X és una v. a. contínua amb funció de densitat $f(x)$, aleshores:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \qquad Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)dx$$

Exercici

Calculeu $E(X)$ i $Var(X)$ en els exemples anteriors.

Tema 1. Teoria del Mostreig Estadístic
Francisco Guerrero Cortina
Variables aleatòries: El cas discret
Variable aleatòria
Models probabilístics discrets
Variables aleatòries: El cas continu
Variable aleatòria
Models probabilístics continus
La normal i la seua importància
La distribució normal
Sumes de variables aleatòries
Teorema central del límit

Continguts

Distribució uniforme contínua

$X \approx U[a, b]$

La seua funció de densitat és:

$$f(x) = \frac{1}{b-a}\mathbb{I}_{[a,b]}(x), \qquad F(x) = \begin{cases} 0 & x \leq a \\ \dfrac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b \end{cases}$$

$$E(X) = \frac{a+b}{2}, \qquad Var(X) = \frac{(b-a)^2}{12}$$

Distribució exponencial

$X \approx Exp(\lambda)$:

$$f(x) = \lambda e^{-\lambda x} \mathbb{I}_{[0,+\infty]}, \quad F(x) = \begin{cases} 0 & x \leq 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

$$E(X) = \frac{1}{\lambda}, \qquad Var(X) = \frac{1}{\lambda^2}$$

Distribució normal

$X \approx \mathcal{N}(\mu, \sigma)$

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

$$E(X) = \mu, \qquad Var(X) = \sigma^2$$

Continguts

Definició

Es diu que una v. a. (uniformement) contínua té una distribució normal estàndard o normal tipificada si la seua funció de densitat de probabilitat és:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

No hi ha una fòrmula tancada per a la funció de distribució

$$F(x) = \int_{-\infty}^{x} f(z)dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\left(-\frac{1}{2}z^2\right)} dz$$

És fàcil calcular l'esperança i la variància d'aquesta v. a.

$$E(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x e^{\left(-\frac{1}{2}x^2\right)} dx = 0,$$

$$Var(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{\left(-\frac{1}{2}x^2\right)} dx = 1$$

Així doncs, aquesta v. a. s'anomena per $X \approx \mathcal{N}(0,1)$.

Distribució normal de mitjana μ i variància σ^2

Es diu que una v. a. (uniformement) contínua té una distribució normal de mitjana μ i variància $\sigma^2, X \approx \mathcal{N}(\mu, \sigma)$ si la seua funció de densitat de probabilitat és:

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

És molt senzill (canvi de variables en la integral) comprovar que

$$E(X) = \mu, \qquad Var(X) = \sigma^2$$

També és senzill comprovar (canvi de variables en la integral) que si

$$X \approx \mathcal{N}(\mu, \sigma), \quad \Leftrightarrow \quad Z = \frac{X-\mu}{\sigma} \approx \mathcal{N}(0, 1)$$

Aquesta propietat era essencial en els temps de càlculs amb taules.

Propietats

1. És simètrica respecte de la mitjana.
2. La funció de densitat té punts d'inflexió en $\mu \pm \sigma$.
3. La funció de densitat tendeix assimptòticament a 0 quan $x \to \pm\infty$.
4. $P(a \leq X \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$

   En $[\mu - 2\sigma, \mu + 2\sigma]$ es troba el 95.5 % de l'àrea sota $f(x)$.

   En $[\mu - 3\sigma, \mu + 3\sigma]$ es troba el 99.7 %.

Exemple

Moltes variables aleatòries, com pesos, altures, etc., segueixen una distribució normal. De la mateixa manera, els errors accidentals o aleatoris que es presenten en assajos de laboratori i observacions experimentals també es distribueixen normalment.

Continguts

V. a. independents

Hi ha moltes situacions en les quals la variable aleatòria que interessa es pot escriure com la suma de diverses variables aleatòries.

Exemple

Temps total que es necessita per a dur a terme una tasca que consisteix en diverses subtasques.

Propietats generals

- Si X i Y són dues v. a. $\boxed{E(X+Y)=E(X)+E(Y)}$.
- Si, a més, X i Y són independents $\boxed{Var(X+Y)=Var(X)+Var(Y)}$.

Variables aleatòries independents

Es diu que dues v. a. X i Y són independents si qualsevol succés associat amb X és independent de qualsevol succés associat amb Y:

$$P((X \in A) \cap (Y \in B)) = P(X \in A) \cdot P(Y \in B)$$

per a qualsevol $A \subset \mathbb{R}$, $B \subset \mathbb{R}$.

Propietats

En general:

$X_1, \ldots, X_n$ són variables aleatòries amb $E(X_i) = \mu_i$, aleshores

$$E(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} \mu_i$$

si a més $X_1, \ldots, X_n$ són independients, i $Var(X_i) = \sigma_i^2$, aleshores

$$Var(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} \sigma_i^2$$

Sumes de variables aleatòries independents

Bernoulli i binomials

Si X_i són v. a. de Bernoulli amb paràmetre p, és a dir,

$$X_i \approx \mathcal{B}(p), \qquad i = 1, 2, ...$$

i X_i són independents, aleshores la suma és una v. a. binomial

$$X_1 + X_2 + \ldots + X_k \approx \mathcal{B}(k, p)$$

La suma de binomials independents amb el mateix paràmetre és una binomial.

$$X_i \approx \mathcal{B}(n_i, p), \qquad i = 1, 2, ... \text{y } X_i \text{ independents, aleshores}$$

$$X_1 + X_2 + \ldots + X_k \approx \mathcal{B}(\sum_{i=1}^{k} n_i, p)$$

Sumes de variables aleatòries independents

Poisson i normal

La suma de v. a. de Poisson independents és una Poisson

$$X_i \approx \mathcal{P}(\lambda_i), \qquad i = 1, 2, \ldots \text{ i } X_i \text{ independents, aleshores}$$

$$X_1 + X_2 + \ldots + X_k \approx \mathcal{P}(\sum_{i=1}^{k} \lambda_i)$$

La suma de normals independents és una normal

$$X_i \approx \mathcal{N}(\mu_i, \sigma_i), \qquad i = 1, 2, \ldots \text{ i } X_i \text{ independents, aleshores}$$

$$X_1 + X_2 + \ldots + X_k \approx \mathcal{N}(\sum_{i=1}^{k} \mu_i, \sqrt{\sum_{i=1}^{k} \sigma_i^2})$$

Aquesta propietat additiva no s'estàn en uns altres models probabilístics.

Continguts

Teorema

Si $X_1, \dots, X_n, \dots$ és una successió de variables aleatòries independents amb $E(X_i) = \mu_i$ i $Var(X_i) = \sigma_i^2$, aleshores per a $n \to \infty$, la successió de variables aleatòries:

$$S_n = \sum_{k=1}^{n} X_k \overset{n\to\infty}{\approx} \mathcal{N}\left(\sum_{k=1}^{n} \mu_k, \sqrt{\sum_{k=1}^{n} \sigma_i^2}\right),$$

També s'escriu

$$Z_n = \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i}{(\sum_{i=1}^{n} \sigma_i^2)^{1/2}} \overset{n\to\infty}{\approx} \mathcal{N}(0,1)$$

És a dir, si $F_n(x)$ és la funció de distribució de la variable aleatòria Z_n, $n = 1, 2, \dots$ i $\phi(x)$ és la funció de distribució d'una variable $\mathcal{N}(0,1)$, aleshores per a cada $x \in \mathbb{R}$

$$lim_{n\to\infty} F_n(x) = \phi(x)$$

Aplicacions del TCL

Si $X_1, \ldots, X_n, \ldots$ és una successió de variables aleatòries i. i. d.

i. i. d. $==$ independents, idénticament distribuïdes,

amb mitjana μ i variància σ^2, aleshores per a $n \to \infty$

$$S_n \approx \sum_{k=1}^{n} X_k \overset{n\to\infty}{\approx} \mathcal{N}\left(n\mu, \sigma\sqrt{n}\right)$$

Si apliquem el TCL a una successió $X_1, X_2, \ldots,$ de variables de Bernoulli i. i. d.

$$X_i \approx \mathcal{B}(p) \quad E(X_i) = p, Var(X_i) = pq \quad \Rightarrow \quad X_1 + X_2 + \ldots X_n \overset{n\to\infty}{\approx} \mathcal{N}(np, \sqrt{npq})$$

Teorema de De Moivre

$$\mathcal{B}(n,p) \quad \overset{n\to\infty}{\approx} \quad \mathcal{N}(np, \sqrt{npq})$$

Aplicacions: importància de la normal

En moltes situacions, els resultats d'un experiment són conseqüència de *múltiples causes de petita incidència individual però els efectes de la qual se sumen, fet que dóna lloc als resultats de l'experiment*.

Exemple

Suposem que un error d'arredoniment es representa mitjançant una v. a. uniforme en $[-0.5, 0.5]$. Si se sumen 100 nombres, trobar la probabilitat que l'error d'arredoniment excedisca 1 en valor absolut.

X_i= error d'arredoniment associat al nombre i-èsim.

$$X_i \approx U([-0.5, 0.5]), \qquad E(X_i) = 0; \quad Var(X_i) = 1/12$$

$X = \sum_{i=1}^{n} X_i$ representa l'error d'arredoniment associat a la suma de n nombres.

Com que les X_i són i. i. d., segons el TCL, per a

$$n >> 1, \quad \sum_{i=1}^{n} X_i \approx \mathcal{N}(n \cdot 0, \sqrt{n/12})$$

Aplicacions: importància de la normal

Exemple

$$\sum_{i=1}^{100} X_i \approx \mathcal{N}(0, \sqrt{1000/12} = \mathcal{N}(0, 2.89)$$

$$P(X > a) = 1 - P(X \leq a); \qquad P(X > 1) = 1 - F(1) = 1 - 0.63534 = 0.3647$$

$$P(X > 2) = 0.2446; \qquad P(X > 5) = 0.0418$$

En aquestes situacions, el model normal sol aproximar bé el comportament dels resultats de l'experiment.

En la majoria de les aplicacions es pot aplicar el TCL a partir de $n \geq 30$.

Tema 1. Teoria del Mostreig Estadístic (Part II)

Matemàtiques II. Grau de Química

Francisco Guerrero Cortina

Dept. Matemàtica Aplicada, Universitat de València

Curs 2012-2013

Tema 1. Teoria del Mostreig Estadístic (Part III)

Matemàtiques II. Grau de Química

Francisco Guerrero Cortina

Dept. Matemàtica Aplicada, Universitat de València

Curs 2012-2013

Continguts

Continguts

Introducció

Procediments que permeten prendre decisions en problemes de naturalesa aleatòria.

Objectiu

Obtenir <u>informació</u> d'un conjunt d'elements al qual solament es pot accedir parcialment, és a dir, un conjunt del qual podem (o volem) analitzar solament alguns elements.

Informació

Un cert model probabilístic es considera apropiat per descriure el fenomen aleatori objecte d'estudi. Es vol esbrinar algun o alguns dels paràmetres que defineixen la funció de distribució del model, a partir de dades (experimentals) basades en mostres de la població sota estudi.

Exemple

Determinar l'altura mitjana dels xiquets de 10 anys residents a la ciutat de València, a partir de les altures d'una mostra formada per 400 xiquets.

L'altura mitjana obtinguda a partir de les dades de la mostra és un estimador de la mitjana de la població. Necessitem instruments per a mesurar-ne la fiabilitat.

Definicions

Situació habitual: Es coneix com és el model probabilístic adequat, però manca informació sobre algun o alguns dels paràmetres que en caracteritzen la funció de distribució.

Exemple

X = altura d'un xiquet de 10 anys, triat a l'atzar $\approx \mathcal{N}(\mu, \sigma)$

- Es volen obtenir estimacions raonables d'un (o diversos) dels paràmetres que defineixen el model probabilístic, a partir de les dades d'una mostra de la població. (μ, σ) en el cas d'un model normal.
  - Estimació puntual: determinació de valors que representen estimacions raonables del paràmetre en qüestió, a partir d'una mostra de la població.
  - Intervals de confiança: determinació d'intervals en els quals s'espera trobar el paràmetre amb un cert nivell de confiança.

- Contrast d'hipòtesis: Estimacions sobre afirmacions sobre la distribució de probabilitat del fenomen estudiat.

Exemples

Exemple

Estimar la mitjana de les alçades dels xiquets de 10 anys en la ciutat de València.

$X =$ alçada d'un xiquet de 10 anys que viu a València. Determinar $E(X)$, el valor esperat per a l'altura o mitjana poblacional.

Com podem fer açò?

Prenem una *mostra* de n xiquets i en mesurem les altures. Esperem que la mitjana de les altures de la mostra represente la mitja de la població, $E(X)$?

Necessitem un model probabilístic per a X.

Necessitem entendre els rudiments de la teoria del mostreig aleatori.

Exemples

Exemple

Estimar la mitjana de les alçades dels xiquets de 10 anys en la ciutat de València.

$X =$ alçada d'un xiquet de 10 anys que viu a València. Determinar $E(X)$, el valor esperat per a l'altura o mitjana poblacional.

Com podem fer açò?

Prenem una *mostra* de n xiquets i en mesurem les altures. Esperem que la mitjana de les altures de la mostra represente la mitja de la població, $E(X)$?

Necessitem un model probabilístic per a X.

Necessitem entendre els rudiments de la teoria del mostreig aleatori.

Exemples

Exemple

Estimar la mitjana de les alçades dels xiquets de 10 anys en la ciutat de València.

$X =$ alçada d'un xiquet de 10 anys que viu a València. Determinar $E(X)$, el valor esperat per a l'altura o mitjana poblacional.

Com podem fer açò?

Prenem una *mostra* de n xiquets i en mesurem les altures. Esperem que la mitjana de les altures de la mostra represente la mitja de la població, $E(X)$?

Necessitem un model probabilístic per a X.

Necessitem entendre els rudiments de la teoria del mostreig aleatori.

Continguts

Definicions

Un procediment de mostreig no és mes que un procés que permet seleccionar elements (mostres) en una població.

Mostreig aleatori simple

Selecciona elements d'una població de manera que cada element té idèntica probabilitat de ser triat en cada extracció.
En poblacions finites, un mostreig aleatori simple és equivalent a un procés d'extraccions amb reemplaçament.

Si X és una variable aleatòria i Ω es el seu espai mostral,

$$\begin{aligned} X : \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega) = x \end{aligned}$$

Una mostra aleatòria de grandària n és un vector

$$(\omega_1, \ldots, \omega_n) \in \Omega^n$$

L'espai mostral Ω^n és el conjunt de totes les mostres de grandària n que es poden extraure d'Ω.

Mostreig aleatori simple

Exemple

Se sap que l'altura dels xiquets de 10 anys de València segueix una distribució normal de mitjana μ i variància σ^2.
X= altura d'un xiquet de 10 anys triat a l'atzar $\approx \mathcal{N}(\mu, \sigma)$.

μ és la mitjana poblacional i σ^2 la variància poblacional

Prenem una mostra de 400 xiquets triats aleatòriament entre els xiquets de la població = xiquets de 10 anys de València.

$X_i =$ altura del xiquet i en la mostra. $X_i \approx \mathcal{N}(\mu, \sigma)$, $i = 1, \ldots, n$
$(X_1, X_2, ...X_{400})$ és un m. a. s. de grandària 400 de X.

Definició

Un mostreig aleatori simple (*m. a. s.*) de grandària n per a una variable aleatòria X és un vector aleatori $(X_1, X_2, \ldots, X_n)$ on X_i, $i = 1, \ldots, n$ són variables aleatòries i. i. d. de manera que la distribució de probabilitat de X_i és la mateixa que la de X.

Cada component d'un *m. a. s.* representa els possibles valors de la component i -èsima d'una mostra aleatòria de grandària n de la població.

Exemple

Exemple

Se sap que l'altura dels xiquets de 10 anys de València segueix una distribució normal de mitjana μ i variància σ^2.
X= altura d'un xiquet de 10 anys triat a l'atzar $\approx \mathcal{N}(\mu, \sigma)$.

μ és la mitjana poblacional i σ^2 la variància poblacional

Prenem una mostra de 400 xiquets triats aleatòriament entre els xiquets de la població = xiquets de 10 anys de València.

$X_i =$ altura del xiquet i en la mostra. $X_i \approx \mathcal{N}(\mu, \sigma)$, $i = 1, \ldots, n$
$(X_1, X_2, ...X_{400})$ és un m. a. s. de grandària 400 de X.

La mitjana mostral

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

és una estimació raonable per a μ?

D'acord amb aquestes definicions $\bar{X}$ és una v. a.

Continguts

Preliminars

- X, Y v. a.; $a, b \in \mathbb{R} \rightarrow E(aX + bY) = aE(X) + bE(Y)$
- X, Y v. a. independents $\rightarrow$ $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$

Per tant:

$X_i, i = 1, \ldots, n$ v. a. $\rightarrow$ $E(\sum_{i=1}^n X_i) = \sum_{i=1}^n E(X_i)$

$X_i, i = 1, \ldots, n$ v. a. $\rightarrow$ $E(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{n}\sum_{i=1}^n E(X_i)$

$X_i, i = 1, \ldots, n$ v. a. independents $\rightarrow$ $Var(\sum_{i=1}^n X_i) = \sum_{i=1}^n Var(X_i)$

$X_i, i = 1, \ldots, n$ v. a. independents $\rightarrow$ $Var(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i)$

Estadístics i estimadors

Estadístics

Estadístic

X v. a. $(X_1, \ldots, X_n)$ una *m. a. s.* de grandària n. Un estadístic és una nova variable aleatòria definida mitjançant una aplicació

$$T : (X_1, \ldots, X_n) \quad \rightarrow \quad \mathbb{R}$$

Exemple

Si la grandària mostral és $n = 3$,

- $T_1(X_1, X_2, X_3) = \dfrac{1}{3}(X_1 + X_2 + X_3) = \bar{X}$ mitjana mostral
- $T_2(X_1, X_2, X_3) = \dfrac{1}{3}\sum_{i=1}^{3}(X_i - \bar{X})^2$ variància mostral

Estadístics i estimadors

Estimadors

Estimador

Un estimador $\hat{\theta}$ és un estadístic que resulta apropiat per a estimar el valor d'un paràmetre θ.

