

1- TEORIA DEL MOSTREIG

INTRODUCCIÀ

Diem, que l'estadística es pot considerar sota dos tipus de raonaments. El primer és l'estadística descriptiva i la definim com una col·lecció de dades numèriques ordenades que estudien el comportament i característiques d'un conjunt d'elements de una població o mostra, com per exemple, l'ordenació d'hotels per províncies respecte la seva categoria. El segon tipus, és l'estadística inductiva o inferència estadística, on a partir de les característiques recollides d'una mostra podem estimar com és una població en general, recolzant-nos en el càlcul de probabilitats i acceptant un determinat marge d'error, ja que es un raonament predictiu. Per exemple mitjançant enquestes a un grup de ciutadans podem interpretar els llocs preferits per fer turisme.

TEORIA DEL MOSTREIG

Una part fonamental de un estudi estadístic és obtenir uns resultats fiables i que puguin ser aplicables, però molt sovint es fa pràcticament impossible l'estudi de tota una població, definint població com el conjunt de tots els elements que tenen una característica determinada i s'ha de prendre el màxim d'element que podem observar. En general, suposem que la població és molt gran. Per solucionar aquest problema, es dur a terme l'estudi basant-se en un subconjunt de la població que anomenem mostra. Llavors es fa necessari, que aquest subconjunt de la població, tinguin algunes característiques específiques per tal de poder generalitzar els resultats a tot el conjunt de la població. Direm aleshores que la mostra és representativa. La relació entre aquesta mostra representativa i el conjunt de la població, es l'estudi fonamental de la teoria del mostreig. A continuació explicarem els diferents mètodes que podem seguir alhora d'escollir una mostra:

- Atenent a l'extracció de la mostra, trobem:
 - Mostreig amb reemplaçament: seran aquells, en els quals un cop seleccionat un element de la població, el tornarem a tenir en compte quan seleccionem el següent element a ser estudiat, per tant, s'haurà de tenir en compte que si la població a estudi és petita, aquesta possibilitat de repetir observacions pot afectar de manera important al resultat de l'estudi.
 - Mostreig sense reemplaçament: seran aquells, en els quals un cop seleccionat un element de la població per formar part de la mostra, no se'l tornarà a tenir en compte.
- Atenent a la selecció de la mostra com un element de la població, trobem:

MOSTREIG ALEATORI: tots els elements de la mostra tenen les mateixes oportunitats de ser escollits per formar part de la mostra. En trobem tres tipus:

- Mostreig aleatori simple: es determina a l'atzar els individus que constituïran la mostra, es a dir, cada element de la població tindrà la mateixa probabilitat de ser inclòs en aquesta.
- Mostreig aleatori sistemàtic: es selecciona a l'atzar un element de la població i a partir d'aquest element seleccionem els elements següents de k en k .
- Mostreig aleatori estratificat: la població es divideix en grups homogenis que anomenem estrats, i posteriorment s'extreu una mostra aleatòria simple de cada estrat. L'estratificació augmenta la precisió de la mostra. Per fer els estrats s'haurà de determinar en primer terme el nombre d'estrats i en segon terme determinar el nombre d'elements de la mostra que hi haurà a cada estrat.

MOSTREIG NO ALEATORI: només es recullen les dades d'aquells elements que pel investigador son

representatiu de la població a estudiar.

- Atenent a la selecció de mostra d'un grup d'elements de la població, trobem:
 - **Mostreig per conglomerats i Àrees:** S'utilitza quan es té una idea clara del total de la població, o seria massa costós generar-ne la llista. Les unitats conglomerats (municipis, ciutats...) contenen nombrosos diferents d'elements o subunitats (regions geogràfiques, escoles, habitatges, persones). Un cop determinats els conglomerats i els individus a seleccionar dins de cadascun es pot aplicar qualsevol dels mètodes anteriors.