Exemple

La mitjana mostral $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ és un estadístic que s'utilitza com a estimador per a la mitjana poblacional.

Els estimadors són estadístics, és a dir són v. a.

Estimació per intervals de confiança

Introducció

- El valor d'un estimador per a una mostra concreta és una estimació.
- Una estimació depén de la mostra poblacional triada, de manera que si s'extrauen dues mostres diferents, les estimacions seran diferents.
- és desitjable proporcionar una mesura de la precisió de l'estimació del paràmetre a partir de l'estadístic (estimador) considerat.

Els intervals de confiança proporcionen aquesta mesura.

Estimació per intervals de confiança

Exemple: IC per a μ, coneixent σ

Exemple

Siga $X \approx \mathcal{N}(\mu, \sigma)$ amb μ desconeguda i σ coneguda.

Volem estimar μ amb un nivell de confiança del 95% a partir d'una m. a. s.:

$(X_1, \ldots, X_n)$, $\quad X_i$ i.i.d, $\approx \mathcal{N}(\mu, \sigma)$.

Considerem l'estadístic mitjana mostral

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad E(\bar{X}) = \mu, \quad Var(\bar{X}) = \frac{\sigma^2}{n}$$

Com que $X_i \approx \mathcal{N}(\mu, \sigma) \to \bar{X} \approx \mathcal{N}(\mu, \sigma/\sqrt{n})$. Estandarditzem

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx \mathcal{N}(0, 1)$$

Z és una v. a. amb distribució coneguda que no depén de cap paràmetre. Podem calcular fàcilment un interval centrat $[-z^*, z^*]$ de manera que

$$P(-z^* \leq Z \leq z^*) = .95$$

Estimació per intervals de confiança

Exemple

Exemple

$$P(-z^* \leq Z \leq z^*) = .95$$

utilitzant matlab/taules/..., obtenim $z^* = 1.96$.

$$-z^* \leq Z \leq z^* \quad \equiv \quad -z^* \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z^* \quad \equiv \quad \bar{X} - z^* \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - z^* \frac{\sigma}{\sqrt{n}}$$

$$P(-z^* \leq Z \leq z^*) = P(\bar{X} - z^* \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - z^* \frac{\sigma}{\sqrt{n}})$$

Per tant,

$$[\bar{X} - z^* \frac{\sigma}{\sqrt{n}}, \bar{X} - z^* \frac{\sigma}{\sqrt{n}}]$$

amb $z^* = 1.96$ és un interval de confiança per a la mitjana poblacional μ amb un nivell de significació del 95%.

El significat d'aquest IC és el següent: Si es prenen mostres de n elements, hi ha una probabilitat del 95% que la mitjana de la mostra triada estiga en l'IC.

Continguts

Recordatori

Els I.C. serveixen per a estimar el valor d'un paràmetre poblacional a partir d'una m. a. s. i proporcionen una mesura de l'error d'estimació.

La seua construcció es basa en

- un estadístic (v. a. relacionada amb la m. a. s.) que siga un estimador (no esbiaixat) del paràmetre poblacional;
- una v. a. relacionada amb l'estadístic que tinga una distribució coneguda i independent de qualsevol paràmetre desconegut, excepte aquell que es vol estimar.

IC per a μ en una distribució normal

μ desconegut i σ conegut

$X \approx N(\mu, \sigma)$ amb μ desconegut i σ conegut

$(X_1, X_2, \ldots, X_n)$ m. a. s. de X

Calcular IC per a μ con nivell de confiança $1 - \alpha$.

- Estimador: Mitjana mostral $\boxed{\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathcal{N}(\mu, \sigma/\sqrt{n})}$

$$E(\bar{X}) = \mu, \qquad Var(\bar{X}) = \sigma^2/n$$

- Tipifiquem $\boxed{Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx \mathcal{N}(0, 1)}$

Per a calcular l'IC per a μ amb nivell de confiança $1 - \alpha$, determinem un interval centrat que continga l'$(1 - \alpha)100\%$ de la probabilitat de la v. a. Z, és a dir, determinem z^* de manera que

$$P(-z^* \leq Z \leq z^*) = 1 - \alpha$$

IC per a μ en una distribució normal

μ desconegut i σ conegut

$$P(-z^* \leq Z \leq z^*) = 1 - \alpha \quad \equiv \quad P(-z^* \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z^*) = 1 - \alpha$$

Manipulem: $$-z^* \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z^* \qquad \equiv \qquad \bar{X} - z^* \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z^* \frac{\sigma}{\sqrt{n}}$$

Qui és z^*, tal que $P(-z^* \leq Z \leq z^*) = 1 - \alpha$?

Per la simetria de la $f(x)$ per a $Z \approx \mathcal{N}(0,1)$ s'ha de complir:

$$P(Z \leq -z^*) = \frac{\alpha}{2}, \qquad P(Z > z^*) = \frac{\alpha}{2} \quad \rightarrow \quad P(Z \leq z^*) = 1 - \frac{\alpha}{2}$$

Se sol utilitzar la nomenclatura $z^* = z_{1-\frac{\alpha}{2}}$

IC per a μ amb nivell de confiança $1 - \alpha$: $\quad [\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \;\; \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$

IC per a μ en una distribució normal

μ desconegut i σ conegut

Exemple

S'ha registrat el valor (en Kg) de la reducció del pes de cadascun dels pacients d'una mostra triada a l'atzar en una determinada població, després d'una setmana de tractament. La mitjana dels 16 valors obtinguts és de 3.42Kg. Suposant que la pèrdua de pes és una v. a. amb una distribució $\mathcal{N}(\mu, \sigma)$, construir un interval de confiança del 95% per al valor mitjà poblacional de la reducció de pes després d'una setmana de tractament

1. en el cas en què $\sigma = .68$Kg
   `norminv(.025,0,1)=-1.96` `Sol:` `IC=` $[3.0868, 3.7532]$
2. en el cas en què σ és desconeguda, però s'ha mesurat la variància poblacional i s'ha obtingut que $S = 0.68$Kg
   La solució NO ÉS $[3.0868, 3.7532]$.

IC per a μ en una distribució normal

μ desconegut i σ desconegut

$X \approx N(\mu, \sigma)$ amb μ desconegut i σ desconegut

$(X_1, X_2, \ldots, X_n)$ m. a. s. de X

Calcular IC per a μ amb nivell de confiança $1 - \alpha$.

Estimador: Mitjana mostral $\boxed{\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i}$

Si σ és desconeguda, la variable $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx \mathcal{N}(0, 1)$ no serveix per a establir l'IC per a μ, ja que no podem calcular els extrems de l'IC

$$\bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

a partir d'informació coneguda.

Necessitem un estimador per a σ!

La distribució χ^2_n

Proposició

Siguen $Z_1, Z_2, \ldots, Z_n$ v. a. i. i. d. $\mathcal{N}(0,1)$. Aleshores la suma dels quadrats té una distribució khi-quadrat amb n graus de llibertat

$$X = Z_1^2 + Z_2^2 + \ldots Z_n^2 \quad \approx \chi^2_n$$

Aquesta distribució té les propietats següents:

- $S_X = [0, \infty)$
- $E(X) = n$, $Var(X) = 2n$
- Propietat de reproductivitat Si $X_1, X_2, \ldots, X_k$ són v. a. independents, i

$$X_i \approx \chi^2_{n_i} \quad \Rightarrow \quad X = X_1 + \ldots + X_k \quad \approx \quad \chi^2_m, \quad m = n_1 + \ldots + n_k$$

- Resta Si $Z = X + Y$, amb X, Y independents, i $Z \approx \chi^2_n$, $\quad X = \chi^2_{n_1}$ amb $n > n_1$, aleshores

$$Y \quad \approx \chi^2_{n-n_1}$$

L'estadístic variància mostral

Si $(X_1, \dots, X_n)$ és una m. a. s. d'una v.a. X, es defineix l'estadístic

Variància mostral $$S^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Proposició

Siga $X \approx \mathcal{N}(\mu, \sigma)$, y $(X_1, \dots, X_n)$ una m. a. s de X. La v. a.

$$\frac{nS^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \quad \approx \quad \chi^2_{n-1}$$

La distribució T de Student

Definició

Definició

Siga $Z \approx \mathcal{N}(0,1)$, $Y \approx \chi^2_n$. Si Z i Y són independents, la v. a.

$$\frac{Z}{\sqrt{Y/n}} \approx T_n$$

és una T de Student amb n graus de llibertat.

Aquesta distribució té les propietats següents:

- $S_X = (-\infty, \infty)$
- $E(X) = 0$, $Var(X) = \frac{n}{n-2}$
- És simètrica respecte de $x = 0$ i similar a la normal, però amb cues més àmplies
- $T_n \to \mathcal{N}(0,1)$ quan $n \to \infty$. L'aproximació és OK per a $n > 30$.

IC per a μ en una distribució normal

μ desconegut i σ desconegut

$X \approx N(\mu, \sigma)$ amb μ desconegut i σ desconegut

$(X_1, X_2, \ldots, X_n)$ m. a. s. de X

Calcular IC per a μ amb nivell de confiança $1 - \alpha$.

- Estimador Mitjana mostral: $$\boxed{\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathcal{N}(\mu, \sigma/\sqrt{n})}$$

- Utilitzem $$\boxed{Z = \frac{\bar{X} - \mu}{S/\sqrt{n-1}} \approx T_{n-1}}$$

IC per a μ en una distribució normal

IC per a μ amb nivell de confiança $1-\alpha$

Determinem un interval centrat que continga el $(1-\alpha)100\%$ de la probabilitat de la v. a. Z, és a dir, determinem z^* de manera que

$$P(-z^* \leq Z \leq z^*) = 1-\alpha \quad \equiv \quad P\left(-z^* \leq \frac{\bar{X}-\mu}{S/\sqrt{n-1}} \leq z^*\right) = 1-\alpha$$

Manipulem: $$-z^* \leq \frac{\bar{X}-\mu}{S/\sqrt{n-1}} \leq z^* \quad \equiv \quad \bar{X} - z^* \frac{S}{\sqrt{n-1}} \leq \mu \leq \bar{X} + z^* \frac{S}{\sqrt{n-1}}$$

Qui és z^*, de manera que $P(-z^* \leq Z \leq z^*) = 1-\alpha$?

Per la simetria de la $f(x)$ per a una T de Student, s'ha de complir:

$$P(Z \leq -z^*) = \frac{\alpha}{2}, \qquad P(Z > z^*) = \frac{\alpha}{2} \quad \rightarrow \quad P(Z \leq z^*) = 1 - \frac{\alpha}{2}$$

Se sol utilitzar la nomenclatura $z^* = t_{n-1,1-\frac{\alpha}{2}}$

IC per a μ amb nivell de confiança $1-\alpha$:

$$\left[\bar{X} - t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}}, \;\; \bar{X} + t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}}\right]$$

IC per a μ en una distribució normal

Exemple

Exemple

S'ha registrat el valor (en Kg) de la reducció del pes de cadascun dels pacients d'una mostra triada a l'atzar en una determinada població, després d'una setmana de tractament. La mitjana dels 16 valors obtinguts és de 3.42Kg. Suposant que la perduda de pes és una v. a. amb una distribució $\mathcal{N}(\mu, \sigma)$, construir un interval de confiança del 95% per al valor mitjà poblacional de la reducció de pes després d'una setmana de tractament

1. en el cas que $\sigma = .68$Kg

   Sol: IC$= [3.0868, 3.7532]$

2. en el cas que σ siga desconeguda, però s'haja mesurat la variància poblacional i s'haja obtingut que $S = 0.68$Kg

   Sol: $\bar{X} = 3.42$ $t_{15,0.975} = 2.1314$ IC$= [3.0458, 3.7942]$

Error quadràtic mitjà d'un estimador

Si $X_1, X_2, \ldots, X_n$ és una m. a. s. de X, una v. a. que depén d'un paràmetre θ desconegut i $\hat{\theta} = g(X_1, \ldots, X_n)$ un estimador de θ.

La *bondat* de l'estimació es pot mesurar calculant l'error quadràtic mitjà

$$\text{EQM}\ (\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

Es pot provar que $\boxed{\text{EQM}\ (\hat{\theta}) = Var(\hat{\theta}) + \left(E(\hat{\theta}) - \theta\right)^2}$

Per a minimitzar l'EQM, convé que els estimadors tinguen

- La variància petita
- El biaix petit $\boxed{\text{biaix de } \hat{\theta} = E(\hat{\theta}) - \theta}$

Exemple

$X \approx \mathcal{N}(\mu, \sigma)$, $X_1, \ldots, X_n$ m. a. s. de X.

$\bar{X}$ és un estimador no esbiaixat per a μ perquè $E(\bar{X}) = \mu$, $\rightarrow$ biaix de $\bar{X} = 0$.

$\text{EQM}\ (\bar{X}) = \sigma^2/n$

Estimadors per a la variància

$$X \approx \mathcal{N}(\mu,\sigma) \quad \Rightarrow \quad E(X)=\mu,\ Var(X)=\sigma^2. \qquad (X_1,\dots,X_n) \text{ m. a. s. de } X.$$

- $\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ és un estimador no esbiaixat per a σ^2.
- La variància mostral:
  $S^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$ és un estimador esbiaixat per a σ^2.
- La quasivariància mostral:
  $S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ és un estimador no esbiaixat per a σ^2.

Noteu que

$$\boxed{\frac{nS^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma^2} = \frac{(n-1)S_c^2}{\sigma^2} \approx \chi^2_{n-1}}$$

IC per a la variància en una distribució normal

μ coneguda i σ desconeguda

$X \approx \mathcal{N}(\mu, \sigma)$ amb μ coneguda i σ desconeguda. Volem estimar σ^2.

Estimador: $\boxed{\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2}$ és un estimador no esbiaixat per a σ^2.

Variable aleatòria que cal considerar: $Z = \dfrac{n\hat{S}^2}{\sigma^2} \approx \chi^2_n$

que sols depén d'un paràmetre desconegut: σ, i de la m. a. s.

Amb ella calculem l'IC al nivell de confiança desitjat.

Compte, la χ^2 no és simètrica.

IC per a la variància en una distribució normal

μ coneguda i σ desconeguda

Per la falta de simetria de la $f(x)$ per a una χ^2, triem z_1 i z_2 de manera que

$$P(z_1 \leq Z \leq z_2) = 1 - \alpha$$

i de manera que $P(Z \leq z_1) = \frac{\alpha}{2}$, $\quad P(Z > z_2) = \frac{\alpha}{2} \quad \rightarrow \quad P(X \leq z_2) = 1 - \frac{\alpha}{2}$

Se sol utilitzar la nomenclatura $z_1 = x_{n,\frac{\alpha}{2}}$, $\quad z_2 = x_{n,1-\frac{\alpha}{2}}$

Manipulem:

$$z_1 \leq Z \leq z_2 \quad \equiv \quad z_1 \leq \frac{n\hat{S}^2}{\sigma^2} \leq z_2$$

$$\frac{n\hat{S}^2}{z_2} \leq \sigma^2 \leq \frac{n\hat{S}^2}{z_1}$$

IC per a σ^2 amb nivell de confiança $1-\alpha$: $\quad \left[\frac{n\hat{S}^2}{x_{n,1-\frac{\alpha}{2}}}, \frac{n\hat{S}^2}{x_{n,\frac{\alpha}{2}}} \right]$

IC per a la variància en una distribució normal

μ desconeguda i σ desconeguda

$X \approx \mathcal{N}(\mu, \sigma)$ amb μ desconeguda i σ desconeguda. Volem estimar σ^2

Estimador: la quasivariància $S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

és un estimador no esbiaixat de σ^2

V. a. que cal considerar: $Z = \dfrac{(n-1)S_c^2}{\sigma^2} = \chi^2_{n-1}$.

Procedint com abans:

IC per a σ^2 amb nivell de confiança $1-\alpha$: $\left[\dfrac{(n-1)S_c^2}{x_{n-1,1-\frac{\alpha}{2}}}, \quad \dfrac{(n-1)S_c^2}{x_{n-1,\frac{\alpha}{2}}}\right]$

Noteu que $Z = \dfrac{(n-1)S_c^2}{\sigma^2} = \dfrac{(n)S^2}{\sigma^2} = \chi^2_{n-1}$.

Resum

- IC per a μ
  - μ desconegut i σ conegut

$$\boxed{\bar{x} \pm z_{(1-\frac{\alpha}{2})} \frac{\sigma}{\sqrt{n}} \text{ amb } \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

  - μ desconegut i σ desconegut

$$\boxed{\bar{x} \pm t_{(n-1,1-\frac{\alpha}{2})} \frac{s}{\sqrt{n-1}} \text{ amb } s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- IC per a σ
  - μ conegut i σ desconegut

$$\boxed{\left[\frac{n\widehat{S}^2}{x_{(n,1-\frac{\alpha}{2})}}, \quad \frac{n\widehat{S}^2}{x_{(n,\frac{\alpha}{2})}} \right] \text{ amb } \widehat{S}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}$$

  - μ desconegut i σ desconegut

$$\boxed{\left[\frac{(n-1)S_c^2}{x_{(n-1,1-\frac{\alpha}{2})}}, \quad \frac{(n-1)S_c^2}{x_{(n-1,\frac{\alpha}{2})}} \right] \text{ amb } S_C^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Exemples

IC per a μ i σ

Exemple

S'ha mesurat el creixement de 5 plantes d'una determinada espècie en una setmana i s'han obtingut les dades següents: 24, 19, 22, 20 i 20 mm. Suposant una distribució normal de la variable que mesura el creixement, estimem la mitja i la desviació típica mitjançant intervals amb nivell de confiança del 90% i 99%.

$X =$ mesura del creixement d'una planta de l'espècie $\approx \mathcal{N}(\mu, \sigma)$

μ, σ desconeguts.

$(X_1, X_2, \ldots, X_n)$ m. a. s. $\bar{X} \approx \mathcal{N}(\mu, \sigma/\sqrt{n})$, $\quad n = 5$.

- IC per a μ :

Utilitzarem $\boxed{\bar{x} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}} \text{ amb } s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$

Determinem $t_{n-1,1-\alpha/2}$.

Tema 1. Teoria del Mostreig Estadístic

Francisco Guerrero Cortina

Intervals de confiança

Inferència estadística

Teoria elemental del mostreig

Estimació de paràmetres d'una distribució

Intervals de confiança

Exemples

IC per a μ

Exemple

- IC per a μ al 90%:
  $1-\alpha = 0.9$, $\alpha = 0.1$, $\alpha/2 = 0.05$, $1-\alpha/2 = 0.950$, $\quad t_{4,0.950} = 2.132.$

  $\bar{X} = 21 \quad S^2 = \frac{1}{5}\sum_{i=1}^{5}(X_i-\bar{X})^2 = 1.066 \Rightarrow S = 1.032$

  $\boxed{\mu \in [19.899, 22.001]}$

- IC per a μ al 99%: $1-\alpha = 0.99$, $\alpha = 0.01$

  Solament s'ha de recalcular $t_{n-1,1-\alpha/2}$
  $1-\alpha/2 = 0.995$, $\quad t_{4,0.995} = 4.604.$

  $\boxed{\mu \in [18.624, 23.3756]}$

Exemples

IC per a σ

Exemple

Utilitzarem $\boxed{\left[\frac{(n-1)S_c^2}{x_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S_c^2}{x_{n-1,\frac{\alpha}{2}}}\right] \text{ amb } S_C^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2}$

- IC per a σ al 90%: $1-\alpha = 0.90$, $\alpha = 0.1$
  $x_{n-1,1-\frac{\alpha}{2}} = x_{4,0.950} = 9.49$
  $x_{n-1,\frac{\alpha}{2}} = x_{4,0.05} = 0.711$
  $S_C^2 = 4$

$$\sigma^2 \in [1.686, 22.503] \quad \to \quad \sigma \in [1.298, 4.744]$$

- IC per a σ al 99%: $1-\alpha = 0.99$, $\alpha = 0.01$
  Recalculem
  $x_{n-1,1-\frac{\alpha}{2}} = x_{4,0.955} = 9.74$
  $x_{n-1,\frac{\alpha}{2}} = x_{4,0.005} = 0.21$

$$\sigma^2 \in [1.077, 77.299] \quad \to \quad \sigma \in [1.038, 8.792]$$

Tema 1. Teoria del Mostreig Estadístic (Part III)

Matemàtiques II. Grau de Química

Francisco Guerrero Cortina

Dept. Matemàtica Aplicada, Universitat de València

Curs 2012-2013

Tema 1. Teoria del Mostreig Estadístic (Part IV)

Matemàtiques II. Grau de Química

Francisco Guerrero Cortina

Dept. Matemàtica Aplicada, Universitat de València

Curs 2012-2013

Continguts

Continguts

Introducció

Les proves o contrastos d'hipòtesis es duen a terme per a inferir decisions que es refereixen a un paràmetre poblacional basant-se en mostres de la variable.

Exemple

Els científics recomanen que per a preveure l'escalfament global, la concentració de gasos d'efecte hivernacle no ha d'excedir les 350 parts per milió. Una organització de protecció del medi ambient vol determinar si el nivell mitjà, μ, de gasos d'efecte hivernacle en una regió compleix amb les pautes requerides, que estableixen un límit mï¿$\frac{1}{2}$xim de 350 parts per milió.

Per a açò es prendrà una mostra de mesuraments diaris d'aire per a decidir si se supera el límit, és a dir, si $\mu > 350$ o no. Per tant, l'organització vol trobar suport per a la hipòtesi $\mu > 350$, anomenada hipòtesi alternativa (H_1) i obtenir proves en la mostra que indiquen que la hipòtesi contrària, $\mu = 350$ (o $\mu \leq 350$), anomenada hipòtesi nul·la (H_0), és falsa.

Introducció

Exemple

Dit d'una altra manera, l'organització sotmetrà a judici la hipòtesi nul·la $\mu \leq 350$. Partirà de la seua innocència, suposant que és certa, és a dir, suposant que, en principi, no se superen els límits de presència de gasos d'efecte hivernacle, i només la rebutjarà en favor d'H_1 si hi ha proves evidents en les dades de la mostra per a fer-ho.

La decisió de rebutjar o no la hipòtesi nul·la en favor de l'alternativa haurà de basar-se en la informació que dóna la mostra, a través d'alguna mesura associada a aquesta, que s'anomena estadístic de contrast.

Per exemple, si es prenen 30 lectures d'aire i la mitjana mostral és molt més gran que 350, allò més lògic serà rebutjar la hipòtesi nul·la en favor de $\mu > 350$, però si la mitjana mostral és només lleugerament major que 350 o menor que 350, no hi haurà proves suficients per a rebutjar $\mu \leq 350$ en favor de $\mu > 350$.

Definicions

Contrast d'hipòtesi

Un contrast d'hipòtesi és una prova que es basa en les dades d'una mostra d'una variable aleatòria mitjançant la qual podem rebutjar una hipòtesi sobre un paràmetre de la població, anomenada hipòtesi nul·la (H_0), en favor d'una hipòtesi contrària, anomenada hipòtesi alternativa (H_1).

Estadístic de contrast

La prova es basa en una transformació de les dades de la mostra, la qual cosa s'anomena estadístic de contrast.

Regió de rebuig

Es rebutjarà la hipòtesi nul·la en favor de l'alternativa quan el valor de l'estadístic de contrast es situe en una determinada regió, anomenada regió de rebuig.

Tipus de contrast

La hipòtesi H_0 se sol expressar com una igualtat, del tipus $H_0 : \theta = \theta_0$, on θ és un paràmetre d'una població i θ_0 és un valor hipotètic per a aquest paràmetre.

Per la seua banda, H_1 pot tenir tenir dues formes:

- $H_1 : \theta > \theta_0$; en aquest cas es parla de contrast unilateral a la dreta o d'una cua a la dreta o d'un extrem a la dreta, o $H1 : \theta < \theta_0$; en aquest cas es parla de contrast unilateral a l'esquerra o d'una cua a l'esquerra o d'un extrem a l'esquerra.
- $H_1 : \theta \neq \theta_0$; en aquest cas es parla de contrast bilateral o de dues cues o de dos extrems.

Un dels aspectes més importants i que se sol prestar a més confusió es refereix a quina hipòtesi cal considerar H_0 i quina, H_1. Una regla pràctica per a fer-ho correctament pot ser la següent:

1. Si estem intentant provar una hipòtesi, aquesta ha de ser considerada la hipòtesi alternativa.
2. Per contra, si volem desacreditar una hipòtesi, hem d'incloure aquesta com a hipòtesi nul·la.

Resum

- Si el valor de l'estadístic de contrast per a les dades de la mostra cau en la regió de rebuig, podrem afirmar amb un determinat nivell de confiança que les dades de la mostra permeten rebutjar la hipòtesi nul·la en favor de l'alternativa.
- Si el valor de l'estadístic de contrast per a les dades de la mostra no cau en la regió de rebuig, no podrem afirmar amb el nivell de confiança exigit que les dades de la mostra permeten rebutjar la hipòtesi nul·la en favor de l'alternativa.

Continguts

Classificació dels errors

El contrast d'una hipòtesi estadística implica, per tant, una presa de decisió, a favor d'H_0 o en contra d'H_0 i en favor d'H_1. Açò implica que podem equivocar-nos en prendre la decisió de dues maneres.

S'anomena error tipus I o fals negatiu el fet de rebutjar la hipòtesi nul·la quan és certa, i la seua probabilitat es denota per α, anomenat nivell de significació.

S'anomena nivell de confiança la probabilitat d'acceptar la hipòtesi nul·la quan és certa, és a dir, $1 - \alpha$:

		Estat real	
		H_0	H_1
Decisió en el contrast	H_0	Decisió correcta	Error Tipus II
	H_1	Error Tipus I	Decisió correcta

Classificació dels errors

El contrast d'una hipòtesi estadística implica, per tant, una presa de decisió, a favor d'H_0 o en contra d'H_0 i en favor d'H_1. Açò implica que podem equivocar-nos en prendre la decisió de dues maneres.

S'anomena error tipus II o fals positiu el fet d'acceptar la hipòtesi nul·la quan és falsa, i la seua probabilitat es denota per β.

S'anomena potència la probabilitat de rebutjar la hipòtesi nul·la quan és falsa, és a dir, $1 - \beta$.

		Estat real	
		H_0	H_1
Decisió en el contrast	H_0	Decisió correcta	Error Tipus II
	H_1	Error Tipus I	Decisió correcta

Continguts

Introducció

- Una altra manera d'actuar és calcular l'estadístic de contrast i valorar com és d'extrem aquest valor sota la distribució en el mostreig de la hipòtesi nul·la.
- Si és més extrem que el nivell de significació desitjat, es rebutjarà la hipòtesi nul·la en favor de l'alternativa. Aquesta mesura de com d'extrem és el valor de l'estadístic s'anomena *p-valor*.

Definició de *p*-valor

De manera general, suposem que volem contrastar una hipòtesi estadística simple del tipus $H_0 : \theta = \theta_0$, enfront d'alguna de les alternatives següents: $H_1 : \theta \neq \theta_0$; $H_1 : \theta > \theta_0$ o $H_1 : \theta < \theta_0$. Suposem a més que el contrast es du a terme mitjançant un estadístic que notarem S, i que el valor de l'estadístic per a la mostra és s.

p-valor

El *p*-valor associat al contrast es defineix com el mínim nivell de significació amb el qual la hipòtesi nul·la seria rebutjada en favor de l'alternativa.

Propietats

Atès que normalment es tria com a nivell de significació màxim $\alpha = 0.05$, es té de la regla de decisió en un contrast amb aquest nivell de significació, donat el *p*-valor, seria la següent:

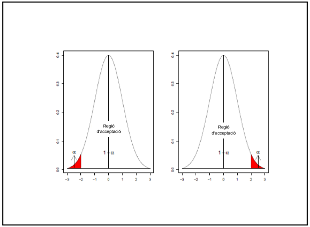
- Si $p < 0.05$, rebutgem H_0 en favor d'H_1 amb més d'un 95% de confiança.
- Si $p \geq 0.05$, no podem rebutjar H_0 en favor d'H_1 amb almenys un 95% de confiança.

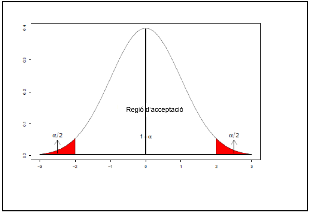
En resum, el *p*-valor permet utilitzar qualsevol altre nivell de significació, ja que si considerem un nivell de significació α:

- Si $p < \alpha$, rebutgem H_0 en favor d'H_1 amb més d'un $(1-\alpha) \times$ % de confiança.
- Si $p \geq \alpha$, no podem rebutjar H_0 en favor d'H_1 amb almenys un $(1-\alpha) \times$ % de confiança.

Càlcul del p-valor

Per a comprendre com es calcula el p-valor d'un contrast és necessari distingir entre contrastos unilaterals o d'una cua enfront de contrastos bilaterals o de dues cues.





- Els contrastos del tipus $H_0 : \theta = \theta_0$ enfront de $H_1 : \theta \neq \theta_0$ són contrastos bilaterals.
- Els contrastos del tipus $H_0 : \theta = \theta_0$ enfront de $H_1 : \theta > \theta_0$ o $H_1 : \theta < \theta_0$ són contrastos unilaterals.

Càlcul del p-valor

Per tant, tenint en compte la definició de p-valor, el càlcul d'aquest es fa de la manera següent:

- Si el contrast és unilateral a l'esquerra ($H_1 : \theta < \theta_0$),

$$p = P[S \leq s|_{H_0}].$$

- Si el contrast és unilateral a la dreta ($H_1 : \theta > \theta_0$),

$$p = P[S > s|_{H_0}].$$

- Si el contrast és bilateral ($H_1 : \theta \neq \theta_0$),

$$p = 2 \times \min\{P[S \leq s|_{H_0}], P[S > s|_{H_0}]\}.$$

Continguts

Introducció

Suposem que tenim una mostra $x_1, \ldots, x_n$ d'una variable aleatòria amb mitjana poblacional μ. Denotarem per $\overline{x}$ la mitjana mostral i s_{n-1}^2 la variància mostral.

El quadre següent inclou un resum del procediment per al contrast. En aquest, z_p és el valor d'una $N(0,1)$ de manera que $P[Z < z_p] = p$.

Tipus de prova	A l'esquerra	Bilateral	A la dreta
Hipòtesi	$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$	$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$	$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$
Estadístic	$z = \frac{\bar{x}-\mu_0}{s_c/\sqrt{n}}$		
Rebuig	$z < z_\alpha$	$\lvert z \rvert > z_{1-\alpha/2}$	$z > z_{1-\alpha}$
p-valor	$P[Z < z]$	$2P[Z > \lvert z \rvert]$	$P[Z > z]$
Supòsits	$n \geq 30$		

Exemple

Exemple

Suposem que una espècie comuna en la zona on s'enclava un jaciment, la *Bichus localis*, té una raó mitjana longitud/amplària de 9. Els arqueòlegs encarregats del jaciment han trobat 50 húmers fòssils, les dades dels quals apareixen en el quadre següent.

9.23	10.38	9.76	7.58	9.99	9.46	10.18	9.08	7.09	9.25
12.57	8.71	9.16	10.80	9.86	7.61	8.98	10.81	9.05	9.39
8.42	7.84	9.16	9.40	9.03	9.00	9.25	10.39	8.50	9.51
9.59	8.63	7.48	7.75	8.92	12.85	11.01	8.19	7.44	11.66
11.37	10.06	8.09	9.19	10.79	9.82	9.37	9.66	9.75	9.66

Tenen els arqueòlegs indicis suficients per a concloure que han descobert en el jaciment una espècie diferent de la *Bichus localis*?

En primer lloc, observem que no ens han especificat cap nivell de significació en l'enunciat. En aquest cas, allò habitual és considerar $\alpha = 0.05$. En cas que la decisió siga molt rellevant, triaríem un nivell més baix.

A continuació hem de plantejar les hipòtesis del contrast. La hipòtesi nul·la serà $H_0 : \mu = 9$, on per μ estem notant la mitjana de la raó longitud/amplària de l'húmer de l'espècie del jaciment. Com a hipòtesi alternativa ens plantegem que es tracte d'una altra espècie, és a dir, $H_1 : \mu \neq 9$. Es tracta, per tant, d'un contrast de dues cues.

Exemple

Exemple

Per a dur-ho a terme, hem de calcular en primer lloc l'estadístic de contrast. Aquest, al seu torn, requereix el càlcul de la mitjana i de la desviació típica mostral de les dades. Aquests valors són, respectivament, 9.414 i 1.239. Per tant,

$$z = \frac{9.414 - 9}{1.239/\sqrt{50}} = 2.363.$$

Per a saber si rebutgem la hipòtesi nul·la en favor de l'alternativa podem fer-ho de dues maneres.

1. Obtenint la regió de rebuig. Atès que $z_{1-0.05/2} = 1.96$, la regió de rebuig és $|z| > 1.96$. Veiem que, en efecte, $2.363 > 1.96$, per la qual cosa podem rebutjar la hipòtesi nul·la en favor de l'alternativa amb un 95% de confiança i concloure amb aquest nivell de confiança que es tracta d'una nova espècie.
2. Mitjançant el p-valor. Tenim que

$$p = 2 \times P[Z > |2.363|] = 0.004$$

Atès que és inferior al 5%, podem rebutjar la hipòtesi nul·la en favor de l'alternativa amb un 95% de confiança i concloure amb aquest nivell de confiança que la raó mitjana longitud/amplària dels húmers del jaciment és diferent de la del *Bichus localis*.

Introducció

Suposem que tenim una mostra $x_1, \ldots, x_n$ d'una variable aleatòria amb mitjana poblacional μ. Denotarem per $\overline{x}$ la mitjana mostral i s^2_{n-1} la variància mostral.

La principal diferència és que en no poder emprar el teorema central del límit, hem d'afegir com a hipòtesi la normalitat de les dades. En aquest cas, la distribució en el mostreig de l'estadístic ja no és normal, sinó t-student. El resum apareix en el quadre següent. En aquest, $t_{p;v}$ és el valor d'una t de Student amb v graus de llibertat de manera que $P[T_v < t_{p;v}] = p$.

Tipus de prova	A l'esquerra	Bilateral	A la dreta
Hipòtesi	$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$	$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$	$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$
Estadístic	$t = \frac{\bar{x}-\mu_0}{s_c/\sqrt{n}}$		
Rebuig	$t < t_{n-1;\alpha}$	$\|t\| > t_{n-1;1-\alpha/2}$	$t > t_{n-1;1-\alpha}$
p-valor	$P[T_{n-1} < t]$	$2P[T_{n-1} > \|t\|]$	$P[T_{n-1} > t]$
Supòsits	Distribució de probabilitat aproximadament normal		

Exemple

Exemple

El nivell màxim de concentració de benzè segons la Directiva Europea de Qualitat de l'Aire és de 5 micrograms per metre cúbic. Tenim unes dades corresponents a 20 cases on es va mesurar la concentració de benzè la qual va donar una mitjana mostral de 5.1 mircrograms per metre cúbic i una desviació típica mostral de 1.7.

Escollirem un nivell de significació de $\alpha = 0.05$. El contrast que es farà és la hipòtesi nul·la $H_0 : \mu = 5$ enfront de $H_1 : \mu > 5$.

L'estadístic de contrast és:

$$t = \frac{5.1 - 5}{1.7/\sqrt{20}} = 0.263$$

1. Si volem concloure amb la regió de rebuig, aquesta està formada pels valors $t > t_{0.95;19} = 1.729$, aleshores, atès que $0.263 < 1.729$, no podem afirmar amb un 95% de confiança que s'estiga incomplint la normativa.
2. El p-valor és encara més informatiu. El seu valor és $p = P[T_{19} > 0.263] = 0.398$, per la qual cosa hauríem d'arribar fins a quasi un 40% de significació per a rebutjar la hipòtesi nul·la en favor de l'alternativa afirmant que s'incompleix la normativa.

Continguts

Contrast amb mostres grans $(n_1, n_2 \geq 30)$

Siguen dues mostres $x_1, \ldots, x_{n_1}$ i $y_1, \ldots, y_{n_2}$ de variables aleatòries amb mitjanes poblacionals μ_1 i μ_2 i variàncies σ_1^2 i σ_2^2. Denotarem per $\overline{x}$ i $\overline{y}$ les mitjanes mostrals i $(s_c^1)^2$ i $(s_c^2)^2$ les variàncies mostrals.

El quadre següent inclou un resum del procediment per al contrast.

Tipus de prova	A l'esquerra	Bilateral	A la dreta
Hipòtesi	$\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 < D_0 \end{cases}$	$\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 \neq D_0 \end{cases}$	$\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 > D_0 \end{cases}$
Estadístic	$z = \dfrac{(\bar{x}-\bar{y})-D_0}{\sqrt{\dfrac{(s_c^1)^2}{n_1}+\dfrac{(s_c^2)^2}{n_2}}}$		
Rebuig	$z < z_\alpha$	$\lvert z \rvert > z_{1-\alpha/2}$	$z > z_{1-\alpha}$
Supòsits	$n_1, n_2 \geq 30$		

Exemple

Exemple

Imaginem que un enginyer inventa un nou mètode de producció amb el qual creu que poden reduir-se els temps de producció. Per a comprovar-ho, produeix 50 unitats amb el nou procés i 30, amb l'antic, i comptabilitza el temps (en segons) que es tarda a produir cada unitat.

Procés nou	Procés antic
$n_1 = 50$	$n_2 = 30$
$\bar{x} = 1255$	$\bar{y} = 1330$
$s_1 = 215$	$s_2 = 238$

Proporcionen aquestes mostres proves suficients per a concloure que la mitjana de temps de producció disminueix amb el procés nou ? Proveu-ho amb $\alpha = 0.05$.
Anomenem μ_1 el temps mitjà de producció sota el procés nou i μ_2 al temps mitjà de producció sota el procés antic. Ens demanen que contrastem $H_0 : \mu_1 = \mu_2$ enfront d'$H_1 : \mu_1 < \mu_2$ o, el que és el mateix, $H_1 : \mu_1 - \mu_2 < 0$. Es tracta, per tant, d'un test unilateral a l'esquerra.
L'estadístic de contrast és:

$$z = \frac{1255 - 1330}{\sqrt{\frac{215^2}{50} + \frac{238^2}{30}}} = -1.41$$

La regió de rebuig és $z < z_{0.05} = -1.65$. Atès que $z = -1.41$ no cau en aquesta regió, no podem rebutjar la hipòtesi nul·la en favor de l'alternativa amb $\alpha = 0.05$, és a dir, no tenim un 95% de confiança en el fet que el procés nou haja disminuït el temps mitjà de producció.

Contrast amb mostres menudes ($n_1 < 30$ o $n_2 < 30$) i variàncies iguals

Tipus de prova	A l'esquerra	Bilateral	A la dreta
Hipòtesi	$\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 < D_0 \end{cases}$	$\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 \neq D_0 \end{cases}$	$\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 > D_0 \end{cases}$
Estadístic	$t = \dfrac{(\bar{x}-\bar{y})-D_0}{\sqrt{s_p^2(\frac{1}{n_1}+\frac{1}{n_2})}}, \quad s_p^2 = \dfrac{(n_1-1)(s_c^1)^2+(n_2-1)(s_c^2)^2}{n_1+n_2-2}$		
Rebuig	$t < t_{n_1+n_2-2;\alpha}$	$\lvert t\rvert > t_{n_1+n_2-2;1-\alpha/2}$	$t > t_{n_1+n_2-2;1-\alpha}$
Supòsits	Variables normals amb $\sigma_1^2 = \sigma_2^2$		

Exemple

Exemple

Anem a considerar com a exemple el d'un enginyer que vol comparar dos equips de treball per a analitzar si es comporten de manera homogènia. Per fer-ho, du a terme una prova de destresa entre els treballadors d'ambdós equips: 13 de l'equip 1 i 15 de l'equip 2, les puntuacions dels quals apareixen en el quadre següent. Hi ha indicis suficients que hi haja diferències entre les puntuacions mitjanes dels dos equips? ($\alpha = 0.05$).

Equip 1	59	73	74	61	92	60	84	54	73	47	102	75	33		
Equip 2	71	63	40	34	38	48	60	75	47	41	44	86	53	68	39

Ens demanen que contrastem la igualtat de les mitjanes ($H_0 : \mu_1 = \mu_2$), enfront de l'alternativa $H_1 : \mu_1 \neq \mu_2$, per la qual cosa es tracta d'un contrast bilateral. En primer lloc, obtenim els estadístics mostrals d'ambdós equips. Les mitjanes són, respectivament, 68.2 i 53.8, mentre que les desviacions típiques mostrals són 18.6 i 15.8. Amb aquests valors podem calcular s_p^2

$$s_p^2 = \frac{12 \times 18.6 + 14 \times 15.8}{13 + 15 - 2} = 294.09$$

$$t = \frac{68.2 - 53.8}{\sqrt{294.09 \frac{1}{13} + \frac{1}{15}}} = 2.22$$

La regió de rebuig és $|t| < t_{26;0.975} = 2.055$. Atès que $t = 2.22$ cau en aquesta regió, podem rebutjar la hipòtesi nul·la en favor de l'alternativa amb un 95% de confiança.

Contrast amb mostres menudes ($n_1 < 30$ o $n_2 < 30$), variàncies distintes i mateixa grandària mostral

Tipus de prova	A l'esquerra	Bilateral	A la dreta
Hipòtesi	$\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 < D_0 \end{cases}$	$\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 \neq D_0 \end{cases}$	$\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 > D_0 \end{cases}$
Estadístic	$t = \dfrac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{1}{n}((s_c^1)^2+(s_c^2)^2)}}$		
Rebuig	$t < t_{2(n-1);\alpha}$	$\|t\| > t_{2(n-1);1-\alpha/2}$	$t > t_{2(n-1);1-\alpha}$
Supòsits	Les dues mostres tenen la mateixa grandària, $n_1 = n_2 = n$		

Contrast amb mostres menudes ($n_1 < 30$ o $n_2 < 30$), variàncies distintes i distinta grandària mostral

Tipus de prova	A l'esquerra	Bilateral	A la dreta
Hipòtesi	$\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 < D_0 \end{cases}$	$\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 \neq D_0 \end{cases}$	$\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 > D_0 \end{cases}$
Estadístic	$t = \dfrac{(\bar{x}-\bar{y})-D_0}{\sqrt{\dfrac{(s_c^1)^2}{n_1}+\dfrac{(s_c^2)^2}{n_2}}} \qquad v = \dfrac{\left(\dfrac{(s_c^1)^2}{n_1}+\dfrac{(s_c^2)^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{(s_c^1)^2}{n_1}\right)^2}{n_1-1}+\dfrac{\left(\dfrac{(s_c^2)^2}{n_2}\right)^2}{n_2-1}}$		
Rebuig	$t < t_{v;\alpha}$	$\|t\| > t_{v;1-\alpha/2}$	$t > t_{v;1-\alpha}$
Supòsits	Les distribucions són aproximadament normals		

Continguts

Contrast per a la proporció en una població

En aquesta ocasió, tenim una població on una proporció donada presenta una determinada característica, que denominem èxit, i la probabilitat del qual és p. Volem fer inferència sobre aquesta proporció. Per a açò seleccionem una mostra aleatòria simple de grandària n i comptabilitzem la proporció d'èxits en la mostra, $\widehat{p}$.

Tipus de prova	A l'esquerra	Bilateral	A la dreta
Hipòtesi	$\begin{cases} H_0 : p = p_0 \\ H_1 : p < p_0 \end{cases}$	$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$	$\begin{cases} H_0 : p = p_0 \\ H_1 : p > p_0 \end{cases}$
Estadístic	$z = \dfrac{\widehat{p}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$		
Rebuig	$z < z_\alpha$	$\lvert z\rvert > z_{1-\alpha/2}$	$z > z_{v;1-\alpha}$
Supòsits	$np_0\,,\ n(1-p_0) \geq 10$		

Contrast per a la proporció en una població

Exemple

Exemple

Anem a considerar un primer exemple relatiu a la relació entre el gènere i els accidents de trànsit. S'estima que el 60% dels conductors són homes. D'altra banda, un estudi fet sobre les dades de 120 accidents de trànsit mostra que el 70% d'aquests accidents van ser provocats per un home conductor. Podem, amb aquestes dades, confirmar que els homes són més perillosos al volant?

Si notem per p la proporció d'homes causants d'accidents de trànsit, la pregunta es respondrà afirmativament si aconseguim contrastar la hipòtesi $H1 : p > 0.6$. El valor de l'estadístic és

$$z = \frac{0.7 - 0.6}{\sqrt{\frac{0.6 \times 0.4}{120}}} = 2.236$$

Per la seua banda, la regió de rebuig seria $|z| > 1.96$ per a un $\alpha = 0.05$, ja que, en efecte, podem concloure que la proporció d'homes causants d'accidents és superior a la proporció d'homes conductors en general.

Continguts

Contrast per a la diferència de proporcions

En aquesta ocasió, partim de dues poblacions dins de les quals hi ha proporcions p_1 i p_2 d'individus amb la característica èxit. Pretenem comparar aquestes proporcions mitjançant la presa de mostres de grandària n_1 i n_2. Denotarem per $\widehat{p_1}$ i $\widehat{p_2}$ les proporcions d'èxits en les mostres

Tipus de prova	A l'esquerra	Bilateral	A la dreta
Hipòtesi	$\begin{cases} H_0 : p_1 - p_2 = D_0 \\ H_1 : p_1 - p_2 < D_0 \end{cases}$	$\begin{cases} H_0 : p_1 - p_2 = D_0 \\ H_1 : p_1 - p_2 \neq D_0 \end{cases}$	$\begin{cases} H_0 : p_1 - p_2 = D_0 \\ H_1 : p_1 - p_2 > D_0 \end{cases}$
Estadístic	$z = \dfrac{\widehat{p_1} - \widehat{p_2} - D_0}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, \qquad \widehat{p} = \dfrac{n_1\widehat{p_1}+n_2\widehat{p_2}}{n_1+n_2}$		
Rebuig	$z < z_\alpha$	$\|z\| > z_{1-\alpha/2}$	$z > z_{v;1-\alpha}$
Supòsits	Almenys 10 èxits i 10 fracassos		

Contrast per a la diferència de proporcions

Exemple

Exemple

Es van dur a terme proves d'alcoholèmia en 274 conductors implicats en accidents de trànsit amb ferits, dels quals, 88 van donar positiu. Per la seua banda, la Guàrdia Civil de Trànsit va dur a terme en la mateixa zona 1044 controls d'alcoholèmia a l'atzar, dels quals 15 van donar positiu.
El que la DGT vol demostrar és que l'alcohol és causant dels accidents de trànsit. No obstant açò, des del punt de vista estadístic només podem contrastar la hipòtesi que la proporció de positius en la prova d'alcoholèmia és major en el grup de conductors implicats en accidents de trànsit.
Notem per p_1 i p_2 les vertaderes proporcions en el grup d'implicats en accidents i en el grup de conductors no implicats. Se'ns demana contrastar $H_0 : p_1 = p_2$ enfront d'$H_1 : p_1 > p_2$. L'estadístic de contrast és

$$z = \frac{\frac{88}{274} - \frac{15}{1044}}{\frac{88+15}{274+1044}(1 - \frac{88+15}{274+1044})(\frac{1}{274} + \frac{1}{1044})} = 904.29$$

Està clar que el valor de l'estadístic és molt gran, sense necessitat de valorar la regió de rebuig, que seria $z > z_{0.95} = 1.645$, després podem rebutjar la hipòtesi nul·la en favor de l'alternativa amb, almenys, el 95% de confiança.

Continguts

Contrast per a la variància d'una població

De nou considerem que tenim una variable aleatòria X amb variància σ^2 i que prenem una mostra de grandària n, la variància mostral de la qual notem per s_c^2. Tractarem de fer inferència sobre σ^2. $\chi^2_{p;v}$ és el valor d'una χ^2 de v graus de llibertat.

Tipus de prova	A l'esquerra	Bilateral	A la dreta
Hipòtesi	$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{cases}$	$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases}$	$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{cases}$
Estadístic	$\chi^2 = \frac{(n-1)s_c^2}{\sigma_0^2}$		
Rebuig	$\chi^2 < \chi^2_{n-1;\alpha}$	ó $\begin{cases} \chi^2 < \chi^2_{n-1;\alpha/2} \\ \chi^2 > \chi^2_{n-1;1-\alpha/2} \end{cases}$	$\chi^2 > \chi^2_{n-1;1-\alpha}$
Supòsits	Distribució aproximadament normal		

Contrast per a la variància d'una població

Exemple

Exemple

L'empresa Sidel afirma que la seua màquina d'ompliment HEMA posseeix una desviació típica en l'ompliment de contenidors de 500ml de producte homogeni inferior a 0.8 gr. Anem a suposar que el supervisor de control de qualitat vol dur a terme una comprovació sobre aquest tema. Amb aquest fi, recopila una mostra de l'ompliment de 50 contenidors i obté una variància mostral de 0.6. Aquesta informació proporciona proves suficients que la desviació típica del seu procés d'ompliment és realment inferior a 0.8 gr?

Plantegem, en primer lloc, les hipòtesis del contrast. Se'ns demana que contrastem $H_0 : \sigma = 0.8$ o, equivalentment, $H_0 : \sigma^2 = 0.64$ enfront de l'alternativa $H_1 : \sigma^2 < 0.64$. Es tracta, per tant, d'un test unilateral a l'esquerra. L'estadístic de contrast és

$$\chi^2 = \frac{49 \times 0.6}{0.64} = 45.938.$$

Ara concloem a través de la regió de rebuig (triem $\alpha = 0.05$) que atès que $\chi^2_{9;0.05} = 33.930$ i $\chi^2 = 45.938 > \chi^2_{9;0.05} = 33.930$, no podem concloure amb almenys un 95% de confiança que, en efecte, la desviació típica de la quantitat d'ompliment és inferior a 0.8 gr.

Continguts

Contrast per al quocient de variàncies

Tenim dues mostres $x_1, \ldots, x_n$ i $y_1, \ldots, y_n$ dues variables aleatòries amb variàncies σ_1^2 i σ_2^2. Denotarem per $(s_c^1)^2$ i $(s_c^2)^2$ les variàncies mostrals.

Tipus de prova	A l'esquerra	Bilateral	A la dreta
Hipòtesi	$\begin{cases} H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1 \end{cases}$	$\begin{cases} H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \end{cases}$	$\begin{cases} H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1 \end{cases}$
Estadístic	$f = \frac{(s_c^1)^2}{(s_c^2)^2}$		
Rebuig	$f < f_{n_1-1;n_2-1;\alpha}$	ó $\begin{cases} f < f_{n_1-1;n_2-1;\alpha/2} \\ f > f_{n_1-1;n_2-1;1-\alpha/2} \end{cases}$	$f > f_{n_1-1;n_2-1;1-\alpha}$
Supòsits	Distribució aproximadament normal		

Contrast per al quocient de variàncies

Exemple

Exemple

Per a practicar sobre el contrast, considerem que s'han dut a terme 20 mesures de la duresa en l'escala Vickers d'acer amb un alt contingut en crom i 20 mesures més independents de la duresa d'una soldadura produïda sobre aquest metall.
Les desviacions estàndard de les mostres de duresa del metall i de duresa de la soldadura sobre aquest va ser de $12.06\mu HV$ i $11.41\mu HV$, respectivament. Podem suposar que les dureses corresponen a variables normals i independents. Podem concloure que la duresa del metall bàsic és més variable que la duresa mesurada en la soldadura?
Anomenarem la duresa sobre l'acer X i la duresa sobre la soldadura, Y . Se'ns demana que contrastem $H_0 : \sigma_X^2 = \sigma_Y^2$ enfront de l'alternativa $H_1 : \sigma_X^2 > \sigma_Y^2$ o, equivalentment, $H_1 : \frac{\sigma_X^2}{\sigma_Y^2} > 1$.
Es tracta, per tant, d'una prova unilateral a la dreta. L'estadístic de contrast és

$$f = \frac{12.06^2}{11.41^2} = 1.1172$$

Anem a prendre un nivell de significació de $\alpha = 0.05$. La regió crítica està delimitada pel valor $f_{19;19;0.95} = 2.168$. Atès que $f = 1.1172 < f_{19;19;0.95} = 2.168$, no podem concloure en el nivell de significació $\alpha = 0.05$ que la duresa del metall bàsic siga més variable que la duresa mesurada en la soldadura.

Tema 1. Teoria del Mostreig Estadístic (Part IV)

Matemàtiques II. Grau de Química

Francisco Guerrero Cortina

Dept. Matemàtica Aplicada, Universitat de València

Curs 2012-2013

Tema 2. Ajustament i Regressió

Matemàtiques II. Grau de Química

Francisco Guerrero Cortina

Dept. Matemàtica Aplicada, Universitat de València

Curs 2012-2013

Continguts

Continguts

Introducció

Un dels aspectes més rellevants que aborda l'Estadística es refereix a l'anàlisi de les relacions que es donen entre dues variables aleatòries.

L'anàlisi d'aquestes relacions està molt freqüentment lligat a l'anàlisi d'una variable, anomenada variable dependent (Y); i de l'efecte que sobre ella té una altra (o unes altres) variable(s), anomenada(es) variable(s) independent(s) (X), i permet respondre a dues qüestions bàsiques:

- És significativa la influència que té la variable independent sobre la variable dependent?
- Si, en efecte, aquesta relació és significativa, com és? i podem aprofitar aquesta relació per a predir valors de la variable dependent a partir de valors observats de la variable independent?

Exemple

Exemple

Hi ha un tipus de soldadura anomenada soldadura per fregament que consisteix en el fet que el frec entre dues peces provoca un escalfament que, al seu torn, produeix la soldadura entre ambdues. Suposem que fem un experiment sobre aquest tipus de soldadura fent rodar a una velocitat fixada per endavant (x, en m/mn) una peça i portant-la fins al repòs mitjançant el fregament amb una altra peça.

La calor generada per aquest fregament provoca una soldadura de pressió calenta la resistència de la qual (y) mesurem en *ksi*. Les dades de l'experiment es recullen en el quadre següent:

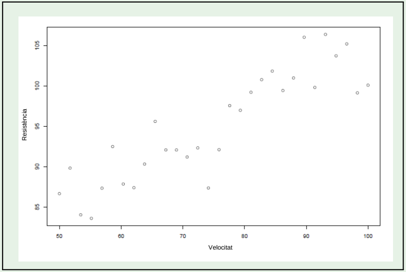
Velocitat	Resistència	Velocitat	Resistència
50.00	86.65	75.86	92.09
51.72	89.81	77.59	97.55
53.45	84.02	79.31	96.97
55.17	83.58	81.03	99.21
56.90	87.32	82.76	100.77
58.62	92.48	84.48	101.83
60.34	87.84	86.21	99.42
62.07	87.38	87.93	100.98
63.79	90.31	89.66	106.03
65.52	95.60	91.38	99.81
67.24	92.06	93.10	106.38
68.97	92.06	94.83	103.73
70.69	91.18	96.55	105.20
72.41	92.31	98.28	99.14
74.14	87.35	100.00	100.09

Exemple

Exemple

S'està tractant d'analitzar l'efecte que té la velocitat (variable independent) sobre la resistï¿$\frac{1}{2}$ncia de la soldadura (variable depenent). Afecta d'una manera rellevant? Si és així, com? Podríem ser capaços de predir la resistència de la soldadura coneguda la velocitat inicial que genera el fregament?

Si dibuixem les dades de x i y en uns eixos cartesians ja intuím que, en efecte, hi ha una relació latent entre les variables, que sembla ser de tipus lineal. Aquesta representació en els eixos cartesians s'anomena núvol de punts.



Regressió lineal simple

Definició

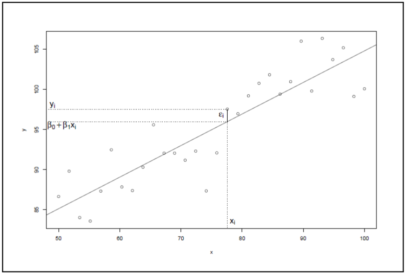
Un model de regressió lineal simple per a una variable, Y (variable dependent), donada una altra variable, X (variable independent), és un model matemàtic que permet obtenir una fòrmula capaç de relacionar Y amb X basat només en relacions lineals, del tipus

$$Y = \beta_0 + \beta_1 X$$

En aquesta expressió:

- Y representa la variable dependent, és a dir, aquella variable que volem estudiar en relació amb unes altres.
- X representa la variable independent, és a dir, aquelles que creiem que poden afectar en alguna mesura la variable dependent.

Recta de regressió



En la Figura podem observar la interpretació dels coeficients del model:

- β_0 és l'ordenada en l'origen del model, és a dir, el punt on la recta intercepta o talla l'eix y.
- β_1 representa el pendent de la recta i, per tant, pot interpretar-se com l'increment de la variable dependent per cada increment en una unitat de la variable independent.

Continguts

Mètode de mínims quadrats

Si volem obtenir el model de regressió lineal que millor s'ajuste a les dades de la mostra, haurem d'estimar els coeficients β_0 i β_1 del model. Per obtenir estimadors d'aquests coeficients, considerarem un nou mètode d'estimació, conegut com mètode de mínims quadrats.

Si tenim una mostra de valors de les variables independent i dependent,

$$(x_1, y_1), \ldots, (x_n, y_n),$$

cercarem valors estimats de β_0 i β_1; que denotarem per $\widehat{\beta}_0$ i $\widehat{\beta}_1$; de manera que en el model ajustat,

$$\widehat{y}_x = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

minimitze la suma dels quadrats dels errors observats.

Aleshores, $\widehat{y}_x$ pot interpretar-se de dues maneres:

- Com una predicció del valor que prendrà Y si $X = x$.
- Com una estimació del valor mitjà de Y quan $X = x$.

Recta de regressió per mÂº’inims quadrats

Segons el que acabem de veure, el que busquem es minimitzar la suma dels quadrats dels errors

$$SSE = \sum_{i=1}^{n} \left(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) \right)^2,$$

És a dir, busquem

$$\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = arg \left[\min_{\beta_0, \beta_1} SSE \right].$$

S’anomena recta de regressió per mínims quadrats (o simplement recta de regressió) de Y donada X en la línia que té la SSE més petita entre tots els models lineals.

Coeficients de la recta

La solució d'aquest problema de mínims s'obté de la manera següent: es deriva SSE respecte de $\widehat{\beta}_0$ i $\widehat{\beta}_1$, s'iguala a zero i s'obté:

$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \tag{1}$$

on

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}$$

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

Exemple

Exemple

Per a les dades sobre l'exemple sobre la resistència de la soldadura, calcularem i interpretarem la recta de regressió.

$$SS_{xy} = 2630.975 \qquad SS_{xx} = 6681.034$$

aleshores

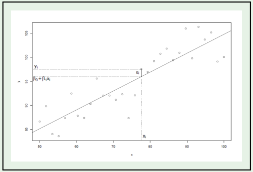
$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = 0.3938$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} = 65.4374,$$

així que la recta de regressió ajustada és

$$\widehat{y}_x = 65.4374 + 0.3938x$$

i està representada en la figura següent:



Continguts

Inferència sobre β_1

Una de les coses més importants és saber si l'efecte de la variable independent és o no significatiu per a la variable dependent. Açò és equivalent a contrastar si el coeficient β_1 és o no significativament diferent de zero.

Tipus de prova	A l'esquerra	Bilateral	A la dreta
Hipòtesi	$\begin{cases} H_0 : \beta_1 = b_1 \\ H_1 : \beta_1 < b_1 \end{cases}$	$\begin{cases} H_0 : \beta_1 = b_1 \\ H_1 : \beta_1 \neq b_1 \end{cases}$	$\begin{cases} H_0 : \beta_1 = b_1 \\ H_1 : \beta_1 > b_1 \end{cases}$
Estadístic	$t = \dfrac{\widehat{\beta}_1 - b_1}{\sqrt{s_e^2/S_{xx}}},\ s_e^2 = \dfrac{SS_{yy} - \widehat{\beta}_1 SS_{xy}}{n-2} = \dfrac{SSE}{n-2}$		
Rebuig	$t < t_{\alpha;n-2}$	$\lvert t\rvert > t_{1-\alpha/2;n-2}$	$t < t_{1-\alpha;n-2}$
p-valor	$P[T_{n-2} < t]$	$2P[T_{n-2} > \lvert t\rvert]$	$P[T_{n-1} > t]$

Si el que volem és contrastar si l'efecte de la variable independent és o no significatiu per a la variable dependent, el valor de b_1 serà zero.

Inferència sobre β_1

Exemple

Exemple

Per a les dades de l'exemple sobre la resistència de la soldadura, provarem si la velocitat és o no significativa ($\alpha = 0.05$):

$$\widehat{\beta}_1 = 0.3938$$

$$s_e^2 = \frac{SS_{yy} - \widehat{\beta}_1 SS_{xy}}{n-2} = 9.8345$$

$$t_{0.975;30-2} = 2.0484, \qquad t_{0.025;30-2} = -2.0484$$

$$t = \frac{0.3938 - 0}{\sqrt{9.8345/6681.034}} = 10.26$$

després, com calia esperar, podem afirmar a la llum de les dades i amb un 95% de confiança que l'efecte de la velocitat sobre la resistència és significatiu. El p-valor, de fet, és

$$p = 2P[T_{28} > 10.26] = 5.41 \times 10^{-11}$$

Inferència sobre β_1

Exemple

Exemple

Un enginyer químic està calibrant un espectròmetre per a mesurar la concentració de CO en mostres d'aire. Aquest calibratge implica que ha de comprovar que no hi ha diferències significatives entre la concentració vertadera de CO (x) i la concentració mesurada per l'espectròmetre (y).

Per a açò pren 11 mostres d'aire en les quals coneix la seua vertadera concentració de CO i les compara amb la concentració mesurada per l'espectròmetre. Les dades són les següents (les unitats són ppm):

x	0	10	20	30	40	50	60	70	80	90	100
y	1	12	20	29	38	48	61	68	79	91	97

Un primer pas en la comprovació que l'espectròmetre està ben calibrat implica contrastar que $\beta = 1$. Per a açò,

$$SS_{xx} = 11000; \quad SS_{yy} = 0.506.73; \quad SS_{xy} = 10740$$

$$\widehat{\beta}_1 = \frac{10460}{11000} = 0.976$$

$$s_e^2 = \frac{SS_{yy} - \widehat{\beta}_1 SS_{xy}}{n-2} = 2.286$$

Inferència sobre β_1

Exemple

Exemple

per tant,

$$t = \frac{0.951 - 1}{\sqrt{1.964/11000}} = -1.639$$

Donat que $t_{1-\frac{0.05}{2};11-2} = t_{0.975;9} = 2.262$ i $|-1.639| < 2.262$, no hi ha raons per a concloure que $\beta_1 \neq 1$. Així doncs, el model podria ser

$$y = \beta_0 + x,$$

encara que el millor seria que fos

$$y = x,$$

És a dir, que el que mesure l'espectròmetre coincidisca amb la quantitat real de CO en l'aire. Açò ocorreria si $\beta_0 = 0$.

Inferència sobre β_0

Aquest últim exemple posa de manifest que també pot tenir interés dur a terme contrastos sobre el valor de β_0. Per a açò, el quadre següent descriu el procediment d'un contrast d'aquest tipus.

Tipus de prova	A l'esquerra	Bilateral	A la dreta
Hipòtesi	$\begin{cases} H_0 : \beta_0 = b_0 \\ H_1 : \beta_0 < b_0 \end{cases}$	$\begin{cases} H_0 : \beta_0 = b_0 \\ H_1 : \beta_0 \neq b_0 \end{cases}$	$\begin{cases} H_0 : \beta_0 = b_0 \\ H_1 : \beta_0 > b_0 \end{cases}$
Estadístic	$t = \frac{\widehat{\beta}_0 - b_0}{\sqrt{s_e^2\left(\frac{1}{n} + \frac{x^2}{SS_{xx}}\right)}}, \; s_e^2 = \frac{SS_{yy} - \widehat{\beta}_1 SS_{xy}}{n-2} = \frac{SSE}{n-2}$		
Rebuig	$t < t_{\alpha;n-2}$	$\lvert t\rvert > t_{1-\alpha/2;n-2}$	$t < t_{1-\alpha;n-2}$
p-valor	$P[T_{n-2} < t]$	$2P[T_{n-2} > \lvert t\rvert]$	$P[T_{n-1} > t]$

Podria ser d'interés un contrast conjunt sobre β_0 i β_1. L'única cosa que podríem fer en un context com el nostre és dur a terme sengles contrastos sobre β_0 i β_1 per separat, tenint en compte el nivell de significació d'ambdós contrastos.

Inferència sobre β_0

Exemple

Exemple

En l'exemple anterior, contrastarem si, en efecte, $\beta_0 = 0$, la qual cosa equivaldrà a concloure que no hi ha raons per a pensar que l'espectròmetre està mal calibrat. Per a açò,

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} = 0.636$$

per tant,

$$t = \frac{0.636 - 0}{\sqrt{2.286(\frac{1}{11} + \frac{50^2}{11000})}} = 0.746$$

Com que $0.746 < t_{0.975;9} = 2.261$ tampoc tenim raons per a pensar que $\beta_0 = 0$ amb un 95% de confiança.

En resum, no hi ha raons per a pensar que l'espectometre està mal calibrat.

Continguts

Definició

$\widehat{\beta}_1$ mesura en certa manera la relació que hi ha entre la variable dependent i la variable independent. No obstant açò, és una mesura subjecta a l'escala de les variables X i Y.

El coeficient de correlació lineal ofereix una mesura quantitativa de la fortalesa de la relació lineal entre X i Y en la mostra, però que a diferència de $\widehat{\beta}_1$ sempre va entre -1 i 1.

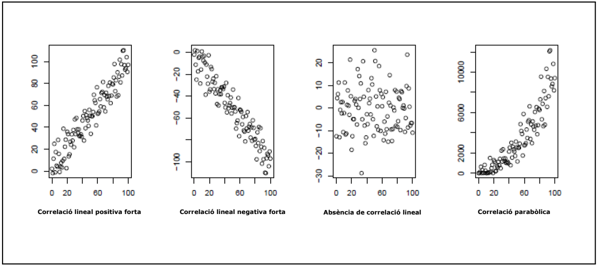
Donada una mostra de valors de dues variables $(x_1, y_1), \ldots, (x_n, y_n)$, el coeficient de correlació lineal mostral r es defineix com

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{\sqrt{SS_{xx}}}{\sqrt{SS_{yy}}}\widehat{\beta}_1.$$

Propietats

La interpretació del valor de r és la següent:

- r proper o igual a 0 implica poca o cap relació lineal entre X i Y.
- Com més s'acoste a 1 o -1, més forta serà la relació lineal entre X i Y.
- Si $r = \pm 1$, tots els punts cauran exactament en la recta de regressió.
- Un valor positiu de r implica que Y tendeix a augmentar quan X augmenta, i aquesta tendència és més acusada com més proper està r d'1.
- Un valor negatiu de r implica que Y disminueix quan X augmenta, i aquesta tendència és més acusada com més proper està r de -1.



Exemple

Exemple

Per a les dades de l'exemple sobre la resistència de la soldadura, calculem r i interpretem-ho. Sabem que

$$SS_{xy} = 2630.975$$
$$SS_{xx} = 6681.034$$
$$SS_{yy} = \sum_{i=1}^{29} y_i^2 - 29\overline{y}^2 = 1311.511$$

Aleshores,

$$r = \frac{2630.975}{\sqrt{6681.034 \times 1131.511}} = 0.8888.$$

Per tant, la resistència de la soldadura i la velocitat que genera el fregament tenen una correlació important per a aquesta mostra de 30 peces soldades, la qual cosa implica que hi ha una relació lineal positiva entre aquestes variables.

Continguts

Error estàndard de l'estimació

Definició

Es defineix l'error estàndard de l'estimació quan es té el valor $X = x$ com

$$s_e^2 = \frac{\sum_i (y_i - \widehat{y}_x)^2}{n-2} = \frac{\sum_i (y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x))^2}{n-2} = \frac{SS_{yy} - \widehat{\beta}_1 SS_{xy}}{n-2}.$$

Com més gran siga aquesta quantitat, pitjors són les prediccions de la recta de regressió.

Interval de confiança per al valor mitjà de Y

Podem garantir amb un $(1-\alpha)\times 100\%$ de confiança que quan $X = x$, el valor mitjà d'Y es troba en l'interval

$$\left(\widehat{y}_x - t_{1-\alpha/2;n-2} \times s_e\sqrt{\frac{1}{n}+\frac{(x-\overline{x})^2}{SS_{xx}}}\,, \widehat{y}_x + t_{1-\alpha/2;n-2} \times s_e\sqrt{\frac{1}{n}+\frac{(x-\overline{x})^2}{SS_{xx}}}\right)$$

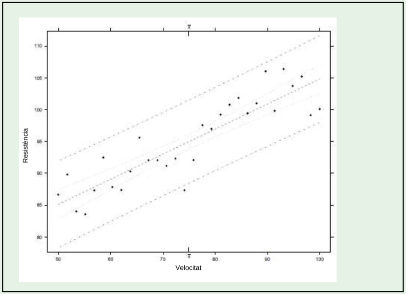
Interval de confiança per al valor Y

Podem garantir amb un $(1-\alpha)\times 100\%$ de confiança que quan $X = x$, el valor Y es troba en l'interval

$$\left(\widehat{y}_x - t_{1-\alpha/2;n-2}\times s_e\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{SS_{xx}}}, \widehat{y}_x + t_{1-\alpha/2;n-2}\times s_e\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{SS_{xx}}}\right)$$

Exemple

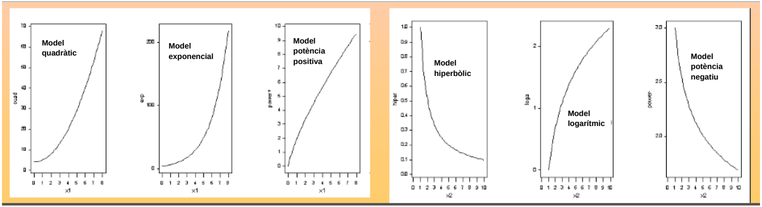
Exemple



En la figura apareix la recta de regressió per a les dades de l'exemple sobre la soldadura juntament amb línies que contenen els intervals de confiança al 95% per a les prediccions i les estimacions associades als diferents valors de X.

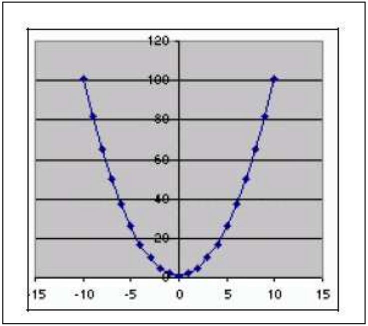
Regressió no lineal

Suposem que, en representar gràficament les variables, s'obté una clara relació entre aquestes, però clarament no lineal.
Per tant, s'haurà de cercar la funció que ha de descriure la dependència entre ambdues variables. Vorem ací les més utilitzades: les funcions parabòlica, hiperbòlica, exponencial i potencial.



Continguts

Regressió parabòlica



En molts casos, és una funció de segon grau la que s'ajusta bastant a la situació real donada. L'expressió general d'un polinomi de segon grau és:

$$Y = a + bX + cX^2$$

on a, b i c són els paràmetres.

Regressió parabòlica

El problema consisteix, per tant, a determinar aquests paràmetres per a una distribució donada. Açò es troba fent que la suma dels quadrats de les desviacions pel que fa a la corba de regressió siga mínima:

$$D = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$$

Per a trobar els valors de a, b i c que que fan mínima l'expressió anterior, s'igualaran les derivades parcials de D pel que fa a aquests paràmetres a zero i es resoldrà el sistema resultant.

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4$$

Exemple

Exemple

Tenim les dades següents:

	X	Y	X^2	X^3	X^4	XY	X^2Y
	1	1.25	1	1	1	1.25	1.25
	2	5	4	8	16	10	20
	3	11.25	9	27	81	33.75	101.5
	4	20	16	64	256	80	320
	5	30.5	25	125	625	152.5	762.5
$\sum$	15	68	55	225	979	277.5	1205

$$\begin{cases} \sum_{i=1}^n y_i = na + b\sum_{i=1}^n x_i + c\sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i y_i = a\sum_{i=1}^n x_i + b\sum_{i=1}^n x_i^2 + c\sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 y_i = a\sum_{i=1}^n x_i^2 + b\sum_{i=1}^n x_i^3 + c\sum_{i=1}^n x_i^4 \end{cases} \Rightarrow \begin{cases} 68 = 5a + 15b + 55c \\ 277.5 = 15a + 55b + 225c \\ 1205 = 55a + 225b + 979c \end{cases}$$

Resolent aquest sistema s'obté: $a = -0.47$, $b = 0.51$, $c = 1.14$, per tant,

$$Y = -0.47 + 0.51X + 1.14X^2$$

és la paràbola que aproxima les dades.

Tema 2. Ajustament i Regressió

Francisco Guerrero Cortina

Regressió lineal simple
Introducció
Estimació dels coeficients del model per mínims quadrats
Inferències sobre el model
El coeficient de correlació lineal
Predicció i estimació a partir del model

Regressió no lineal
Regressió parabòlica
Regressió hiperbòlica
Regressió potencial
Regressió exponencial

Continguts

Regressió hiperbòlica

- Es tracta d'ajustar per un model de la forma $y = a + b\frac{1}{x}$. Aleshores el valor de y no és lineal amb els valor de x, però sí amb els valors d'$1/x$.
- En aquest cas, ajustar les observacions (x_i, y_j) es redueix a ajustar uns nous valors (z_i, y_j) per a la recta $y = a + bz$, fent prèviament el canvi de variable $z = \frac{1}{x}$ i calculant directament a i b per mínims quadrats (eq 1).

Continguts

Regressió potencial

- La forma general d'aquest ajust és $y = ax^b$, la qual es pot reduir al cas lineal prenent logaritmes:

$$\log y = \log a + b \log x$$

- Per tant, ajustarem una recta al conjunt (z_i, u_j) on $z_i = \log x_i$ i $u_j = \log y_j$. Les equacions (eq 1) proporcionen les variables z i u els valors de $\log a$ i b.

Exemple

Exemple

$y = ax^b \Rightarrow \log y = \log a + b \log x \Rightarrow V = A + bU$

	X	Y	$U = \log x$	$V = \log y$	U^2	UV
	1	1.25	0	0.2231	0	0
	2	5	0.6931	1.6094	0.4803	1.1156
	3	11.25	1.0986	2.4203	1.2069	2.6590
	4	20	1.3863	2.9957	1.9215	4.1530
	5	30.5	1.6094	3.4177	2.5901	5.5006
$\sum$	15	68	4.7875	10.666	6.1988	13.428
$1/5\sum$	3	13.6	0.9575	2.1332	1.2397	2.6856

$$b = \frac{S_{UV}}{S_U^2} = \frac{1/5\sum_{i=1}^{n} UV - \overline{U}\,\overline{V}}{1/5\sum_{i=1}^{n} U^2 - \overline{U}^2} = \frac{2.6856 - 0.9575 \times 2.1332}{1.2397 - 0.9575^2} = 1.9902$$

$$A = \overline{V} - b\overline{U} = 2.1332 - 1.9902 \times 0.9575 = 0.2277$$

Es desfà el canvi efectuat

$$a = anti \log A = anti \log 0.2277 = 1.2557$$

Per tant,

$$y = 1.2557x^{1.9902}$$

Continguts

Regressió exponencial

- Tindrem en aquest cas $y = ab^x$, la qual es pot reduir al cas lineal prenent logaritmes:
$$\log y = \log a + x \log b$$

- Ara resulta lineal la relació entre $\log y$ i x, per tant, ajustarem una recta a les observacions (x_i, u_j) on $u_j = \log y_j$. Les equacions (eq 1) proporcionen les variables $\log a$ i $\log b$.

Exemple

Exemple

$y = ab^x \Rightarrow \log y = \log a + x \log b \Rightarrow V = A + BX$

	X	Y	$V = \log y$	X^2	XV
	1	1.25	0.2231	1	0.2231
	2	5	1.6094	4	3.2188
	3	11.25	2.4203	9	7.2609
	4	20	2.9957	16	11.983
	5	30.5	3.4177	25	17.088
$\sum$	15	68	10.666	55	39.774
$1/5\sum$	3	13.6	2.1332	11	7.9548

$$B = \frac{S_{XV}}{S_X^2} = \frac{1/5\sum_{i=1}^{n} XV - \overline{X}\,\overline{V}}{1/5\sum_{i=1}^{n} X^2 - \overline{X}^2} = \frac{7.9548 - 2.1332 \times 3}{11 - 3^2} = 0.7776$$

$$A = \overline{V} - B\overline{X} = 2.1332 - 0.7776 \times 3 = -0.1996$$

Es desfà el canvi efectuat

$$a = anti \log A = anti \log 0.1996 = 0.819$$

$$b = anti \log B = anti \log 0.7776 = 2.176$$

Per tant,

$$y = 0.819\ 2.176^x$$

Tema 2. Ajustament i Regressió

Matemàtiques II. Grau de Química

Francisco Guerrero Cortina

Dept. Matemàtica Aplicada, Universitat de València

Curs 2012-2013

Tema 3. Interpolació polinòmica

Matemàtiques II. Grau de Química

Francisco Guerrero Cortina

Dept. Matemàtica Aplicada, Universitat de València

Curs 2012-2013

Continguts

Continguts

Introducció

Siguen (x_k, y_k) $(k = 0, \ldots, m)$ parelles de valors reals (punts del pla) donats de manera que $x_k \neq x_i$, si $k \neq i$. Ens podem preguntar si hi ha alguna funció p, d'un tipus predeterminat, tal que $p(x_k) = y_k$ $(k = 0, \ldots, m)$.

La qüestió anterior s'anomena problema d'interpolació i al procés de recerca de la funciÂº'o p se li dóna el nom d'interpolació.

- La resolució d'aquest problema és d'utilitat en moltes situacions, en particular quan la procedència dels punts (x_k, y_k) $(k = 0, \ldots, m)$ és experimental.
- Suposem que coneguda experimentalment la resposta y_k obtinguda sota condicions x_k, ens interesa trobar el resultat y que obtindríem en prendre condicions x no experimentades.
- Podem pensar que els punts donats formen part de la gràfica d'una funció f que voldríem conèixer almenys aproximadament i de la qual únicament sabem que $f(x_k) = y_k$ $(k = 0, \ldots, m)$.

Exemple

Exemple

El valor de l'acceleració de la gravetat en un punt de la superfície terrestre, al nivell de la mar, depén de la latitud. Experimentalment s'ha comprovat la correspondència donada en la Taula 1

Latitud (en graus)	Acceleració gravetat (en m/s^2)
0	9.780350
30	9.793238
45	9.806154
60	9.819099
90	9.832072

Table: 1.- Mesures de la gravetat a diverses latituds

El problema consisteix a trobar el valor de la gravetat a Barcelona, que està situada a una latitud de $41^o\,25'$.

Resum

- De quin tipus (polinòmica, trigonomètrica, racional, etc.) ha de ser la funció p buscada?
- Elegit el conjunt de funcions, del qual hem d'extraure p, existeix la funció buscada? I, si existeix, és única?
- És la funció p una bona aproximació de la funció f en els punts en què no coincideixen?

Continguts

Interpolació polinòmica

Interpolació polinòmica

Donada una taula de $m+1$ punts d'interpolació (x_k, y_k) $(k = 0, \ldots, m)$ amb $x_k \neq x_i$, si $k \neq i$, anomenarem interpolació polinòmica la determinació d'un polinomi $p(x)$ de grau més menut o igual que N, tal que

$$p(x_k) = f_k \quad (k = 0, \ldots, m).$$

Continguts

Existència i unicitat del polinomi interpolador

Existeix un únic polinomi $p_m(x)$ de grau més menut o igual que $N = m$, tal que $p_m(x) = f_k$ $(k = 0, \ldots, m)$. El polinomi $p_m(x)$ rep el nom de polinomi interpolador de f en les abscisses x_k $(k = 0, \ldots, m)$.

Per veure això, considerem un polinomi $p_m(x) = a_0 + a_1 x + \cdots + a_m x^m$. Quan imposem les condicions d'interpolació, obtenim un sistema lineal de $m + 1$ equacions en les $m + 1$ incògnites $a_0, a_1, \ldots, a_m$:

$$\begin{cases} a_0 + a_1 x_0 + \cdots + a_m x_0^m & = f_0 \\ a_0 + a_1 x_1 + \cdots + a_m x_1^m & = f_1 \\ \cdots & \cdots \\ a_0 + a_1 x_m + \cdots + a_m x_m^m & = f_m \end{cases}$$

Existència i unicitat del polinomi interpolador

El determinant de la matriu del sistema s'anomena determinant de Vandermonde i té la forma

$$\begin{vmatrix} 1 & x_0 & \cdots & x_0^m \\ 1 & x_1 & \cdots & x_1^m \\ . & . & \cdots & . \\ . & . & \cdots & . \\ 1 & x_m & \cdots & x_m^m \end{vmatrix} = \prod_{k>i}(x_k - x_i)$$

aquest determinant és diferent de 0 ja que $x_i \neq x_k$, si $i \neq k$. Aleshores, el sistema plantejat és compatible i determinat i, per tant, hi ha una única solució: $a_0, a_1, \ldots, a_m$, coeficients de l'únic polinomi interpolador.

Si bé el procés que s'acaba d'escriure és el primer mètode en què es pensa quan s'intenta trobar $p_m(x)$, resulta excessivament laboriós quan m no és petit.

Continguts

Fòrmula d'interpolació de Lagrange

Es pren com a expressió del polinomi interpolador la fòrmula d'interpolació de Lagrange:

$$p_m(x) = \sum_{j=0}^{m} f_k l_k, \qquad l_k(x) = \frac{\prod_{i \neq k}(x - x_i)}{\prod_{i \neq k}(x_k - x_i)} \quad (k = 0, \ldots, m)$$

Els polinomis $l_k(x)$ reben el nom de polinomis de Lagrange. Observeu que el grau de $l_k(x)$ és igual a m i que $l_k(x_i) = \delta_{ik}$ $(i, k = 0, \ldots, m)$; per tant, $p(x_i) = f_i$ $(i = 0, \ldots, m)$, com es volia.

Exemple

Exemple

Siga una funció $f(x)$ desconeguda, que pren els valors 0, -3 i 1 en els punts 1, 2 i 4 respectivament. Trobar un polinomi d'interpolació.

Es té: $x_0 = 1, x_1 = 2, x_2 = 4$ i $f(x_0) = 0, f(x_1) = -3, f(x_2) = 1$

$$\begin{aligned}
p_2(x) &= \sum_{k=0}^{2} f(x_k) \cdot \prod_{j\neq k}^{2} \frac{(x-x_j)}{(x_k-x_j)} = \\
&= f(x_0)\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + f(x_1)\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + f(x_2)\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \\
&= f(1)\frac{(x-2)(x-4)}{(1-2)(1-4)} + f(2)\frac{(x-1)(x-4)}{(2-1)(2-4)} + f(4)\frac{(x-1)(x-2)}{(4-1)(4-2)} = \\
&= \frac{0}{3}(x-2)(x-4) + \frac{3}{2}(x-1)(x-4) + \frac{1}{6}(x-1)(x-2)
\end{aligned}$$

Resultant:

$$p_2(x) = \frac{5}{3}x^2 - 8x + \frac{19}{3}$$

Es comprova facilment que

$$p_2(1) = f(1) = 0, \quad p_2(2) = f(2) = -3, \quad p_2(4) = f(4) = 1$$

Fòrmula d'interpolació de Lagrange

Expressió de l'error

Si la funció que cal interpolar, $f(x)$, és almenys $m+1$ vegades derivable en $[a, b]$, s'obté l'error d'interpolació mitjançant l'expressió següent:

$$E(x) = f(x) - p_m(x) = \frac{f^{m+1}(\xi)}{(n-1)!}\prod_{i=0}^{n}(x - x_i)$$

on $\xi \in [a, b]$.

Aleshores, una fita per a l'error seria:

$$|E(x)| = |\frac{f^{m+1}(\xi)}{(n-1)!}\prod_{i=0}^{n}(x - x_i)| \leq \frac{M}{(n-1)!}\prod_{i=0}^{n}|(x - x_i)|$$

on $M \geq sup\{|f^{m+1}(t)|, \quad t \in [a, b]\}$.

Exemple

Exemple

Siga una funció $f(x) = 2xe^{-(4x+2)}$, definida en $[0.2, 1]$.
(a) Calcular el polinomi de Lagrange $p(x)$ que interpola $f(x)$ en els punts 0.2 i 1.0.
(b) Obtenir l'expressió de l'error d'interpolació.

Es té: $x_0 = 0.2, x_1 = 1$ i $f(x_0) = 0.4, f(x_1) = 2e^{-6}$

$$\begin{aligned} p_1(x) &= \sum_{k=0}^{1} f(x_k) \cdot \prod_{j\neq k}^{1} \frac{(x - x_j)}{(x_k - x_j)} = f(x_0)\frac{(x - x_1)}{(x_0 - x_1)} + f(x_1)\frac{(x - x_0)}{(x_1 - x_0)} = \\ &= f(0.2)\frac{1 - x}{0.8} + f(1)\frac{x - 0.2}{0.8} = (0.4)e^{-2.8}\frac{1 - x}{0.8} + 2e^{-6}\frac{x - 0.2}{0.8} = -0.0242x + 0.0292 \end{aligned}$$

Per a obtenir

$$f'(x) = (2 - 8x)e^{-(4x+2)} \qquad f''(x) = (-16 + 32x)e^{-(4x+2)}$$

Aleshores,

$$E(x) = \frac{(-16 + 32\xi)e^{-(4\xi+2)}}{2!}(x-0.2)(x-1) = \frac{(-16 + 32\xi)e^{-(4\xi+2)}}{2!}(x^2-1.2x+0.2)$$

Continguts

Introducció

Ambdós mètodes obtenen el polinomi interpolador construint una successió de polinomis de graus creixents que van interpolant cada vegada en més abscisses. Es basen en el principi següent:

Siga $C = \{x_0, x_1, \ldots, x_m\}$ i siga $S \subset C$ i $T \subset C$ tal que S i T tenen el mateix nombre d'elements i que aquests coincideixen llevat de dos (això és, existeixen $x_j \in S$ i $x_i \in T$ tals que $x_i \notin T$ i $x_j \notin S$).

Suposem coneguts els polinomis p_S i p_T tals que $p_S(x_k) = f_k$ (si $x_k \in S$) i $p_T(x_k) = f_k$ (si $x_k \in T$). Llavors es pot construir un nou polinomi $p_{S \cup T}$ tal que $p_{S\cup T}(x_k) = f_k$ (si $x_k \in S \cup T$), de la manera següent:

$$p_{S\cup T}(x) = \frac{(x_j - x)p_S(x) - (x_i - x)p_T(x)}{x_j - x_i}.$$

Els dos mètodes que estudiem construeixen el polinomi interpolador gràcies a successives aplicacions de la fòrmula anterior, però seguint un esquema diferent.

Mètode d'Aitken

Es construeixen polinomis $p_{i,j}(x)$ $(i = j, \ldots, m)$ de grau més petit o igual que j, que interpolen en les abscisses $\{x_0, x_1, \ldots, x_{j-1}, x_i\}$ $(j = 0, \ldots, m)$; s'arriba al polinomi $p_m(x) \equiv p_{m,m}(x)$, de grau menor o igual que m, que interpolarà en $\{x_0, x_1, \ldots, x_m\}$:

$$
\begin{aligned}
p_{i,0} &= f_i \qquad (i = 0, \ldots, m); \\
p_{i,j+1} &= \frac{(x_i - x)p_{j,j}(x) - (x_j - x)p_{i,j}(x)}{x_i - x_j} \quad (i = j+1, \ldots, m) \quad (j = 0, \ldots, m-1) \\
p_{i,j+1} &= \frac{1}{x_i - x_j}\begin{vmatrix} p_{j,j} & x_j - x \\ p_{i,j} & x_i - x \end{vmatrix} \quad (i = j+1, \ldots, m) \quad (j = 0, \ldots, m-1)
\end{aligned}
$$

Mètode d'Aitken

Esquema per a $m = 3$

Esquema per al mètode d'Aitken amb $m = 3$

$$
\begin{array}{llll}
p_{0,0}(x) & & & \\
p_{1,0}(x) & \rightarrow p_{1,1}(x) & & \\
p_{2,0}(x) & \rightarrow p_{2,1}(x) & \rightarrow p_{2,2}(x) & \\
p_{3,0}(x) & \rightarrow p_{3,1}(x) & \rightarrow p_{3,2}(x) & \rightarrow p_{3,3}(x)
\end{array}
$$

Mètode d'Aitken

Exemple

Exemple

Trobar una aproximació de $\sqrt{1.5}$ utilitzant el polinomi d'interpolació d'Aitken de grau 3, per als valors següents:

$x_0 = 1, x_1 = 2, x_2 = 3, x_3 = 4,$
$f(x_0) = 1, f(x_1) = 1.414213, f(x_2) = 1.732050, f(x_3) = 2$

Utilitzant les fòrmules, tenim:

$$p_{1,1} = \frac{1}{x_1 - x_0}\begin{vmatrix} p_{0,0} & x_0 - x \\ p_{1,0} & x_1 - x \end{vmatrix} = \frac{1}{2-1}\begin{vmatrix} 1 & 1\text{-}1.5 \\ 1.414213 & 2\text{-}1.5 \end{vmatrix} = 1.20710678$$

$$p_{2,1} = \frac{1}{x_2 - x_0}\begin{vmatrix} p_{0,0} & x_0 - x \\ p_{2,0} & x_2 - x \end{vmatrix} = \frac{1}{3+1}\begin{vmatrix} 1 & 1\text{-}1.5 \\ 1.73205081 & 3\text{-}1.5 \end{vmatrix} = 1.1830127$$

$$p_{3,1} = \frac{1}{x_3 - x_0}\begin{vmatrix} p_{0,0} & x_0 - x \\ p_{3,0} & x_3 - x \end{vmatrix} = \frac{1}{4-1}\begin{vmatrix} 1 & 1\text{-}1.5 \\ 2 & 4\text{-}1.5 \end{vmatrix} = 1.166667$$

Mètode d'Aitken

Exemple

Exemple

$$p_{2,2} = \frac{1}{x_2 - x_1}\begin{vmatrix} p_{1,1} & x_1 - x \\ p_{2,1} & x_2 - x \end{vmatrix} = \frac{1}{3-2}\begin{vmatrix} 1.20710678 & 2\text{-}1.5 \\ 1.1830127 & 3\text{-}1.5 \end{vmatrix} = 1.22205934$$

$$p_{3,2} = \frac{1}{x_3 - x_1}\begin{vmatrix} p_{1,1} & x_1 - x \\ p_{3,1} & x_3 - x \end{vmatrix} = \frac{1}{4-2}\begin{vmatrix} 1.20719678 & 2\text{-}1.5 \\ 1.16666667 & 3\text{-}1.5 \end{vmatrix} = 1.21721681$$

$$p_{3,3} = \frac{1}{x_3 - x_2}\begin{vmatrix} p_{2,2} & x_2 - x \\ p_{3,2} & x_3 - x \end{vmatrix} = \frac{1}{4-3}\begin{vmatrix} 1.22205934 & 3\text{-}1.5 \\ 1.21721681 & 4\text{-}1.5 \end{vmatrix} = 1.229323135$$

El valor real de $\sqrt{1.5}$ és 1.224744871.

Mètode de Neville

Es construeixen polinomis $p_{i,j}(x)$ ($i = 0, \ldots, m-j$) de grau més petit o igual que j, que interpolen en les abscisses $\{x_i, x_{i+1}, \ldots, x_{i+j}\}$ ($j = 0, \ldots, m$); s'arriba al polinomi $p_m(x) \equiv p_{0,m}(x)$, de grau menor o igual que m, que interpolarà en $\{x_0, x_1, \ldots, x_m\}$:

$$
\begin{aligned}
p_{i,0} &= f_i \qquad (i = 0, \ldots, m);\\
p_{i,j+1} &= \frac{(x_{i+j+1} - x)p_{i,j}(x) - (x_i - x)p_{i+1,j}(x)}{x_{i+j+1} - x_i}\\
&\quad (i = 0, \ldots, m-j-1) \quad (j = 0, \ldots, m-1)
\end{aligned}
$$

Mètode de Neville

Esquema per a $m = 3$

Esquema per al mètode de Neville amb $m = 3$

$$
\begin{tikzcd}
p_{0,0}(x) \arrow[r] & p_{0,1}(x) \arrow[r] & p_{0,2}(x) \arrow[r] & p_{0,3}(x) \\
p_{1,0}(x) \arrow[r] \arrow[ur] & p_{1,1}(x) \arrow[r] \arrow[ur] & p_{1,2}(x) \arrow[ur] & \\
p_{2,0}(x) \arrow[r] \arrow[ur] & p_{2,1}(x) \arrow[ur] & & \\
p_{3,0}(x) \arrow[ur] & & &
\end{tikzcd}
$$

Mètode de Neville

Exemple

Exemple

Aproxima $f(2.5)$ donada la taula següent:

	x	$f(x)$
x_0	2.0	0.5103757
x_1	2.2	0.5207843
x_2	2.4	0.5104147
x_3	2.6	0.4813306
x_4	2.8	0.4359160

$$\begin{array}{llllll}
x_0 & P_0 & & & & \\
x_1 & P_1 & P_{0,1} & & & \\
x_2 & P_2 & P_{1,2} & P_{0,1,2} & & f(2.5) \\
x_3 & P_3 & P_{2,3} & P_{1,2,3} & P_{0,1,2,3} & \downarrow \\
x_4 & P_4 & P_{3,4} & P_{2,3,4} & P_{1,2,3,4} & P_{0,1,2,3,4}
\end{array}$$

Mètode de Neville

Exemple

Exemple

2.0	0.5103757				
2.2	0.5207843	0.5363972			
2.4	0.5104147	0.5052299	0.4974380		
2.6	0.4813306	0.4958726	0.4982119	0.4980829	
2.8	0.4359160	0.5040379	0.4979139	0.4980629	0.49807047

On $f(2.5) \approx 0.49807047$. Un exemple del càlcul de la matriu és:

$$P_{0,1}(x) = \frac{(x - x_0)P_1 - (x - x_1)P_0}{(x_1 - x_0)} =$$

$$= \frac{(2.5 - 2.0)0.5207843 - (2.5 - 2.2)0.5103757}{(2.2 - 2.0)} = 0.5363972$$

Continguts

Introducció

Expressem ara el polinomi interpolador de la forma següent:

$$p_m(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \ldots + c_m(x - x_0)(x - x_1)\cdots(x - x_{m-1})$$

El mètode de les diferències dividides ens permet calcular els coeficients c_j $(j = 0, \ldots, m)$ mitjançant la construcció de les anomenades diferències dividides, definides per:

$$\begin{aligned} f[x_i] &= f_i \qquad (i = 0, \ldots, m); \\ f[x_i, x_{i+1}, \ldots, x_{i+j}, x_{i+j+1}] &= \frac{f[x_{i+1}, \ldots, x_{i+j+1}] - f[x_i, \ldots, x_{i+j}]}{x_{i+j+1} - x_i} \\ & \quad (i = 0, \ldots, m - j) \qquad (j = 0, \ldots, m - 1) \end{aligned}$$

Esquema per a $m = 3$

Esquema per al mètode de les diferències dividides amb $m = 3$

$$
\begin{array}{c|cccccc}
x_0 & f[x_0] & & & & & \\
 & & \searrow & & & & \\
 & & & f[x_0,x_1] & & & \\
 & & \nearrow & & \searrow & & \\
x_1 & f[x_1] & & & & f[x_0,x_1,x_2] & \\
 & & \searrow & & \nearrow & & \searrow \\
 & & & f[x_1,x_2] & & & & f[x_0,x_1,x_2,x_3] \\
 & & \nearrow & & \searrow & & \nearrow \\
x_2 & f[x_2] & & & & f[x_1,x_2,x_3] & \\
 & & \searrow & & \nearrow & & \\
 & & & f[x_2,x_3] & & & \\
 & & \nearrow & & & & \\
x_3 & f[x_3] & & & & &
\end{array}
$$

Fòrmula d'interpolació de Newton

Usant la definició de les diferències dividides sobre $x \in [a, b]$, per a $x \neq x_k$ ($k = 0, \ldots, m$) s'arriba, per inducció sobre m, a la fòrmula següent:

$$\begin{aligned} f(x) &= f[x_0] + f[x_1, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \cdots \\ &+ f[x_0, x_1, \ldots, x_m](x - x_0)(x - x_1)\cdots(x - x_{m-1}) \\ &+ f[x_0, x_1, \ldots, x_m, x](x - x_0)(x - x_1)\cdots(x - x_{m-1})(x - x_m) \qquad (1) \end{aligned}$$

També, per inducció sobre m, es comprova que el polinomi $p_m(x)$ de grau menor o igual que m està també donat per la fòrmula d'interpolació de Newton:

$$\begin{aligned} p_m(x) &= f[x_0] + f[x_0, x_1](x - x_0) + \\ &+ f[x_0, x_1, x_2](x - x_0)(x - x_1) + \cdots \\ &+ f[x_0, x_1, \ldots, x_m](x - x_0)(x - x_1)\cdots(x - x_{m-1}) \end{aligned}$$

que és, de fet, el polinomi $p_{0,m}(x)$ construït pel mètode de Neville i, com a conseqüència, és el polinomi interpolador de f en les abscisses $\{x_0, \ldots, x_m\}$; és a dir, en (1)

$$c_j = f[x_0, x_1, \ldots, x_j]$$

Error d'interpolació

La fòrmula (1) dóna una nova expressió per a l'error d'interpolació

$$R(x) = f(x) - p_m(x) = f[x_0, x_1, \ldots, x_m, x](x - x_0)(x - x_1)\cdots(x - x_{m-1})(x - x_m)$$

Si les abscisses d'interpolació x_k $(k = 0, \ldots, m)$ i x mateix pertanyen a (a, b) i $f \in C^{m+1}(a, b)$, tenim:

$$R(x) = \frac{f^{(m+1)}(\xi(x))}{(m+1)!}(x - x_0)(x - x_1)\cdots(x - x_{m-1})(x - x_m)$$

on $\xi(x)$ depén de x i està contingut a $< x_0, x_1, \ldots, x_m, x >$.

Exemple

Exemple

Considerem el cas de $f(x) = \ln(x)$ amb

$$x_0 = 1, \quad x_1 = 2, \quad x_2 = 3, \quad x_4 = 5$$

El càlcul de les diferències dividides

$$\begin{array}{cccccc}
1 & 0 & & & & \\
 & & f[x_0, x_1] & & & \\
2 & 0.6931 & & f[x_0, x_1, x_2] & & \\
 & & f[x_1, x_2] & & f[x_0, x_1, x_2, x_3] & \\
3 & 1.0986 & & f[x_1, x_2, x_3] & & f[x_0, x_1, x_2, x_3, x_4] \\
 & & f[x_2, x_3] & & f[x_1, x_2, x_3, x_4] & \vdots \\
4 & 1.6863 & & f[x_2, x_3, x_4] & \vdots & \vdots \\
 & & f[x_3, x_4] & \vdots & \vdots & \vdots \\
5 & 1.6094 & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{array}$$

Exemple

Exemple

Considerem el cas de $f(x) = \ln(x)$ amb

$$x_0 = 1, \quad x_1 = 2, \quad x_2 = 3, \quad x_4 = 5$$

El càlcul de les diferències dividides

$$\begin{array}{cccccc}
1 & 0 & & & & \\
 & & 0.6931 & & & \\
2 & 0.6931 & & f[x_0,x_1,x_2] & & \\
 & & 0.4055 & & f[x_0,x_1,x_2,x_3] & \\
3 & 1.0986 & & f[x_1,x_2,x_3] & & f[x_0,x_1,x_2,x_3,x_4] \\
 & & 0.2877 & & f[x_1,x_2,x_3,x_4] & \vdots \\
4 & 1.6863 & & f[x_2,x_3,x_4] & \vdots & \vdots \\
 & & 0.2231 & \vdots & \vdots & \vdots \\
5 & 1.6094 & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{array}$$

Exemple

Exemple

Considerem el cas de $f(x) = \ln(x)$ amb

$$x_0 = 1, \quad x_1 = 2, \quad x_2 = 3, \quad x_4 = 5$$

El càlcul de les diferències dividides

1	0				
		0.6931			
2	0.6931		-0.1438		
		0.4055		$f[x_0, x_1, x_2, x_3]$	
3	1.0986		-0.0589		$f[x_0, x_1, x_2, x_3, x_4]$
		0.2877		$f[x_1, x_2, x_3, x_4]$	$\vdots$
4	1.6863		-0.0323	$\vdots$	$\vdots$
		0.2231	$\vdots$	$\vdots$	$\vdots$
5	1.6094	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Exemple

Exemple

Considerem el cas de $f(x) = \ln(x)$ amb

$$x_0 = 1, \quad x_1 = 2, \quad x_2 = 3, \quad x_4 = 5$$

El càlcul de les diferències dividides

1	0				
		0.6931			
2	0.6931		-0.1438		
		0.4055		0.0283	
3	1.0986		-0.0589		$f[x_0, x_1, x_2, x_3, x_4]$
		0.2877		0.0089	$\vdots$
4	1.6863		-0.0323	$\vdots$	$\vdots$
		0.2231	$\vdots$	$\vdots$	$\vdots$
5	1.6094	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Exemple

Exemple

Considerem el cas de $f(x) = \ln(x)$ amb

$$x_0 = 1, \quad x_1 = 2, \quad x_2 = 3, \quad x_4 = 5$$

El càlcul de les diferències dividides

1	0				
		0.6931			
2	0.6931		-0.1438		
		0.4055		0.0283	
3	1.0986		-0.0589		-0.0485
		0.2877		0.0089	⋮
4	1.6863		-0.0323	⋮	⋮
		0.2231	⋮	⋮	⋮
5	1.6094	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮

Calculeu l'expressió del polinomi interpolador.

Tema 3. Interpolació polinòmica

Matemàtiques II. Grau de Química

Francisco Guerrero Cortina

Dept. Matemàtica Aplicada, Universitat de València

Curs 2012-2013

Tema 4. Integració Numèrica

Matemàtiques II. Grau de Química

Francisco Guerrero Cortina

Dept. Matemàtica Aplicada, Universitat de València

Curs 2012-2013

Continguts

Continguts

Introducció

Matemàticament, la integració es representa per:

$$I = \int_a^b f(x)dx$$

que és la integral de la funció $f(x)$ pel que fa a la variable independent x, avaluada en els límits $x = a$ i $x = b$.

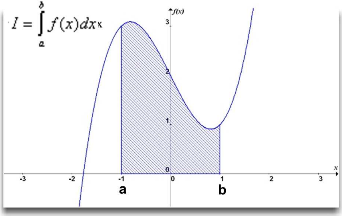


Fig 1

El seu significat és l'àrea de la funció $f(x)$ sobre el rang $x = a$ a $x = b$.

Tema 4. Integració Numèrica
Francisco Guerrero Cortina

Integració numèrica
Introducció
Mètodes d'integració numèrica
Regla del trapezi
Regla de Simpson
Regla de Simpson 1/3
Regla de Simpson 3/8

Integració numèrica d'equacions diferencials
Mètode d'Euler
Procediment d'Euler-Cauchy millorat
Algorismes de Runge-Kutta
Algorisme de Runge-Kutta de segon ordre
Algorisme de Runge-Kutta de quart ordre

Continguts

Fòrmules de Newton-Cotes

Les fòrmules d'integració de Newton-Cotes són els esquemes d'integració numèrica més comuns.
Es basen en l'estratègia de reemplaçar una funció complicada o dades tabulades amb una funció aproximada que siga fàcil d'integrar:

$$I = \int_a^b f(x)dx \cong \int_a^b f_n(x)dx$$

on $f_n(x)$ és un polinomi de la forma

$$f_n(x) = a_0 + a_1 x + \cdots + a_n x^n$$

on n és el grau del polinomi.

Fòrmules de Newton-Cotes

Per exemple, en la figura de l'esquerra s'usa el polinomi de primer ordre (una línia recta) com una aproximació. Mentre que en la figura de la dreta s'empra una paràbola per al mateix propòsit.

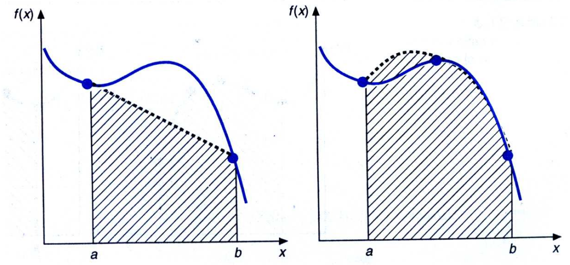
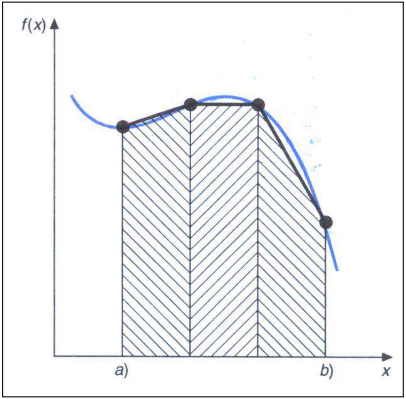


Fig 2 Fig 3

Fòrmules de Newton-Cotes

Per exemple, en la figura següent s'usen tres segments de línia recta per a aproximar la integral. Poden utilitzar-se polinomis d'ordre superior per als mateixos propòsits.



Continguts

Definició

La regla trapezoïdal és la primera de les fòrmules d'integració tancada de Newton-Cotes. Correspon al cas on el polinomi de l'equació següent és de primer ordre:

$$I = \int_a^b f(x)dx \cong (b-a)\left[\frac{f(a)+f(b)}{2}\right]$$

Geomètricament, la regla del trapezi és equivalent a aproximar l'àrea per la del trapezi sota la línia recta que connecta f(a) i f(b) com es mostra en la figura.

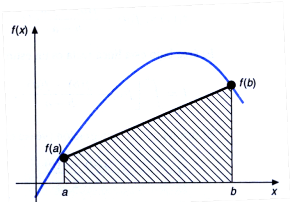


Fig. 7

Regla del trapezi

Recordeu que la fòrmula per a calcular l'àrea d'un trapezoide és l'altura per la mitjana de les bases, tal com es mostra en la figura següent:

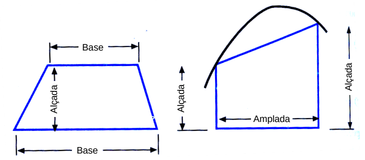


Fig. 8 **Fig. 9**

Exercicis

Exercici

Empreu la regla del trapezi per aproximar els valors de les integrals següents:

a) $\int_0^1 e^{x^2} dx$

b) $\int_2^4 \frac{e^x}{x} dx$

Aplicació múltiple de la regla del trapezi

La regla del trapezi es pot ampliar si subdividim l'interval $[a, b]$ en n subintervals, tots de la mateixa longitud $h = \frac{b-a}{n}$.
Siga $P = \{x_0, x_1, \ldots, x_n\}$ la partició que es forma en fer aquesta subdivisió. Usant les propietats de la integral, tenim que:

$$\int_a^b f(x)d(x) = \int_{x_0}^{x_1} f(x)d(x) + \int_{x_1}^{x_2} f(x)d(x) + \cdots + \int_{x_{n-1}}^{x_n} f(x)d(x)$$

Aplicant la regla del trapezi a cadascuna de les integrals, obtenim:

$$\int_a^b f(x)d(x) \approx (x_1 - x_0)\left[\frac{f(x_0)+f(x_1)}{2}\right] + \cdots + (x_n - x_{n-1})\left[\frac{f(x_{n-1})+f(x_n)}{2}\right]$$

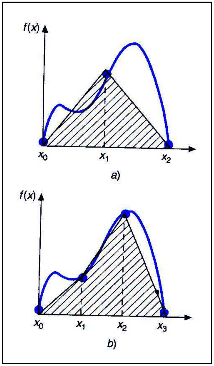
Ara bé, ja que els subintervals tenen la mateixa longitud h, tenim que:

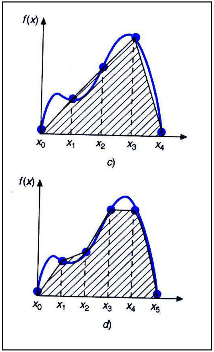
$$\int_a^b f(x)d(x) \approx \frac{h}{2}[f(x_0) + 2f(x_1) + 2f(x_2) + \cdots + 2f(x_{n-1}) + f(x_n)]$$

Aplicació múltiple de la regla del trapezi

Exemple

Il·lustració de la regla trapezoïdal d'aplicació múltiple: a) dos segments, b) tres segments, c) quatre segments i d) cinc segments:





Aplicació múltiple de la regla del trapezi

Error

L'error d'integració s'obté per l'expressió

$$R = -\frac{nh^3}{12} y''(\xi) = -\frac{(b-a)h^2}{12} y''(\xi)$$

on $\xi \in [a, b]$

Aplicació múltiple de la regla del trapezi

Exercici

Exercici

Empreu la regla del trapezi per a aproximar el valor de la integral següent:

$$\int_0^1 e^{x^2}\,dx$$

si subdividim en 5 intervals.

Tema 4. Integració Numèrica

Francisco Guerrero Cortina

Integració numèrica
Introducció
Mètodes d'integració numèrica
Regla del trapezi
Regla de Simpson
Regla de Simpson 1/3
Regla de Simpson 3/8

Integració numèrica d'equacions diferencials
Mètode d'Euler
Procediment d'Euler-Cauchy millorat
Algorismes de Runge-Kutta
Algorisme de Runge-Kutta de segon ordre
Algorisme de Runge-Kutta de quart ordre

Continguts

Introducció

A més d'aplicar la regla trapezoïdal amb segmentació més fina, una altra manera d'obtenir una estimació més exacta de la integral és amb l'ús de polinomis d'ordre superior per a connectar els punts.

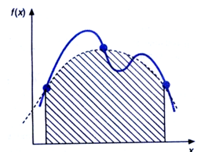


Fig. 12

Per exemple, si hi ha un punt extra a la meitat del camí $f(a)$ i $f(b)$, els tres punts es poden connectar en una paràbola.

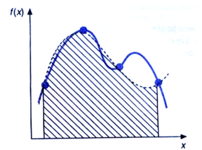


Fig. 13

Si hi ha dos punts igualment espaiats entre $f(a)$ i $f(b)$, els quatre punts es poden connectar amb un polinomi de tercer ordre.

Les fòrmules que resulten en prendre les integrals sota aquests polinomis són conegudes com a regla de Simpson.

Regla de Simpson 1/3

La regla de Simpson 1/3 resulta quan una interpolació polinòmica de segon ordre és substituïda en l'equació:

$$\int_a^b f(x)dx \approx \int_a^b f_2(x)dx$$

Si a i b es designen com x_0 i x_2 i $f_2(x)$ és representada per un polinomi de Lagrange de segon ordre i la integral es transforma en:

$$I = \int_{x_0}^{x_2} \left[\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2) \right] dx$$

Després de la integració i el maneig algebraic, resulta la fòrmula següent:

$$I \cong \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)]$$

on $h = \frac{b-a}{2}$ i x_1 és el punt mitjà entre a i b.

Exercici

Exercici

Empreu la regla de Simpson d'1/3 per a aproximar el valor de les integrals següents:

a) $\int_0^1 e^{x^2}\,dx$

b) $\int_2^4 \frac{e^x}{x}\,dx$

Regla de Simpson 1/3 múltiple

La regla de Simpson es pot ampliar si subdividim l'interval $[a, b]$ en n subintervals, tots de la mateixa longitud $h = \frac{b-a}{n}$.
Siga $P = \{x_0, x_1, \dots, x_n\}$ la partició que es forma en fer aquesta subdivisió i siga $P_m = \{x_{i-1}, x_i\}$ el conjunt dels punts mitjans dels subintervals.
Usant les propietats de la integral, tenim que:

$$\int_a^b f(x)d(x) = \int_{x_0}^{x_1} f(x)d(x) + \int_{x_1}^{x_2} f(x)d(x) + \cdots + \int_{x_{n-1}}^{x_n} f(x)d(x)$$

Aplicant la regla de Simpson 1/3 a cadascuna de les integrals, obtenim:

$$\int_a^b f(x)d(x) \approx 2h\frac{f(x_0)+4f(x_1)+f(x_2)}{6} + 2h\frac{f(x_1)+4f(x_2)+f(x_3)}{6} + \cdots + 2h\frac{f(x_{n-2})+4f(x_{n-1})+f(x_n)}{6}$$

Combinant termes ens queda:

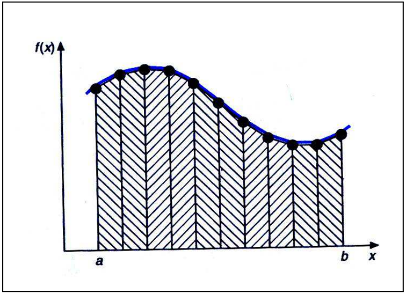
$$I \approx \frac{h}{3}[y_0 + y_m + 4\left(y_1 + y_3 + \cdots y_{2m-1}\right) + 2\left(y_2 + y_4 + \cdots y_{2m-2}\right)]$$

on $n = 2m$ i $y_i = f(x_i)$.

Regla de Simpson 1/3 múltiple

Exemple

Representació gràfica de la regla de Simpson 1/3 d'aplicació múltiple. Observeu que el mètode es pot emprar només si el nombre de segments és parell:



Regla de Simpson 1/3 múltiple

Error

L'error d'integració s'obté per l'expressió

$$R = \left| -\frac{mh^5}{90} y^{iv}(\xi) \right| = \left| -\frac{(b-a)h^4}{180} y^{iv}(\xi) \right|$$

on $\xi \in [a, b]$.

Regla de Simpson 1/3 múltiple

Exercici

Exercici

Empreu la regla de Simpson d'1/3 per a aproximar el valor de les integrals següents:

a) $\int_0^1 e^{x^2}\,dx$ en 6 intervals

b) $\int_2^4 \frac{e^x}{x}\,dx$ en 4 intervals

Regla de Simpson 3/8

En una manera similar a la derivació de la regla trapezoïdal i regla de Simpson 1/3, un polinomi de Lagrange de tercer ordre es pot ajustar a quatre punts i integrar-se:

$$\int_a^b f(x)dx \approx \int_a^b f_3(x)dx$$

Per obtenir

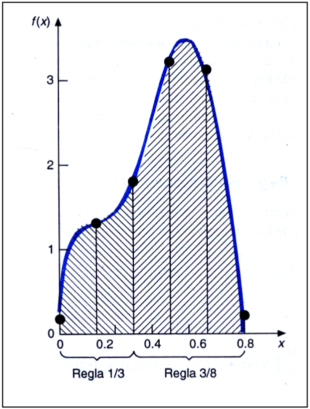
$$I \cong \frac{3}{8}h[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)]$$

on $h = \frac{(b-a)}{3}$.

Aquesta equació s'anomena regla de Simpson 3/8.

Regla de Simpson 3/8

En la Figura veiem com es poden usar en conjunt les regles de Simpson d'1/3 i 3/8 per a manejar aplicacions múltiples amb nombres senars d'intervals.



Regla de Simpson 3/8

Exercici

Exercici

Aproximeu la integral següent usant la regla de Simpson de 3/8:

$$\int_1^4 e^x \ln x \, dx$$

Regla de Simpson 3/8 múltiple

Igual que en els casos anteriors, la regla de Simpson de 3/8 es pot estendre si subdividim l'interval $[a, b]$ en n intervals de la mateixa longitud h.
Siga $P = \{x_0, x_1, \ldots, x_n\}$ la partició que es forma en fer aquesta subdivisió. Cada subinterval el dividirem en tres parts iguals. Aleshores tindrem $n = 3m$ punts de manera que $y_i = f(x_i) \quad i = 1, \ldots, 3m$.
Emprant les mateixes propietats que als casos anteriors, tindrem que l'aproximació és la següent:

$$I \approx \frac{3h}{8}[y_0 + 3y_1 + 3y_2 + 2y_3 + 3y_4 + 3y_5 + 2y_6 + 3y_7 + \cdots + 3y_{3m-1} + y_{3m}]$$

Regla de Simpson 3/8 múltiple

Error

L'error d'integració s'obté per l'expressió

$$R = \left| -\frac{3h^5}{80} y^{iv}(\xi) \right| = \left| -\frac{(b-a)h^4}{80} y^{iv}(\xi) \right|$$

on $\xi \in [a, b]$.

Regla de Simpson 3/8 múltiple

Exercici

Exercici

Aproximeu la integral següent usant la regla de Simpson de 3/8:

$$\int_1^4 e^x \ln x \, dx$$

subdividint en 9 intervals.

Continguts

Introducció

Considerem una equació diferencial de primer ordre amb condició inicial, és a dir,

$$y' = f(x, y), \quad y(x_0) = y_0 \tag{1}$$

Volem construir la solució $y(x)$ que satisfà aquesta equació juntament amb la condició inicial.

Mètode d'Euler

El mètode bàsic d'Euler consisteix a construir les aproximacions a la funció solució a partir de la primera derivada, que coneixem per l'equació (1). Això ens proporciona aquest algorisme:

$$y_{i+1} = y_i + hf(x_i, y_i)$$

Mètode d'Euler

Exemple

Exemple

Considereu el problema de valors inicials

$$\begin{cases} y'(t) &= 2 \cdot t + e^t, \quad t \in [0,1] \\ y(0) &= 0 \end{cases}$$

Considerem els instants $t_0 = 0.0$, $t_1 = 0.1$, $t_2 = 0.2$, $t_3 = 0.3$, ... $t_9 = 0.9$ i $t_{10} = 1.0$.
Per tant, $h = 0.1$
Si apliquem el mètode d'Euler, tenim

$$\begin{cases} y_0 &= 0 \\ y_{n+1} &= y_n + 0.1 \cdot (2 \cdot t_n + e^{t_n}) \quad (n = 0, 1, 2, \ldots, 9) \end{cases}$$

Per tant, els primers quatre valors calculats seran:

$$\begin{aligned} y_0 &= 0 \\ y_1 = y_0 + 0.1 \cdot (2 \cdot t_0 + e^{t_0}) = 0 + 0.1 \cdot (2 \cdot 0 + e^0) &= 0.1 \\ y_2 = y_1 + 0.1 \cdot (2 \cdot t_1 + e^{t_1}) = 0 + 0.1 \cdot (2 \cdot 0.1 + e^{0.1}) &= 0.23052 \\ y_3 = y_2 + 0.1 \cdot (2 \cdot t_2 + e^{t_2}) = 0 + 0.1 \cdot (2 \cdot 0.2 + e^{0.2}) &= 0.39265 \\ y_4 = y_3 + 0.1 \cdot (2 \cdot t_3 + e^{t_3}) = 0 + 0.1 \cdot (2 \cdot 0.3 + e^{0.3}) &= 0.58764 \end{aligned}$$

Mètode d'Euler

Exemple

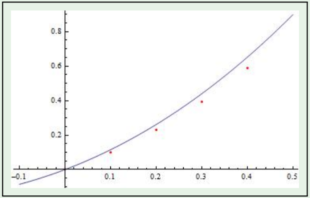
Exemple

Considereu el problema de valors inicials

$$\begin{cases} y'(t) &= 2 \cdot t + e^t, \quad t \in [0,1] \\ y(0) &= 0 \end{cases}$$

Solució

	x_i	y_i
$i=0$	0	0
$i=1$	0.1	0.1
$i=2$	0.2	0.23052
$i=3$	0.3	0.39265
$i=4$	0.4	0.58764



Sabent que la solució exacta és $y(x) = -1 + e^x + x^2$

Mètode d'Euler

Error

L'error comés és

$$R = \left| \frac{h^2}{2} y''(\xi) \right|$$

amb $x_i < \xi < x_i + h$

Continguts

Procediment d'Euler-Cauchy millorat

Les equacions d'aquest són

$$
\begin{aligned}
y^*_{n+1} &= y_n + f(x_n, y_n) \\
x_{n+1} &= x_n + h \\
y_{n+1} &= y_n + \frac{1}{2} \cdot h \cdot [f(x_n, y_n) + f(x_{n+1}, y^*_{n+1})]
\end{aligned}
$$

Procediment d'Euler-Cauchy millorat

Exemple

Exemple

Aproximar el PVI $y' = x + y$, $y(0) = 0$ amb $h = 0.2$ pel mètode d'Euler-Cauchy millorat. Anomenem

$$k_1 = hf(x_n, y_n) = 0.2(x_n + y_n)$$

$$\begin{aligned} k_2 = hf(x_{n+1}, y^*_{n+1}) = 0.2(x_{n+1} + y^*_{n+1}) &= 0.2(x_n + 0.2 + y_n + 0.2(x_n + y_n)) \\ &= 0.04 + 0.24x_n + 0.24y_n \end{aligned}$$

Aleshores,

$$y_{n+1} = y_n + \frac{k_1 + k_2}{2}$$

$$y_{n+1} = y_n + \frac{0.2x_n + 0.2y_n + 0.04 + 0.24x_n + 0.24y_n}{2} = y_n + \frac{0.44x_n + 0.44y_n + 0.04}{2}$$

Procediment d'Euler-Cauchy millorat

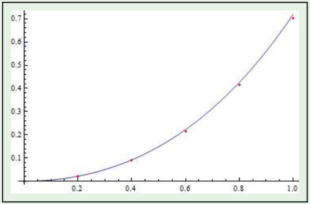
Exemple

Exemple

Aproximar el PVI $y' = x + y$, $y(0) = 0$ amb $h = 0.2$ pel mètode d'Euler-Cauchy millorat.

Solució

	x_i	y_i
$i = 0$	0	0
$i = 1$	0.2	0.02
$i = 2$	0.4	0.0884
$i = 3$	0.6	0.215848
$i = 4$	0.8	0.415335
$i = 5$	1.0	0.702708



Sabent que la solució exacta és $y(x) = -1 + e^x - x$.

Continguts

Algorisme de Runge-Kutta de segon ordre

Les equacions d'aquest són:

$$\begin{aligned} k_1 &= f(x_i, y_i) \\ k_2 &= f(x_i + h, y_i + hk_1) \\ y_{i+1} &= y_i + \frac{h}{2}[k_1 + k_2] \end{aligned}$$

Algorisme de Runge-Kutta de segon ordre

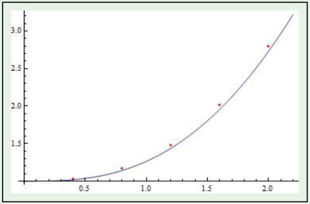
Exemple

Exemple

Aproximar el PVI $y'(x) = -y + x^2 + 1$, $y(0) = 1$, $h = 0.4$. Traure els valors fins a $x = 2.0$.

Solució

	x_i	y_i
$i = 0$	0	1
$i = 1$	0.4	1.032
$i = 2$	0.8	1.16896
$i = 3$	1.2	1.47969
$i = 4$	1.6	2.01099
$i = 5$	2.0	2.79467



Sabent que la solució exacta és $y(x) = -2e^{-x} + x^2 - 2x + 3$.

Algorisme de Runge-Kutta de quart ordre

Les seues equacions són

$$
\begin{aligned}
k_1 &= f(x_i, y_i) \\
k_2 &= f(x_i + \frac{h}{2}, y_i + h \cdot \frac{k_1}{2}) \\
k_3 &= f(x_i + \frac{h}{2}, y_i + h \cdot \frac{k_2}{2}) \\
k_4 &= f(x_i + h, y_i + h \cdot k_3) \\
y_{i+1} &= y_i + \frac{h}{6}[k_1 + 2k_2 + 2k_3 + k_4]
\end{aligned}
$$

Algorisme de Runge-Kutta de quart ordre

Exemple

Exemple

Amb el mètode Runge-Kutta amb $h = 0.1$ obtenir una aproximació a $y' = 2xy$, $y(1) = 1$. Calcularem el cas $i = 0$.

$$
\begin{aligned}
k_1 &= f(x_0, y_0) = 2.0 \\
k_2 &= f(x_0 + \frac{1}{2}0.1, y_0 + 0.1 \cdot \frac{k_1}{2}) = 2.31 \\
k_3 &= f(x_0 + \frac{1}{2}0.1, y_0 + 0.1 \cdot \frac{k_2}{2}) = 2.34255 \\
k_4 &= f(x_0 + 0.1, y_0 + 0.1 \cdot k_3) = 2.71536 \\
y_1 &= y_0 + \frac{0.1}{6}[k_1 + 2k_2 + 2k_3 + k_4] = 1.23367435
\end{aligned}
$$

Aleshores, $x_1 = 1.10$, $y_1 = 1.2337$ i fent la resta de càlculs:

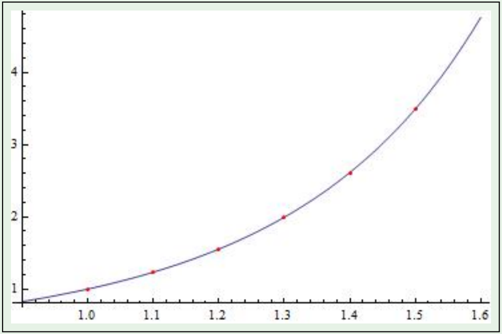
x_i	1.0	1.10	1.20	1.30	1.40	1.50
y_i	1.0	1.2337	1.5527	1.9937	2.6116	3.4902

Algorisme de Runge-Kutta de quart ordre

Exemple

Exemple

x_i	1.0	1.10	1.20	1.30	1.40	1.50
y_i	1.0	1.2337	1.5527	1.9937	2.6116	3.4902



Sabent que la solució exacta és $y(x) = e^{-1+x^2}$.

Tema 4. Integració Numèrica

Matemàtiques II. Grau de Química

Francisco Guerrero Cortina

Dept. Matemàtica Aplicada, Universitat de València

Curs 2012-2013