La selecció d'un tipus o un altre d'aquests mètodes es fa segons l'estudi que volem fer, el tipus de dades i les característiques de la població. Haurem de tenir en compte que es poden produir determinats errors, que poden ser deguts no només a errors de càlcul, si no a diverses situacions, que podem evitar com són els errors de plantejament o planificació en l'estudi, errors en les respostes degut a una formulació errònia de les preguntes, errors per insuficiència en la representació de parts de la població, etc. Però hi ha un tipus d'error quan treballem amb subconjunts de la població que és l'error mostral que no es evitable, encara que es pot reduir a nivells acceptables. Abans de definir matemàticament aquest tipus d'error, per tal de poder calcular-lo i entendre-ho millor introduïrem alguns conceptes bàsics en l'estadística.

CONCEPTES BÀSICS

Un cop fet el mostreig de l'estudi en podem calcular els seus paràmetres i estadístics:

- Els paràmetres són valors numèrics que descriuen una característica de la població. Els anomenem paràmetres de localització o centralització i són els següents:
 - **Moda:** La moda (M_o) d'una distribució estadística és el valor (o valors) de freqüència més alta, es a dir, el valor que mes vegades apareix a la mostra. És la que menys informació conté, i pot no ser única.
 - **Mediana:** S'anomena mediana d'una distribució estadística el valor de la variable tal que la freqüència absoluta dels valors més petits que ell és igual a la freqüència absoluta dels valors més grans que ell. No es veu afectada per valors atípics i necessita criteri d'ordenació de dades. Es representa per M .
 - **Mitjana:** S'anomena mitjana aritmètica d'una distribució estadística el quocient entre la suma de tots els valors de X_i observats i el nombre total d'observacions, N . Es a dir, els valors que 'prometgen' el conjunt de valors de la mostra. És la més utilitzada, ja que conté molta informació. Només té sentit en variables quantitatives. Es representa per $\bar{x}$.

Aritmètica: Geomètrica: Harmònica:

- Els estadístics són valors numèrics que descriuen una característica de la mostra. Els anomenem estadístics de dispersió i són els següents:
 - **Rang:** És la diferència entre el valor màxim i el mínim de la distribució de la mostra.
 - **Desviacions:** És la mitjana aritmètica de les desviacions preses en valor absolut.
 - **Variança:** És la mitjana aritmètica dels quadrats de les desviacions de cada valor. La representem per S^2 .
 - **Desviació típica:** És l'arrel quadrada positiva de la variança. Es representa per S_x .

- **Coefficient de variació:** És el quocient entre la desviació típica i el valor absolut de la mitjana aritmètica.

Error Mostral:

És la diferència entre un estadístic i el seu paràmetre corresponent. És una mesura de la variabilitat de les estimacions de mostres que es repeteixen al voltant del valor de la població, ens dona una idea clara de fins a on i amb quina probabilitat una estimació d'una mostra s'allunya del valor que haguéssim obtingut en un cens complet. Aleshores, un estadístic serà més precís com més petit sigui el seu error. Podríem dir que és la desviació de la distribució mostral de un estadístic i la seva fiabilitat. El calculem amb la següent fórmula:

Hem vist la importància d'escollir un mètode de mostreig adequat i la importància de no cometre errors en la obtenció de la mostra, càlculs, etc. Ara veurem un altre punt de importància per tal de que la mostra sigui representativa, i és la seva grandària. Per tal de escollir els valors necessaris d'elements per la mostra haurem d'establir dos suposats, en primer lloc el nivell de confiança que volem assolir i en segon lloc el error màxim que estem disposats a cometre en el nostre estudi. Aplicarem un cop fet això, la següent fórmula:

On Z és el nivell de confiança escollit, σ és la variància i e és l'error màxim. Aquesta equació ha de complir; . Si no compleix això haurem de trobar la n adequada de la següent manera: . A partir d'aquí podem dir que tenim la mida correcta per la qual farem el mostreig.

VARIABLES ALEATÒRIES

Es diu que una variable és aleatòria si els possibles valors que pren són determinats per l'atzar, es a dir coneixem els valors que adoptarà la variable però no sabem si tindran lloc, només sabem que es produiran amb una certa probabilitat.

Les variables aleatòries es classifiquen en:

- Discretes: les que resulten de comptar el nombre de casos en els quals el esdeveniment té lloc. Per exemple; el nombre de fills d'una família. És un conjunt discret de valor $X = \{x_1, x_2, \dots, x_n\}$
- Continues: les que resulten com a producte d'una medició. Per exemple: el pes, l'alçada, duració d'una operació, etc. Són un conjunt infinit continu de nombres reals (normalment intervals)

Classificació i propietats de les variables aleatòries discretes:

Distribució Binomial, $B(n, p)$: Aquesta distribució es basa en el procés de Bernoulli, és tot aquell experiment que consisteix en una sèrie de proves repetides que donen com a resultat un èxit o un fracàs de la propietat en estudi, sent aleatoris e independents. Per identificar aquest tipus de distribució, s'han de donar les següents condicions: Els resultats s'han de poder classificar en èxits si compleixen la condició i en fracàs en cas contrari. Els resultats són independents entre ells, i la probabilitat d'obtenir un resultat concret és constant duran tot el procés.

Distribució de Poisson de paràmetre λ : compta el nombre de successos en un interval de temps.

Distribució Geomètrica: En aquesta distribució es realitzen les proves de Bernoulli i compta el nombre de fracassos abans de l'èxit.

Classificació i propietats de les variables aleatòries contínues:

Distribució Normal o de Gauss, $N(\mu, \sigma^2)$: Aquesta distribució queda definida per el paràmetre μ , que és la esperança matemàtica i σ^2 , que és la desviació típica. Es representa mitjançant la corba de Gauss i necessita de intervals ja que és una variable contínua. És aplicable a molt camps i la seva probabilitat ve donada per la següent fórmula:

MODELS DE DISTRIBUCIÓ DE PROBABILITATS

Fins ara hem vist el tipus de mostreig i variables aleatòries, les hem estudiat i calculat tots els seus estadístics. Això ens dona molta informació pel que fa a l'experiència real. Ara les estudiarem de forma concreta veient els seus models de probabilitat i comparant els resultats obtinguts que hem vist fins ara a la primera part del treball, amb el càlcul dels estadístic de la mostra estudiada en funció de les probabilitats.

• Distribució de Poisson:

La variable que conta les persones que estan dins d'un ascensor en un període de temps, s'adapta molt a una distribució de Poisson, per tant, s'espera que s'ajusti a una distribució de Poisson amb paràmetre λ . Una bona estimació de λ és la mitjana:

$$\hat{\lambda} = \bar{x}$$

Tot seguit explicarem en què consisteix aquesta distribució de Poisson. La distribució de Poisson és una distribució de probabilitat discreta. Aquesta expressa la probabilitat d'un nombre de successos produïts en un període de temps fix. Aquesta distribució va ser descoberta per Simon-Denis Poisson (1781-1840). El treball estava enfocat a determinades variables aleatòries N que expliquen, entre altres coses, un nombre d'esdeveniments discrets (molts cops anomenats "arribos") que tenen lloc durant un interval de temps de duració determinada. Si el nombre esperat d'esdeveniments es aquest interval és λ esdeveniments (sigui k un enter no negatiu, $k = 0, 1, 2, \dots$) es igual a:

On:

e és la base del logaritme natural ($e = 2.71828\dots$),

$k!$ és el factorial de k ,

k és el nombre d'esdeveniments d'un succés,

λ és un nombre real positiu, equivalent al nombre esperat d'esdeveniments durant un interval donat.

La distribució de Poisson pot ser vista com un cas limitant de la distribució binomial, es a dir, que una distribució binomial en la que $n \rightarrow \infty$ i $p \rightarrow 0$ es pot aproximar per una distribució de Poisson de valor $\lambda = np$. La distribució de Poisson és també coneguda com Poissoniana, anomenada així en honor de Gauss per una distribució de Gauss o distribució normal. Aquesta distribució té connexió amb els processos de Poisson. S'aplica a varis fenòmens discrets de la naturalesa (així, aquells fenòmens que es donen 0, 1, 2, 3, ... vegades durant un període definit de temps o en una àrea determinada) quan la probabilitat de que el esdeveniment tingui lloc és constant en el temps o en l'espai.

Alguns exemples d'aquests esdeveniments son:

- el nombre de cotxes que passen a través de un cert punt en una ruta en un temps determinat (1 dia).
- El nombre d'errors d'ortografia que hom comet al escriure una única pàgina.

- El nombre de trucades telefòniques d'una central telefònica per minut.
- El nombre d'estrelles en un determinat volum d'espai

Els valors de $P(x)$ poden ser calculats utilitzant els logaritmes;

Mitjana	$\hat{\mu} = \bar{x}$
Variança	$= \hat{\sigma}^2$
Desviació estàndard	$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$
Coeficient de Biais	
Coeficient de apuntament	

- *Distribució Normal:*

La segona variable que hem escollit, el temps que tarden els alumnes a arribar a la facultat, es una variable continua que s'espera que s'ajusti millor a una distribució normal, on : $\hat{\mu} = \bar{x}$ i $\hat{\sigma} = S$

La distribució normal, també coneguda com a distribució gaussiana, és una important família de distribucions de probabilitat contínues i és aplicable a molts camps. Cada membre de la família queda definit per dos paràmetres: la mitjana i la desviació estàndard. Un cas particular és la distribució normal estàndard, pel qual la mitjana és 0 i la desviació estàndard és 1.

Fou Carl Friedrich Gauss qui descobrí la distribució normal quan analitzava dades astronòmiques, i definí l'equació de la seva funció de densitat de probabilitat. Aquesta distribució també s'anomena campana de Gauss, doncs el gràfic de la seva funció de densitat de probabilitat s'assembla a una campana.

Te molta importància la distribució normal en les ciències naturals i en el comportament del teorema central del límit. Aquest teorema estableix que la suma d'un elevat nombre de efectes independents segueix una distribució normal. D'aquesta manera, és útil en processos en els quals hi ha errors de mesura que es deuen a un elevat nombre de factors, tots ells contribuint una petita porció a l'error total. En la teoria de probabilitat i d'inferència estadística, el teorema central del límit garanteix que un llarg nombre d'estadístics segueixen la distribució normal, si més no aproximadament. Per exemple, la mitjana mostral o els estimadors més semblants segueixen aproximadament una distribució normal sota certes condicions matemàtiques que són força generals.

Funció de densitat de probabilitat:

on $\hat{\sigma}$ és la desviació estàndard, $\hat{\mu}$ és l'esperança matemàtica, i

és la funció de densitat de probabilitat de la distribució normal estàndard, és a dir, la distribució normal amb $\hat{\mu} = 0$ i $\hat{\sigma} = 1$.

Funció de distribució:

No existeix una fórmula tancada per a la funció de distribució, però pot aproximar-se amb diversos mètodes, com integració numèrica, sèries de Taylor, sèries asimptòtiques i fraccions continuades.

Tipificació de variables aleatòries normals:

És possible relacionar totes les variables aleatòries normals amb la distribució normal tipificada. Si, aleshores és una variable aleatòria normal tipificada: .

Una conseqüència important és que la funció de distribució de Z és:

on l'À©s la funciÀ³ de distribuciÀ³ normal tipificada per a tot real k .

La funciÀ³ de distribuciÀ³ normal tipificada l'À© ha estat tabulada, i les altres funcions de distribuciÀ³ normals en sÀ³n simples transformacions. Per tant, un pot emprar valors tabulats de la funciÀ³ de distribuciÀ³ normal tipificada per a trobar el valor de la funciÀ³ de distribuciÀ³ de qualsevol altre distribuciÀ³ normal.

Observant la grÀfica de la campana de Gauss, si l'alÀçada de la corba À©s petita podem dir que la desviaciÀ³ tÀpica À©s gran i que per tant la mostra no À©s significativa. Aleshores podem afirmar que la mostra À©s significativa quan l'alÀçada de la campana de Gauss sigui considerable.

- *DistribuciÀ³ Uniforme Discreta :*

Una variable aleatÀria discreta X que pren valors enters $1, 2, \dots, n$ amb probabilitats: $P(X=k)=1/n$; $k=1, 2, \dots, n$ rep el nom de variable uniforme discreta. DepÀn d'un Ànic parÀ metre que À©s n . I la seva mitja i varianÀa son les segÀents:

La seva funciÀ³ probabilitat es descriu de la manera segÀent: