Mètodes i tècniques quantitatives en turisme

PID_00269603

M. Encarnación André Romero
Raquel Camprubí Subirana

Temps mínim de dedicació recomanat: 6 hores

Universitat Oberta de Catalunya

M. Encarnación André Romero

Professora d'Economia aplicada i membre del grup de recerca AQR-IREA (Anàlisi Quantitativa Regional-Institut de Recerca en Economia Aplicada) de la Universitat de Barcelona. Durant cinc anys va ser coordinadora de l'Observatori del Turisme de Catalunya.

Raquel Camprubí Subirana

Professora Agregada de l'Àrea d'Organització d'Empreses a la Facultat de Turisme de la Universitat de Girona, i membre de l'Institut de Recerca en Turisme.

La revisió d'aquest recurs d'aprenentatge UOC ha estat coordinada per la professora: Julie Wilson (2019)

Segona edició: octubre 2019

Av. Tibidabo, 39-43, 08035 Barcelona
Realització editorial: FUOC

Índex

1. L'estadística com a eina de recerca aplicada al turisme

El turisme, com sabem i s'ha dit repetidament, és una activitat transversal, que abasta multitud d'àmbits d'activitat i sectors econòmics, però que també té conseqüències ambientals, geogràfiques, socials i culturals. És per aquest motiu que són moltes les disciplines d'estudi que es preocupen de la seva anàlisi. Realment es fa difícil arribar a abastar l'estudi del turisme en tota la seva amplitud i complexitat. En tot cas, del que no hi ha dubte és que el turisme és un fenomen humà, que s'emmarca dins del que denominem ciències socials.

Sovint parlem o sentim parlar al voltant del terme «estadística». L'estadística és una eina que ens acompanya en el dia a dia per tal d'ajudar-nos a conèixer la realitat i prendre decisions de tot tipus, tant professionals com en l'àmbit privat i domèstic. Tanmateix, aquesta intuïció i utilitat quotidiana de l'estadística té el suport de tot un cos científic i una metodologia pròpia, com s'ha exposat en l'apartat primer. Així doncs, atesa la rellevància de l'estadística com a eina per a la investigació de les ciències socials, humanes, geogràfiques o ambientals, tot seguit analitzarem breument el concepte d'estadística. La paraula «estadística» acostuma a tenir dues accepcions:

1) D'una banda es fa servir per a referir-nos a un conjunt d'informació numèrica.

2) D'altra banda, fa referència a la disciplina científica que a través del llenguatge matemàtic ens proporciona les eines per a tractar aquesta informació i extreure'n conclusions.

Per què és important l'estadística? L'entorn en el qual ens movem és ple d'incertesa i cada decisió que prenem genera unes conseqüències, que intentem mesurar i tenir en compte abans d'optar per una o altra decisió. La incertesa es fa encara més gran en l'àmbit de les ciències amb un component humà, dins de les quals es troba el turisme, ja que en aquestes intervenen elements propis dels individus, que no són presents en les ciències exactes, en la física o en la química. Dit d'altra manera, davant d'una acció o una decisió, no coneixem del cert quina serà la resposta obtinguda.

Exemple

Quan decidim endegar una campanya de promoció turística desconeixem quin efecte tindrà sobre la demanda, però sí que disposem d'altres experiències similars que ja han estat dutes a terme i la seva repercussió. Per tant, en podem recollir les dades i analitzar-les per tal d'orientar la nostra campanya i mirar d'optimitzar la nostra actuació.

L'aplicació de l'estadística en les ciències d'àmbit social implica acostumar-nos a treballar amb l'aleatorietat. Expliquem aquest concepte: parlem de fenòmens aleatoris quan fem referència a aquells dels quals no coneixem el resultat fins que no hem dut a terme una acció determinada. En l'extrem oposat es troben els fenòmens deterministes, que serien aquells el resultat dels quals és conegut abans de dur a terme l'acció. Els fenòmens deterministes són propis de les ciències exactes.

Exemple

En química sabem que sempre que barregem en les mateixes condicions dues molècules d'hidrogen amb una d'oxigen obtindrem una molècula d'aigua. Aquest seria un exemple de fenomen determinista. Per contra, no podem assegurar que una mateixa campanya de promoció turística tingui el mateix resultat si es du a terme en un mercat o en un altre. Ens trobem, per tant, davant d'un fenomen aleatori.

Així doncs, en la recerca aplicada al turisme, l'estadística constitueix l'eina imprescindible per a una correcta recollida de la informació, la seva gestió i la corresponent transformació en coneixement.

Igual que s'ha dit abans en referir-nos al mètode científic, l'aplicació de l'estadística comporta també la necessitat de seguir un mètode i fer unes passes, i incorpora tècniques que ens hauran de permetre,

- una recollida d'informació correcta, sistemàtica, ordenada i suficient,
- el seu tractament i descripció,
- l'anàlisi a partir de tècniques de causalitat, univariants, multivariants, etc.,

i que ens conduiran finalment a la interpretació i la presa de decisions a partir dels resultats obtinguts.

Com ja s'ha explicat, cadascuna d'aquestes etapes és fonamental, i cal que siguin plantejades d'una manera rigorosa i científica. Segurament més d'una vegada tots hem estat objecte d'una enquesta, o potser fins i tot hem participat en l'elaboració d'una. Doncs bé, cal sempre tenir molt present que aquesta informació que proveïm o recopilem és la matèria primera amb la qual després es treballarà. La bondat i seriositat de tot el procediment, i per tant de les conclusions que se'n desprenguin, depèn del fet que aquesta informació sigui veraç, seriosa i hagi estat recollida segons uns criteris que més endavant seran exposats. Altrament, tots els recursos destinats a aquest fi hauran estat malaguanyats i la presa de decisions estarà fonamentada sobre uns criteris sense cap fiabilitat.

1.1. El procés estadístic

Entrant ja en el que implica el procés estadístic, a continuació s'exposen algunes de les consideracions principals que caldrà que tinguem en compte a l'hora d'aplicar aquesta eina a la nostra investigació.

El primer que cal recordar, és que el procés estadístic s'inicia amb la definició del fenomen que volem estudiar o investigar, ja que aquesta qüestió determinarà tota la resta de les etapes i la metodologia triada, que caldrà desenvolupar.

Exemple

Abans d'iniciar una recerca, caldrà que definim si l'objectiu de la nostra anàlisi es conèixer el perfil dels turistes d'un municipi determinat (perfil de demanda); si volem saber el grau de congestió d'un territori (estudi de capacitat de càrrega); o si desitgem explorar les possibilitats d'atraure nous visitants d'un determinat segment (estudi de mercat), etc.

Podem dir que l'objectiu suposa el «què» d'una investigació (què volem estudiar?); per tant, s'hauran de definir també el "qui" (qui volem estudiar?) i el "com" (com obtindrem des dades?; i com les analitzarem?).

Aquests són elements que requereixen una atenció especial, i com hem dit abans, treballar-los amb seriositat, sistematització i rigorositat és un factor molt important, ja que tenen un paper decisiu en els resultats finals de la investigació. En els propers apartats s'explica detalladament en què consisteix cadascun d'aquests elements, i quins són els elements a tenir en consideració per a una correcta aplicació.

2. Població i mostres

2.1. La població o univers en una investigació

Un dels elements més importants en el desenvolupament d'una investigació és definir «qui» volem estudiar. Com hem vist, aquest és un elements que va directament relacionat amb l'objectiu de la recerca que es planteja. Per tant, caldrà definir el qui és la nostra població o univers d'anàlisi. Què vol dir això? La població o univers és el conjunt d'individus o elements amb unes característiques comunes sobre els quals volem saber alguna cosa; el conjunt d'individus –no necessàriament persones– sobre els quals volem obtenir un coneixement determinat i extreure'n per tant unes conclusions validades i contrastades.

Exemple

Seguint els exemples exposats en el punt anterior, les poblacions corresponents sobre les quals es vol treure conclusions serien, respectivament, tots els turistes del municipi, tots els elements de pressió deguts al turisme que hi ha a cadascun dels punts del territori de referència, ja siguin de component ambiental, social o cultural, i tots els individus que formen part del mercat en el qual estem interessats.

No es pot oblidar que, en definitiva, les conclusions que obtindrem comportaran una presa de decisions que afectaran tots i cadascun d'aquests individus. Per tant, cal establir molt clarament quina és la població objectiu de la recerca. Tanmateix, és evident que tot sovint no podrem recopilar informació exhaustiva de tots els elements de la població d'interès, per limitacions de temps, econòmiques, de recursos humans o tècniques. Per això és molt freqüent treballar amb mostres.

Una mostra és un subconjunt de la població, que ha de ser representatiu del total i ha de tenir una mida suficientment gran per a explicar el conjunt de la població. És a dir, a la mostra s'han de trobar reflectits els diversos comportaments i perfils de tots els individus que componen la població. Altrament la investigació que es dugui a terme no ens permetrà treure i extrapolar conclusions fiables per a tota la població. La construcció d'una bona mostra és per tant un element cabdal en el procés de recollida de la informació. En posteriors apartats s'exposaran quins són els criteris que cal tenir en compte per a dur a terme un bon mostreig, així com per a determinar la mida mostral.

Exemple

Volem saber l'ocupació hotelera de tota una comarca que disposa de platja i d'interior, i decidim, per limitacions de temps, enquestar tan sols una part dels establiments, és a dir, treballar amb una mostra. Si volem garantir que, en treure l'ocupació mitjana de la comarca, aquesta resulti significativa, caldrà assegurar-se que la mostra sigui significativa, i incloure-hi establiments de diferents categories, dimensions i, per descomptat, tant establiments d'interior com de la costa.

En tot cas, quan la nostra recerca impliqui la recollida d'informació exhaustiva per a tota la població, parlarem de cens. Tots hem sentit parlar per exemple del cens electoral o dels cens de població, ambdues estadístiques són d'abast universal, en el sentit que abasten tots i cadascun dels individus objecte d'anàlisi, per tant, el conjunt de la població o univers. En canvi, quan basem l'estudi en una fracció de la població, és a dir, estarem davant d'una mostra.

2.2. El mostreig: mètodes i aplicacions

2.2.1. El mostreig i els errors del procés

Com s'ha explicat, no sempre es pot estudiar la totalitat de la població. En aquest cas hem de basar l'estudi en una mostra. Doncs bé, quan parlem d'una mostra fem referència a un subconjunt dels individus d'una població.

El recurs a l'estudi d'una mostra és una opció d'investigació que és fonamental en les ciències observacionals. Quan no resulta factible estudiar tots els individus que integren una determinada població, caldrà que intentem extreure conclusions generals a partir de l'anàlisi d'uns quants casos.

En quines circumstàncies resulta adient centrar l'estudi en una mostra? Principalment, quan no és viable fer un estudi censal, per diversos motius:

- No es disposa de prou recursos humans. Enquestar tota la població requeriria un nombre d'enquestadors que no tenim al nostre abast.

- No tenim prou temps. Si la població és nombrosa ens caldria molt de temps per a fer el treball de camp i normalment és necessari disposar de la informació amb més rapidesa.

- No es disposa de prou recursos econòmics perquè l'estudi comprengui tota la població.

Ens hem d'assegurar que el mostreig, és a dir, el procediment per a l'obtenció d'una mostra, ens ha de proporcionar la col·lecció d'una mostra que permeti generalitzar els resultats obtinguts al conjunt de la població. La mostra s'obté amb la finalitat de poder inferir resultats vàlids per a una determinada població, a partir dels resultats obtinguts per a uns pocs individus (els quals componen la mostra) que pertanyen a l'esmentada població. En conseqüència, la mostra representa un subconjunt de mida manejable, que ha de ser representatiu de la població de la qual s'ha extret. Per això ens hem de plantejar dues qüestions fonamentals:

- Com s'han de seleccionar els individus que integraran la mostra?
- Quina ha de ser la grandària de la mostra?

La resposta a aquestes qüestions ens la proporcionen els mètodes de mostreig. Abans, però, de presentar les característiques dels diferents mètodes de mostreig, s'ha d'esmentar un fet que tots tenen en comú: quan es treballa amb una mostra s'ha de tenir present que aquesta, tot i haver estat dissenyada per a ser representativa de la població, no en proporciona un coneixement exacte.

Les tècniques de mostreig així com l'estadística matemàtica ens garanteixen que podrem fer una extrapolació del comportament de la població a partir de la mostra. Ara bé, s'ha d'introduir un matís: com que la mostra no ens proporciona un coneixement complet i perfecte de la població, aquesta extrapolació no serà perfecta. Hem de tenir clar, per tant, que les extrapolacions a la població no estaran mancades d'error.

Dit d'altra manera, inferir les característiques d'una població a partir d'una mostra comporta introduir un cert component d'error.

Si el disseny de la mostra és correcte, aquest component d'error disminueix a mesura que la mostra es fa més gran, i desapareix completament quan la mostra inclou tots els individus que componen la població. En tot cas intentarem aconseguir el màxim d'informació de la població, per obtenir unes estimacions de les seves característiques que siguin com més exactes millor.

Així, a partir de la informació mostral podrem aproximar-nos a quines són les característiques de la població, però no ho podrem fer amb una completa exactitud. Si no podem evitar l'error, hem d'aprendre a conviure-hi. L'hem de controlar i tenir-lo present quan intentem aproximar quines poden ser les característiques d'una població. Més endavant introduirem els conceptes relatius a l'error de mostreig.

Quan s'ha d'extreure una mostra d'una determinada població, se'n poden fer moltes seleccions alternatives. En aquest sentit, són les tècniques de mostreig les que permeten seleccionar una mostra entre els individus que integren la població. Hi ha dos grans blocs de tècniques:

- Mètodes de mostreig aleatoris o probabilístics
- Mètodes de mostreig no aleatoris o no probabilístics

Aquesta classificació es sustenta en un aspecte diferenciador clau, la probabilitat de ser escollit que té cadascun dels individus que formen part la població.

En el cas dels mètodes de mostreig probabilístics, l'obtenció de la mostra es basa en un procés de selecció aleatòria de les unitats mostrals que pren com a fonament l'atzar. Quan s'aplica una tècnica de mostreig aleatori, la selecció de les unitats mostrals es fa de manera que, abans de fer una extracció, tots els individus que pertanyen a la població tenen la mateixa probabilitat (major que zero) de ser seleccionats. Cal tenir en compte que per a poder garantir que

tots els individus tinguin la mateixa probabilitat de ser escollits, es planteja un requisit: hem de disposar d'un registre que contingui la relació completa de tots els individus de la població.

Exemple

El padró municipal, la base de dades del clients d'un establiment hoteler, el registre complet d'usuaris d'una instal·lació, etc.

En moltes ocasions, però, ens trobem que no disposem d'un registre complet de la població que desitgem estudiar, bé perquè aquesta població no es pot registrar fàcilment, bé perquè els registres existents són poc exhaustius. En aquest cas haurem de recórrer a les tècniques de mostreig no probabilístic, ja que, en no disposar d'un registre de la població, no podrem determinar quina probabilitat té cada element de la població de pertànyer a la mostra.

Quina és la rellevància de fer servir mètodes probabilístics o no? Recordem que l'objectiu d'una enquesta per mostreig és induir les característiques d'una població a partir de la informació que proporciona una mostra. En aquest context, els mètodes probabilístics són els que millor ens garanteixen la representativitat de la mostra extreta i, en conseqüència, són els més recomanables. De fet, només quan la mostra hagi estat obtinguda aplicant les tècniques de mostreig probabilístic, podrem determinar el grau de fiabilitat –i el grau d'error estimat– que tenen les conclusions a les quals s'arriba.

En resum, per a defensar quina és la validesa i fiabilitat d'un estudi, cal que la mostra sigui aleatòria. Aquestes mostres són les úniques que ens permeten obtenir conclusions de tipus probabilístic.

A continuació es descriuen alguns dels mètodes de mostreig probabilístic:

1) Mostreig aleatori simple (MAS): és la més simple de totes aquestes tècniques. Aquest procediment de mostreig probabilístic és el més senzill i conegut. D'altra banda, el MAS és un procediment bàsic com a component d'altres procediments més complexos que es presenten a continuació. El MAS es caracteritza per atorgar a tot subconjunt d'una població la mateixa probabilitat de selecció. Això és així, ja que el MAS atorga als N individus de la població la mateixa probabilitat de ser seleccionats per a integrar la mostra[1]. Aquest procediment comporta que s'hagi d'identificar un procediment de selecció aleatòria dels individus de la població que han de compondre la mostra. Aquest procés de selecció aleatòria es pot bastir de molt diverses maneres, des de les més rudimentàries (un bombo que conté boles numerades) fins a les més sofisticades (generació de nombres aleatoris utilitzant un programari adequat).

(1)D'ara endavant, quan escrivim N estarem fent referència a la mida de la població o univers d'estudi, d'acord amb la nomenclatura general que s'utilitza per a aquesta dada.

2) Mostreig aleatori sistemàtic: quan s'estudien poblacions que són massa nombroses, s'acostuma a utilitzar una variant del MAS que, conservant l'aleatorietat, simplifica el procés de selecció de la mostra. Es tracta del mostreig aleatori sistemàtic. Aquest mètode s'utilitza comunament quan volem

(2)També cal familiaritzar-se amb la notació «n» com a mida de la mostra.

estudiar una població relativament gran i els seus membres estan enumerats individualment en algun lloc central, com una guia telefònica, un directori d'estudiants, una llista de votants inscrits, un índex, una agenda o un registre de membres. Aquesta tècnica de mostreig només requereix l'extracció d'un nombre aleatori i permet prescindir de la numeració prèvia dels elements de la població. A partir d'aquest nombre aleatori, que anomenarem «a», coneixent N i també la mida de la mostra que volem obtenir, n^2, el que es fa és calcular el rati k (k=N/n, que no és més que el nombre d'elements de la població –N– que equival a cada element de la mostra –n–), i a partir de la primera observació seleccionada, a, les següents sorgeixen directament com a resultat de calcular a+k; a+2k; a+3k... Aquesta manera de seleccionar la mostra evita haver de numerar prèviament la població, ja que permet identificar la mostra directament sobre la llista.

Exemple

Imaginem que tenim el registre de visitants d'un parc natural. Tenim 12 visitants i volem escollir-ne a l'atzar 3. En aquest cas el coeficient d'elevació seria k=12/3=4, de manera que sabem que hem d'escollir un de cada quatre clients. Ara establirem el punt de partida. Imaginem que de manera aleatòria obtenim que aquest serà el segon individu de la nostra llista de visitants (a=segona observació). Així, la mostra de tres persones estaria integrada per les que es situen a la segona posició, la sisena (a+k=2+4=6) i la desena (a+2k=10).

3) Mostreig aleatori estratificat (MAE): com s'ha vist, el mostregi aleatori simple és una tècnica que garanteix que tots els individus de la població tenen la mateixa probabilitat de pertànyer a la mostra. Ara bé, aquest mètode no controla les característiques dels individus que componen la població i, en molts casos, caldrà fer-ho per a obtenir una mostra representativa de la població; és a dir, caldrà que la mostra hagi estat escollida controlant algunes característiques rellevants per a l'estudi dels individus que componen la població. El MAE s'aplica quan els elements de la població es divideixen en subpoblacions o classes, denominades estrats, que s'han de controlar en el procés de selecció de la mostra per tal que aquesta no perdi representativitat.

4) Mostreig aleatori per conglomerats: és una tècnica que resulta adient quan la població es troba dividida, de manera natural, en grups que se suposa que contenen tota la variabilitat de la població. En aquest cas, quan es pot concloure que aquests grups, els conglomerats, constitueixen una bona representació respecte a la població objectiu, podem seleccionar-ne només alguns per a fer l'estudi. Aquesta manera de procedir implica que la unitat mostral no és un individu de la població sinó un grup d'individus, els que constitueixen el conglomerat.

En tot cas, i com s'ha exposat anteriorment, quan es fa una enquesta per mostreig existeixen diversos tipus d'errors que es poden introduir. Alguns són evitables i altres no. A continuació s'expliquen quines són les possibles fonts d'error.

Hi ha dos fonts principals d'error. D'una banda tenim tots aquells errors que es poden introduir quan es fa l'enquesta. Aquests errors es poden evitar fent un disseny correcte de l'enquesta, del seu planejament i de la seva aplicació. No obstant això, s'ha de tenir present que poden ser introduïts pels diversos agents que hi intervenen. És a dir, poden ser introduïts tant pels investigadors quan dissenyen l'enquesta, com pels enquestadors, pels enquestats i pels encarregats de processar les dades. Per això en cada cas caldrà procediments i rutines de control diferents.

Entre els diferents errors, mereix un especial comentari el que té el seu origen en la manca de resposta o el rebuig a participar en l'enquesta. Quan no es pot obtenir la informació desitjada, bé pel rebuig de la persona seleccionada per a formar part de la mostra o bé per no poder-la localitzar, es produeix un doble efecte negatiu: d'una banda, es minva la grandària de la mostra i d'una altra, poden aparèixer biaixos. Tot i que el primer problema es pot solucionar substituint les persones que s'havien d'enquestar, aquest fet no elimina la possibilitat d'introduir biaixos. Per exemple, pot ser que qui refusa fer l'enquesta pertanyi a un col·lectiu específic que té un perfil diferenciat de la resta. Per això, en substituir aquests individus podem estar deixant d'estudiar un subgrup més o menys nombrós de persones que tenen un comportament semblant i diferencial respecte dels que accepten fer l'enquesta. Igualment, tenim els errors que es poden introduir en seleccionar una mostra. Entre aquests hi ha els que sorgeixen quan no s'apliquen correctament les tècniques de mostreig. Aquests errors es poden evitar tenint cura en el procés de selecció de la mostra.

Finalment, tenim els errors aleatoris de mostreig. Com ja hem comentat, aquests són inherents als procediments de mostreig. Per això es recomana utilitzar les tècniques de mostreig aleatori o probabilístic, ja que en aquestes l'error de mostreig es troba lligat a l'atzar que guia la selecció de la mostra, cosa que permet estimar l'error que es pot cometre i determinar la fiabilitat de les conclusions obtingudes. En l'explicació de com es determina la mida de la mostra s'exposa la manera com introduïm en el procés el grau d'error acceptat o el nivell de confiança en els resultats assolits.

2.2.2. Determinació de la mida mostral

Quan planifiquem una enquesta per mostreig, sempre arribem a l'etapa en què hem decidir la mida de la mostra. Aquesta és una decisió clau per a assolir els objectius de l'estudi. Una mostra massa gran comportaria malbaratament de recursos, mentre que una de massa petita minvaria la utilitat de les dades obtingudes. La teoria del mostreig, basada en l'estadística matemàtica i la teoria de les probabilitats, ens proporciona el marc teòric per a poder identificar quina es la grandària de la mostra que s'ajusta a les condicions establertes per l'investigador.

Així, per a determinar quin és el nombre d'unitats que han de compondre la mostra, haurem de prendre decisions que fan referència a:

1) La precisió que pretenem per a l'estudi que hem de dur a terme. És a dir, hem d'establir quins són els límits d'error esperats o tolerables, els que estem disposats a acceptar.

Exemple

Si estem estudiant el grau de satisfacció d'un visitant a un espai natural, a partir d'una valoració numèrica de l'1 al 10, i de la mostra resultant obtenim que la mitjana se situa en el 6, acceptar un error del 10% significa que, encara que la mostra ens hagi donat un valor concret (en aquest cas el 6), acceptem que la dada real del conjunt de la població es pot situar en un interval que s'allunyarà, com a màxim, en un 10% del valor mostral. Cal tenir en compte que aquesta discrepància pot ser per excés (si la dada mostral sobreestima la dada poblacional) o per defecte (si la dada de la mostra subestima la de la població). Un marge d'error «acceptable» del 10% sobre 6 significa 0,6. Si aquest 0,6 el dividim entre dos per a poder establir un interval d'error que inclogui possibles desviacions per dalt o per baix, tenim un 0,3. Així doncs, i a tall d'exemple, podríem dir que donat un resultat mostral igual a 6, si acceptem un error del 10%, la dada real de la població se situaria entre el 5,7 (6-0,3) i el 6,3 (6+0,3).

2) El nivell de confiança, o seguretat amb què es generalitzaran a la població els resultats obtinguts a la mostra. Ja hem comentat que no es podrà estar completament segur que la dada obtinguda tindrà la precisió desitjada pel que fa a la població. La seguretat absoluta tan sols s'obté quan s'estudia la població íntegra i, en aquest cas, la precisió també seria l'òptima, en la mesura que l'error de mostreig seria nul. Cal distingir el nivell de confiança de l'error acceptat, explicat en el punt anterior.

Exemple

Quan triem una mostra, tenim tota una sèrie de tècniques que utilitzem per mirar de garantir que la mostra serà representativa. Malgrat tot, el component d'atzar no ens assegura que, per casualitat, triem una mostra no adequada. Si aquest fos el cas, totes les conclusions que obtinguéssim no serien vàlides per a ser extrapolades a la població. Quan fixem un nivell de confiança estem dient amb quin percentatge ens refiem que la mostra és vàlida. Per tant, si treballem amb un nivell de confiança, per exemple, del 95%, i una discrepància o error acceptat del 10% –seguint l'exemple del punt anterior–, el que estem dient és que la dada de grau mitjà de satisfacció per a la mostra és igual a 6, i a la població se situaria en un rang d'entre 5,7 i 6,3; però que això només serà vàlid si la nostra mostra és bona, cosa que té un 95% de probabilitats de complir-se. Dit d'altra manera, estem avisant que hi ha un 5% de probabilitats que aquests resultats no tinguin cap validesa en relació amb la població estudiada.

3) La uniformitat en el comportament de la població enfront de la seva variabilitat: si en una població tots els individus donen la mateixa resposta, convindrem que n'hi haurà prou de prendre una mostra de mida n=1. En canvi, com més comportaments dispars hi hagi dins la població, més gran serà la mostra que necessitarem, per tal de captar aquesta diversitat i que realment sigui representativa de tota la població. Aquí es planteja un problema afegit. Si estem treballant amb mostres, és precisament perquè no coneixem la població. Per tant, és probable que no sapiguem el nivell de diversitat i variabilitat (dispersió) del seu comportament. En aquest cas, la solució més freqüent per tal de minimitzar els possibles errors comesos, és la de treballar sota el supòsit que la població presenta una dispersió màxima en el seu comportament. És el supòsit de màxima folgança implica que, en estudiar si un individu presenta

o no una característica, suposarem que la probabilitat *a priori* que la presenti és del 50% (0,5). Si ens adonem, aquesta probabilitat del 50% és la de màxima incertesa; de fet, és la que ens trobem per exemple quan llancem una moneda: si no hi ha cap informació que ens inclini a apostar per la cara o la creu (com el fet de saber que la moneda està trucada), ens trobem davant una situació de màxima folgança i incertesa. Doncs bé, aquest és el supòsit que habitualment farem servir per a calcular la mida òptima de la nostra mostra.

Tenint en compte aquests factors que l'investigador pot establir, la formula que s'utilitza comunament per a determinar la grandària de la mostra és:

$$n = \frac{Z^2 \alpha/2 PQ}{e^2}$$

On:

- el nivell de confiança prefixat dóna lloc a un coeficient Za/2, que sorgeix de l'aplicació de conceptes de probabilitat i estadística. Així, per a una confiança del 95%, el coeficient pren el valor Za/2= 1,96; per a una del 95,5%, Za/2 = 2; mentre que per a una confiança del 99%, passa a ser Za/2 = 2,58;

- la quantitat d'error o discrepància que es considera tolerable (e), és a dir, la precisió que es desitja per a l'estudi, s'estableix en termes percentuals, tot i que a la fórmula s'introdueix en tant per u. Per exemple, si l'error màxim tolerat és de 3%, a la fórmula el valor que hi donarem serà de e = 0,03;

- finalment, com s'ha assenyalat, pel que fa a P*Q (nivell de dispersió o variabilitat en el comportament de la població), introduirem la hipòtesi de màxima folgança, en funció de la qual P=Q=0,5.

Exemple

Suposem que volem fer un estudi per a determinar la proporció de visitants rebuts en una destinació atrets per aspectes relatius a la sostenibilitat. Calculem la mida de la mostra, és a dir, n.

- Haurem de decidir amb quin grau de precisió volem conèixer el percentatge assenyalat. Per exemple, tenint en compte altres estudis que han estat fets abans en aquest àmbit, podríem considerar adient que l'error de mostreig no fos més gran de ±5% (10% total de discrepància). Per tant, e = 0,05,

- Igualment haurem de fixar el nivell de confiança, o de seguretat. Suposem que el fixem en el 95%, i per tant el valor de Za/2= 1,96.

- A més treballarem amb el supòsit de màxima incertesa, folgança o dispersió: P = Q = 0,5.

Si fem el càlcul, ens resultarà que, per als paràmetres establerts, necessitarem fer 384 enquestes.

3. Les variables estadístiques

Un altre concepte rellevant en qualsevol procés estadístic és definir les variables necessàries a estudiar per a poder assolir els objectius de la investigació. Així doncs, cal definir el concepte de variable.

Anomenem variable a la característica (o característiques, en aquest cas parlarem de variables) que ens interessa estudiar de cadascun dels individus. Com que cada individu presentarà comportaments o respostes diferents, per això l'anomenem variable. Evidentment, de cada individu ens pot interessar observar més d'una característica, per tant el més habitual és l'observació simultània de més d'una variable per a cada individu.

Exemple

En un estudi de perfil de demanda turística, cadascun dels turistes esdevé un individu de la població. Per la seva banda, podem estar interessats a saber qüestions com procedència, nivell d'estudis, nits d'estada, despesa, activitats que fa. Cadascuna d'aquestes característiques seria una variable a observar. Tindríem per tant la variable procedència, la variable nivell d'estudis, la variable nits d'estada, la variable despesa i la variable relativa a les activitats fetes.

Com es pot veure a l'exemple anterior, les variables poden ser de molts tipus, i segons la seva naturalesa requeriran un tractament o un altre. A continuació s'expliquen els principals tipus de variables que són objecte d'anàlisi, segons diverses classificacions:

1) Variables qualitatives enfront de variables quantitatives

Les variables qualitatives són aquelles que no s'expressen numèricament. Seguint l'anterior exemple, la procedència, el nivell d'estudis o les activitats fetes serien tres variables qualitatives.

Les variables quantitatives són aquelles que es poden expressar numèricament. A l'exemple anterior, doncs, les nits d'estada a la destinació o la despesa feta serien dades quantitatives.

2) Variables ordinals enfront de variables nominals

Les variables ordinals són aquelles en les quals es pot establir una escala d'ordenació dels seus possibles valors a partir d'algun criteri de referència. En general les variables quantitatives són totes ordinals, ja que les podem ordenar numèricament. En el cas de les variables qualitatives, n'hi ha que també són ordinals. El nivell d'estudis esmentat en l'exemple anterior seria un cas de variable qualitativa ordinal, ja que podríem ordenar les respostes de menor a major nivell d'estudis i viceversa.

Les variables nominals són aquelles que no permeten una ordenació de cap tipus. Seguint amb l'exemple, la variable de procedència o la d'activitats fetes serien exemples de variables qualitatives nominals.

3) Variables contínues enfront de variables discretes

Les variables contínues són aquelles que presenten infinits valors entre dos valors determinats. Dit d'una manera senzilla, serien les variables que admeten fraccions i decimals. La despesa seria una variable contínua, ja que entre dos possibles valors de despesa (per exemple, 50 i 51 €) podem trobar infinites respostes, a mesura que anem afegint decimals (50,000 €, 50,001 €, 50,0002 €, etc.).

Les variables discretes són aquelles que prenen valors sencers. La variable de nits d'estada en seria un exemple, ja que un turista passarà 1 nit, 2 nits, etc., però mai podrà passar 1,3 o 4,28 nits en la destinació.

4) Variables transversals o *cross-section* enfront de variables temporals o *time-series*

Les variables transversals observen una característica per a un conjunt d'individus en un determinat moment del temps, de manera estàtica. En l'exemple anterior, si decidim observar les variables per un determinat any, mes o temporada turística, totes serien variables transversals.

Les variables temporals o històriques observen l'evolució d'una característica per a un o diversos individus al llarg del temps, de manera seqüencial. En l'exemple anterior, si decidíssim observar la mitjana de despesa al llarg de diversos anys, estaríem treballant amb una variable temporal.

4. Mètodes de recollida de dades

Un altre aspecte essencial que cal tenir en consideració quan es realitza el disseny d'una investigació són els tipus de mètodes que tenim per a recollir la informació, és a dir, com es farà la recopilació de les dades concretes relatives als individus (població o mostra) sobre els quals centrem la nostra investigació, així com les característiques (variables) en les quals estem interessats. Sobretot, sigui quin sigui el mètode escollit per al treball de camp o la recollida d'informació empírica, sempre cal recordar que la recollida de la informació condicionarà la bondat de tot el procés, ja que les dades són la matèria primera amb la qual treballarem.

Com que més endavant s'inclouen apartats específics sobre aquest aspecte, ara es tracta de començar a familiaritzar-nos amb alguns dels principals mètodes. Hi ha diverses tècniques per a recollir informació, en funció de les seves característiques i la finalitat de la investigació, així com de la disponibilitat de recursos. De totes les tècniques possibles, la que més ens interessa per l'àmbit en què ens movem, i per ser probablement la més utilitzada per a la recerca quantitativa, és l'enquesta. Tanmateix no és l'única.

1) Enquesta: com ja s'ha dit, és la tècnica mes freqüent en la recerca quantitativa i permet obtenir informació tant quantitativa com qualitativa. L'enquesta a partir d'una mostra és efectivament una de les metodologies més emprades per a la investigació en les ciències socials i, més particularment, en l'àmbit del turisme. Atesa aquesta rellevància, hi dedicarem un apartat específic una mica més endavant.

2) Observació directa: aquesta tècnica consisteix a recollir informació a partir d'observar un determinat fenomen que pot fer referència a les característiques, comportament, conducta o actitud de l'objecte de l'anàlisi. De fet, aquesta tècnica té els orígens en l'anomenada etnografia o observació participació en el camp de l'antropologia, la qual és una tècnica que recull dades qualitatives. No obstant això, l'observació directa també pot recollir dades quantitatives quan s'estableix una observació estructurada o sistemàtica. Aquesta tècnica pot fer-se manualment; és a dir, l'investigador pot observar i fer el registre de les observacions; o bé es poden utilitzar aparells que permeten automatitzar el procés com ara comptadors automàtics, GPS, fotografia aèria, etc. Així doncs, aquesta tècnica es pot utilitzar per a estimar el nivell d'ús d'un determinat espai o equipament, o el patró de comportament dels seus usuaris (circuit, temps d'ús, etc.).

4.1. L'enquesta com a eina per a la recollida de dades

L'enquesta és una tècnica d'investigació que et facilita la recollida de dades primàries, les quals després de ser analitzades et permetran obtenir la informació que desitgis sobre les diverses dimensions del tema objecte d'investigació.

Es tracta d'un mètode d'investigació interrogatiu que, prenent com a suport un qüestionari, permet recollir les respostes donades pels individus que constitueixen les unitats d'observació d'una investigació. Com es veurà més endavant, el qüestionari és l'eina que ens permet introduir les preguntes que volem plantejar a la persona enquestada, i per tant, dependrà de la seva correcta elaboració l'obtenció de les dades d'interès. L'enquesta, quan pren com a suport un qüestionari estructurat, es repeteix tants cops com unitats d'observació o individus que cal enquestar haguem determinat, tot garantint que les preguntes es formularan de la mateixa manera, i en el mateix ordre, a tots els enquestats.

Potser el principal avantatge que presenta l'enquesta en el cas de la investigació en l'àmbit del turisme és la seva capacitat per a recollir una gran diversitat de dades, corresponents a diverses variables –de diferents tipologies i naturalesa– i individus alhora. Tenint en compte la transversalitat del turisme, la tècnica de l'enquesta permet obtenir dades que reflecteixen fets, actituds, interessos, motivacions, opinions, valoracions, etc. Aquesta versatilitat, el fet que és una tècnica d'investigació emprada per a estudis d'àmbits molt diversos, explica la seva àmplia difusió en el camp de la recerca turística.

N'hi ha prou de fer una breu cerca per internet per a poder comprovar que el turisme, en els seus diferents aspectes, s'estudia molt sovint mitjançant enquestes. Així, podem trobar estudis en què l'anàlisi del turisme pren en consideració la residència habitual del visitant i la destinació del viatge, distingint entre turisme interior (intern i receptor) i turisme emissor. Per exemple, algunes de les enquestes que s'elaboren per a estudiar les característiques del turisme interior són les de l'IET de *Movimientos turísticos en fronteras* (Frontur), en què s'estudien els visitants estrangers, o l'*Encuesta de gasto turístico* (Egatur), que estudia la despesa també dels visitants estrangers. D'altres analitzen el turisme interior i l'emissor, com l'estudi de Movimientos Turísticos de los Españoles (Familitur).

D'igual manera, com ja s'ha explicat a les unitats precedents, hi ha estudis en què s'analitza el turisme des de la perspectiva de l'oferta i altres que ho fan des de l'enfocament de la demanda. Alguns estudis en què l'enquesta recull la perspectiva de l'oferta són l'*Encuesta de ocupación hotelera* i l'*Encuesta de ocupación en acampamentos turísticos*, ambdues de l'INE, o l'*Enquesta a directors d'hotel de Turisme de Barcelona*. Des del punt de vista de la demanda, alguns

exemples serien l'*Enquesta a turistes que visiten la ciutat de Turisme de Barcelona* o l'*Estudi de quantificació del nombre de visitants a la ciutat de Barcelona* del Pla estratègic de turisme de la ciutat de Barcelona.

Finalment, les enquestes són un instrument força útil perquè les destinacions i les seves entitats o organismes gestors puguin obtenir la informació necessària per a elaborar estadístiques sobre l'ocupació als establiments on s'allotgen els turistes, la valoració i satisfacció dels visitants en relació amb la destinació, el perfil dels visitants: procedència, mitjans de transport, motivació del viatge, etc., i, per descomptat, tots aquells aspectes, ja siguin des del vessant d'oferta o demanda, relatius a la sostenibilitat i els seus diferents vectors.

4.1.1. Aspectes clau en l'aplicació de la metodologia de l'enquesta

Tanmateix, i encara que les enquestes poden ser una eina molt útil, quan es decideix dur a terme una enquesta per a resoldre una investigació, cal tenir en compte que l'aplicació d'aquesta tècnica no està mancada de dificultats, que s'han de mirar d'evitar o resoldre en el seu procés inicial. Així doncs, cal procedir amb molta cura a l'hora de dissenyar la investigació, per tal de no patir –o minimitzar– els perjudicis que es deriven de la renúncia de la persona enquestada a subministrar la informació que es vol obtenir –cosa tot sovint provocada per la forma o moment en què es planteja la pregunta–, de la incapacitat de la persona enquestada per a aportar la informació requerida (bé perquè no se'n recorda, no és capaç de discriminar entre les diferents alternatives de resposta que es plantegen, etc.), o bé, quan un plantejament poc adequat de les preguntes recollides al qüestionari pot influenciar les respostes. Existeixen pautes per a resoldre aquests i altres inconvenients propis de les enquestes. Tot i així, cal insistir que la cura i detall en l'elaboració de tot el procés, en la fase prèvia a l'execució de l'enquesta, esdevenen fonamentals per a obtenir-ne bons resultats.

4.1.2. Etapes per a la correcta aplicació de la recerca per mitjà d'enquestes

Quan un estudi es basa en les dades obtingudes mitjançant una enquesta, es poden identificar uns grans blocs de tasques que cal fer de manera seqüencial. Dit d'altra manera, el disseny d'una enquesta comporta seguir una sèrie d'etapes, totes importants per a assolir un plantejament correcte de l'estudi que permeti obtenir la informació que ens interessa.

1) En primer lloc, cal insistir en la importància de reflexionar sobre l'existència d'una necessitat objectiva d'informació que justifiqui l'estudi. És a dir, cal plantejar-se si aquesta informació no es pot obtenir a partir d'altres fonts de informació que ja es troben disponibles i si l'enquesta és la tècnica més adient per a l'objectiu d'investigació. En definitiva, cal fixar bé què aportarà el nou estudi i quina utilitat tindrà. Una manera de procedir recurrent i força reco-

manable és la d'explorar les solucions que han estat emprades per a recerques similars en altres indrets o en el cas d'altres temàtiques amb les quals es puguin establir paral·lelismes. Una cerca correcta de documentació pot aportar informació útil per a l'adopció de la decisió sobre la metodologia que convé utilitzar en la investigació. Amb aquesta revisió, doncs, es minimitza el risc d'utilitzar metodologies poc adequades per a l'objectiu fixat.

2) Una vegada constatat que existeix una necessitat objectiva d'informació que ha de ser coberta, cal delimitar amb precisió quin són els objectius que s'han d'assolir amb les enquestes. Recordem que, en exposar el mètode científic i de recerca, ja hem explicat la importància de relacionar, des del principi, temàtica amb objectius, hipòtesis i metodologia. En aquest cas, el repàs dels objectius i hipòtesis de la investigació són clau per a enfocar l'enquesta.

Cal dir que en general és recomanable una enquesta amb pocs objectius i clarament definits, ja que les enquestes molt ambicioses acaben generant confusió de conceptes i idees (complicant el qüestionari i en conseqüència l'obtenció de resposta i, a més, una resposta vàlida). Tot i així, també és cert que de vegades estudis de gran interès potencial no han reeixit per no incloure una o dues preguntes més al qüestionari, que haurien resultat molt reveladores pel que fa a relacions i causalitats. En resum, es tracta de no incloure preguntes supèrflues i alhora d'evitar el risc de passar per alt dimensions del turisme importants per al nostre estudi. Així doncs cal trobar el punt exacte d'equilibri, tasca complexa en la qual cal invertir el temps i els recursos necessaris.

També aquí recordem, com s'ha explicat, que tot sovint val la pena aplicar tècniques d'anàlisi qualitativa, que ens poden ajudar a acabar dissenyant una enquesta completa i suficient. Les sessions de *brainstorming*, les entrevistes, la dinàmica de grups, l'observació participativa i/o distant, són algunes d'aquestes tècniques.

Finalment, en aquesta etapa cal tenir present que en el procés d'investigació s'utilitzarà tot un seguit de conceptes i operacions que requereixen una definició explícita, com s'ha dit en un apartat anterior. És, doncs, també el moment d'establir amb exactitud què s'entén per cadascun dels conceptes explicitats, o, per exemple, quines són les variables i unitats amb què les voldrem mesurar (per exemple, si parlem de contaminació caldria explicitar a què fem referència i com la voldrem mesurar: si en emissions a l'atmosfera, si en generació de residus, etc.).

3) Caldrà decidir si l'enquesta serà de caràcter censal (quan s'estudien tots els individus que formen la població o univers), o bé si s'ha de recórrer a la utilització de les tècniques estadístiques de mostreig, per tal d'escollir un subconjunt d'individus que sigui representatiu del col·lectiu objecte d'estudi. En

aquesta decisió, a banda de qüestions estrictament metodològiques, per descomptat que hi intervenen aquelles de caràcter restrictiu, és a dir, les limitacions de recursos.

4) Ara entraríem en la fase d'elaboració del qüestionari. En la seva construcció s'han de tenir en compte factors com ara les característiques dels individus que es vol analitzar, les variables sobre les quals es vol obtenir informació, el mètode d'administració del qüestionari, el calendari, la forma i el lloc o llocs de les enquestes, el pla d'explotació posterior, etc. Tots aquests aspectes, i d'altres, seran tractats més endavant. Cal avançar, però, que es recomana sempre, i abans de passar a la següent etapa, dur a terme una prova pilot del qüestionari, a fi de detectar possibles errors o imprecisions i corregir-los adequadament.

5) A continuació haurem de planificar i dur a terme tota una sèrie de tasques que és el que coneixem com a treball de camp i que és, de fet, la recollida de les dades. És evident, encara que no ens hi hem referit fins ara, que aquest conjunt de tasques ha hagut de ser pensat i previst amb antelació, ja que determina aspectes diversos i fonamentals de la investigació com el calendari de disponibilitat de les dades o els costos que es generaran. Entre aquestes tasques destaquen la fixació del calendari d'enquestes, la formació dels enquestadors, el control de l'operació, la coordinació de la recollida de les dades, etc.

6) Un cop disposem dels qüestionaris emplenats, haurem de confeccionar la base de dades i tot seguit depurar-la d'errors i inconsistències. Cal que decidim quin programari emprarem, quin serà el contingut de la base de dades, en quina forma s'han de disposar i codificar les dades, quines instruccions s'han de donar per tutelar el procés d'introducció de dades, quins mètodes han de fer possible la detecció de possibles errors introduïts, etc.

7) Finalment estarem en disposició de dur a terme l'explotació estadística de les dades (anàlisi descriptiva i explicativa) que ha de conduir a l'obtenció de resultats concloents.

4.1.3. Disseny del qüestionari i redacció de les preguntes

Pel que fa a l'elaboració del qüestionari, cal tenir present que aquest constitueix una veritable base de coneixement, ja que compleix una funció d'enllaç entre els objectius de la investigació i la realitat de la població. En el qüestionari s'han de traduir les necessitats d'informació en preguntes concretes, que no són més que l'expressió en forma interrogativa de les variables (o característiques) que es volen estudiar. Els qüestionaris que s'utilitzen amb més freqüència en la investigació turística són els qüestionaris estructurats, és a dir, els que:

- estableixen una seqüència ordinal en la presentació de les preguntes,
- possibiliten que totes les preguntes es plantegin de la mateixa manera,

- es poden utilitzar com a formulari per a consignar les respostes rebudes.

En relació amb l'enunciat o redacció de les preguntes d'un qüestionari, cal que tenir present que aquestes han de ser formulades de manera que donin lloc a respostes sinceres, clares i exactes. En aquest sentit, abans de redactar les preguntes, s'han de considerar les circumstàncies que defineixen el context en el qual es farà l'enquesta. Per exemple cal tenir present:

- Les característiques sociodemogràfiques de la població que es vol estudiar. Per exemple, el seu nivell de formació condicionarà el llenguatge que cal emprar i la complexitat de les preguntes. Per descomptat també en determinarà l'idioma o idiomes.

- El mitjà escollit per a administrar el qüestionari, que, al seu torn, determinarà els mitjans físics auxiliars que es poden emprar. Per exemple, si l'enquesta es fa per telèfon, no es podrà utilitzar cap suport (en paper o digital) on l'enquestat pugui llegir les opcions de resposta i escollir la que li sembli més adient.

- Finalment, també cal pensar que el tipus de preguntes que s'inclouran al qüestionari dependrà de l'anàlisi estadística que es vulgui fer *a posteriori* amb les dades resultants. Efectivament, segons com siguin formulades les preguntes, permetran un tractament estadístic o un altre.

Exemple

Si demanem als turistes d'una zona protegida la seva valoració pel que fa a les accions per a preservació de l'entorn, podem optar per un plantejament qualitatiu (amb respostes de tipus semàntic, com ara «molt bona, bona, normal, dolenta, molt dolenta»), però hem de ser conscients que això, pel que fa a càlculs, tan sols ens permetrà establir percentatges i poc més. Per contra, la pregunta es pot plantejar demanant una valoració numèrica o quantitativa dins d'una escala determinada (per exemple, del 0 al 10); en aquest cas sí que es podran extreure indicadors quantitatius com la mitjana o la desviació estàndard.

Un cop fetes aquestes consideracions generals, a continuació es descriuen les etapes o pautes que cal seguir per a la construcció del qüestionari.

1) Cal definir quines són les grans àrees temàtiques que volem cobrir amb l'estudi.

2) Cal acotar quines són les característiques (variables) dels individus que volem conèixer, en el context de cadascuna de les àrees temàtiques. Com es pot comprovar a la taula anterior, a cadascun dels blocs s'inclouen aquelles variables sobre les quals es vol obtenir informació.

Exemple

A la taula següent es descriuen les àrees temàtiques i variables utilitzades en l'enquesta *Viatges dels catalans* de l'IDESCAT.

Àrees temàtiques definides	Ítems a cadascuna de les àrees (variables)
Composició de la llar	
Viatges a segona residència	Disposa de segona residència? Hi ha pernoctat en el mes de referència? Nombre de viatges fets en el mes de referència Municipi/s de destinació Dates de sortida i de retorn
Altres viatges	Nombre de viatges de no segona residència fets en el mes de referència Municipi/s de destinació Dates de sortida i retorn Motiu del viatge
Descripció dels «altres viatges»	Contractació del viatge Forma de viatjar (individual, familiar, en grup...) Mitjà de transport Tipus d'allotjament Grau de satisfacció Activitats fetes Grau de fidelitat a la destinació Despesa declarada
Dades sociodemogràfiques	Sexe Estat civil Nivell d'estudis Activitat professional

Font: Idescat. Enquesta dels viatges dels catalans

3) S'haurà de determinar quin tipus de preguntes es faran servir i, d'acord amb l'elecció, redactar-ne l'enunciat. Més endavant s'explicaran aquestes dues qüestions.

4) Cal decidir quina és la seqüència adequada per a presentar les preguntes que fan referència a diversos aspectes d'un mateix tema (el que s'anomena bateria de preguntes), establint filtres que evitin plantejar determinades preguntes als enquestats que no han de respondre-les, o introduint preguntes de control per a contrastar la validesa de les respostes dels enquestats.

5) Finalment, un cop elaborat el primer esborrany del qüestionari, en caldrà determinar els aspectes formals. Per exemple, el tipus i el cos de la lletra, l'espaiat interlineal, la utilització de negreta, cursiva o subratllats per a destacar i diferenciar parts del text, etc.

En relació amb les preguntes del qüestionari, també cal fer algunes consideracions importants:

- Les preguntes poden ser de diferent tipus en funció de la resposta que admeten de la persona enquestada:
  - Tancades: són aquelles la resposta de les quals ha d'encaixar en una categoria prèviament establerta. Són més fàcils de respondre per la persona enquestada, que simplement ha d'escollir la categoria de resposta amb la qual s'identifica. Permeten codificar les respostes *a priori* i, en

conseqüència, faciliten el procés de construcció de la base de dades resultant. No obstant això, constrenyen la resposta, cosa que introdueix el risc de perdre importants dimensions quan aquestes no s'han previst com a categories de resposta. És per aquest motiu que tot sovint s'inclou com una de les possibles respostes la modalitat «altres». Mentre que aquesta categoria tingui un pes testimonial, no suposa un problema. Ara bé, si un percentatge significatiu de les respostes l'escull com a resposta, aquest fet ens indicarà que hem oblidat alguna categoria de resposta rellevant. En general, quan s'inclou la resposta «altres», s'anota quina és la resposta concreta de l'individu enquestat. De fet, de les preguntes en què es deixa oberta una part de les respostes en diem preguntes parcialment obertes.

– Obertes: permeten que la persona enquestada respongui amb plena llibertat a la pregunta que se li planteja. Proporcionen una informació més subtil i rica en matisos que les preguntes tancades. Resulten indispensables per a estudis exploratoris i són útils també perquè no sempre podem predeterminar les alternatives de resposta que configuren una pregunta tancada. Tanmateix presenten també inconvenients, ja que poden donar lloc a respostes incompletes o difícils de comprendre, i el seu tractament posterior és força complex i laboriós. Cadascuna de les respostes obertes s'ha de considerar per separat i després anar agrupant totes les que tenen un significat similar, o fan referència a un mateix àmbit temàtic, a fi de poder fer l'oportú recompte per categories. Donats aquests inconvenients és preferible l'ús de preguntes tancades, sempre que sigui possible i les respostes siguin de tipus qualitatiu. Ara bé, imagineu-vos que voleu preguntar l'edat. En aquest cas és preferible formular la pregunta com una pregunta oberta, ja que la resposta permet un tractament de dades més complet i complex a nivell estadístic, que en el cas que es reculli la dada a través d'intervals.

- Una pregunta s'ha formulat correctament quan no exerceix cap mena d'influència en el sentit de la resposta i no condueix a una resposta inexacta.

- Sempre cal plantejar-se reflexivament si la pregunta és realment necessària per a l'estudi. Cal evitar les preguntes que es puguin considerar supèrflues, ja que com més llarg sigui el qüestionari més serà la dificultat que comportarà la seva administració, i a més, com més preguntes s'incloguin al qüestionari més grans seran les reticències que trobarem a l'hora de respondre'l.

- Les preguntes s'han de redactar de manera que siguin fàcilment comprensibles per a les persones a les quals van destinades.

- S'han de formular de manera concisa i clara, evitant la introducció de dissertacions ambigües que no aporten més claredat al plantejament.

- No s'han d'utilitzar termes que puguin resultar desconeguts, o coneguts tan sols superficialment, per una part important dels entrevistats.

- Una pregunta no pot ser interpretada de diferents maneres pels enquestats.

- Les preguntes han de ser com menys complicades millor i formulades d'acord amb el perfil de l'enquestat.

- No s'han de plantejar preguntes en forma negativa ja que poden dificultar la resposta.

- S'ha d'evitar que una pregunta en realitat inclogui dues.

- Les preguntes no han d'obligar l'enquestat a fer grans càlculs o esforços de memòria. Pensem que la incapacitat per a recordar adequadament dóna lloc, per exemple, a errors per omissió, quan no es pot recordar un fet que realment va succeir; o a errors per dissipació, quan es comprimeix o expandeix el temps en recordar un fet. Així, es pot referir un fet que va succeir més enllà del període fixat, o bé excloure'n un que sí que es va produir en aquest període.

- Les preguntes s'han de fer de manera que no aixequin prejudicis.

- En cap cas les preguntes han de ser indiscretes sense necessitat.

- En el cas de preguntes sobre opinions o valoratives, com poden ser les relatives a satisfacció, podem optar per preguntes obertes, que ens dificultaran molt el seu tractament, o per preguntes tancades. Si decidim que la pregunta ha de ser tancada, tenim diverses opcions de nou. La pregunta es pot tancar proposant diverses categories que reflecteixin el grau de satisfacció del visitant, utilitzant com a suport de la resposta una escala de gradació numèrica. Les escales han de ser imparells a fi que sempre hi hagi una resposta neutra. Aquestes escales s'anomenen *escales Likert*. Altres escales que també es poden utilitzar són les anomenades *escales de diferencial semàntic*, les quals també són imparells, però en cada extrem hi ha un adjectiu oposat, que s'ha de relacionar a un concepte determinat (l'element a valorar), de manera que la valoració indicarà la proximitat i llunyania a cadascun dels adjectius inclosos en la valoració.

- A més de proporcionar les dades que volem obtenir, algunes preguntes assoleixen una funció addicional en el qüestionari:
  - Preguntes filtre: són les que, en funció de la resposta que s'hi doni, discriminen per quin ítem del qüestionari s'ha de continuar l'enquesta. Resulten imprescindibles quan el qüestionari conté preguntes que no poden aplicar-se a tots els individus enquestats.

Exemple de preguntes filtre

Imaginem un qüestionari genèric adreçat a tots els visitants d'una destinació per mitjà del qual volem destriar, en primer lloc, quins d'aquests visitants han triat la destinació per qüestions relacionades amb la seva sostenibilitat i, un cop identificats, volem que estrictament aquests ens responguin una bateria de preguntes relacionades amb el tema. Doncs bé, hauríem de procedir de la següent manera:

Pregunta núm. X: Quin és el motiu principal pel qual ha triat aquesta destinació?

1. Recomanació de coneguts (salti a la pregunta Y)
2. Reportatges/documentals (salti a la pregunta Y)
3. Informació a la xarxa (salti a la pregunta Y)
4. Proposta de l'agent de viatges (salti a la pregunta Y)
5. Per l'aposta per la sostenibilitat (continuï a la següent pregunta)
6. Preu (salti a la pregunta Y)
7. Altres (.....) (salti a la pregunta Y)

Pregunta núm. X+1 i següents –fins a la pregunta Y–: només adreçades als que a l'anterior pregunta han respost que el motiu principal pel qual havien triat la destinació era la seva aposta per la sostenibilitat.

  - Preguntes de control: són les que tenen com a finalitat el control de la veracitat i la fiabilitat de les respostes. Les preguntes de control també es poden utilitzar per a contrastar la consistència en les respostes. Per això, es redacten preguntes en les quals la idea subjacent sobre la qual s'interroga resulta ser la mateixa, i es formulen espaiades entre si per a comprovar si les respostes recollides resulten congruents. Una altra possibilitat que plantegen les preguntes de control és la d'incloure entre les categories de resposta algunes que no són possibles. D'aquesta manera, es pot dubtar de la veracitat de les respostes de qui ha escollit una d'aquestes opcions.

- L'ordre amb el qual apareixen les preguntes al qüestionari no és arbitrari. El bon funcionament del qüestionari depèn d'haver fet una ordenació adequada. Es poden fer alguns suggeriments que orientin sobre com cal ordenar les preguntes:
  - L'agrupació de les preguntes per àrees temàtiques facilita les respostes dels enquestats. En principi, totes les preguntes en les quals es doni una concordança temàtica haurien de configurar un bloc que respectés uns criteris d'ordre temporal, lògic i psicològic. De manera general podem referir la conveniència que no es produeixin canvis temàtics bruscos i, en el cas que aquests no es puguin evitar, haurà de ser l'enquestador qui atenuï la transició perquè resulti tan suau com sigui possible. Amb aquest objectiu es poden incloure en el qüestionari frases de transició o introductòries que seran llegides per l'enquestador.

  - S'han d'evitar les ordenacions que puguin influir en algun sentit en les respostes. Aquest problema es manifesta quan es col·loquen consecutivament dues preguntes la resposta d'una de les quals pot influir en el sentit de la resposta a l'altra.

– S'ha de fer una presentació seqüencial segons el moment del temps a què es fa referència i segons l'especificitat temàtica. Les preguntes més properes en el temps es formulen primer i a continuació les preguntes poden anar fent referència a períodes més llunyans. Pel que fa al nivell d'especialització, primer es formulen les preguntes més genèriques i després les més especialitzades. Per norma general, se sol començar el qüestionari formulant preguntes senzilles i poc sensibles, a fi que els enquestats es trobin còmodes. Resulta contraproduent començar el qüestionari formulant preguntes amb un grau de dificultat elevat o que puguin fer sentir incòmode la persona que respon, ja que poden provocar el rebuig a l'enquesta. S'ha de procurar que les preguntes que podem etiquetar com a «sensibles» (com les que tenen per objecte els ingressos, el posicionament polític...) no figurin al principi del qüestionari, ja que un cop l'enquesta es troba avançada resulta més improbable que l'entrevistat es mostri intransigent a contestar aquest tipus de preguntes. Seguint aquests criteris, les preguntes relatives al perfil sociodemogràfic (edat, gènere, nivell d'ingressos, etc.), com que n'hi ha moltes que són preguntes sensibles, se situaran al final del qüestionari.

Un cop revisats alguns dels aspectes fonamentals al voltant del plantejament de les preguntes, repassarem ara algunes qüestions importants relatives al disseny del qüestionari:

- Ha de contenir un número de sèrie per a poder-lo identificar. Aquest número s'inclourà a la base de dades i permetrà fer les comprovacions necessàries. A més sempre cal saber data, horari, lloc i persona que ha dut a terme l'enquesta.

- El qüestionari ha de ser com més breu millor. És a dir, no ha de contenir més preguntes que les necessàries i ha de prescindir de totes les que es puguin considerar supèrflues.

- Cada pregunta ha d'estar numerada i la numeració ha de ser consecutiva.

- També és aconsellable la precodificació dels ítems/variables i les seves categories, als quals es poden assignar uns codis que en el qüestionari figuraran al seu costat. La precodificació simplifica les posteriors operacions de translació a la base de dades de les respostes consignades al qüestionari.

- Cal ser curós amb l'aparença del qüestionari, la qual sempre ha de facilitar el treball que s'ha de fer posteriorment.

- En alguns qüestionaris resulta imprescindible preveure un espai lliure en el qual l'enquestador pugui introduir anotacions aclaridores o observacions.

- Per poder treure un millor rendiment en l'anàlisi estadística posterior de les dades recollides, serà preferible l'ús de preguntes que permetin obtenir una resposta numèrica. Si cal, posteriorment ja es faran agrupacions en intervals per presentar els resultats.

- Al final del qüestionari s'ha d'incloure una instrucció per a l'enquestador, que li recorda que ha d'agrair a l'enquestat la seva col·laboració.

4.1.4. L'administració del qüestionari, la prova pilot i l'enregistrament de les dades

Un cop dissenyat el qüestionari, d'acord amb els diferents aspectes que s'han anat comentant, arriba el moment de la seva administració. Cal que en tot el procés s'hagi pensat sobre aquest aspecte, tant per qüestions de logística com d'altra naturalesa, com la pressupostària.

Existeixen diversos mitjans que es poden emprar per a administrar el qüestionari. Els clàssics són en persona, per telèfon o per correu postal o electrònic. Tanmateix, quan es considera el tipus de suport emprat, cal referir-se a la importància creixent que té la utilització del suport informàtic. Avui en dia la majoria de les enquestes que es fan per telèfon compten amb l'ajut de la metodologia CATI (*computer aided telephone interviewing*). De la mateixa manera, cada cop és més freqüent que els enquestadors que fan enquestes presencials ho facin amb l'ajuda de suport informàtic i telemàtic que permet transferir les dades d'una manera molt eficient i ràpida. Aquesta metodologia es sol anomenar CAPI (*computer aided personal interviewing*).

Exemple

A *l'Estadística de movimientos turísticos de los españoles* (Familitur) de l'IET, la recollida de la informació es fa per un sistema mixt en el qual es combinen les enquestes personals a domicili amb metodologia CAPI i les enquestes telefòniques assistides per ordinador (CATI).

L'elecció del mitjà d'administració no es fa després d'haver dissenyat el qüestionari. Com s'ha assenyalat, el mitjà s'ha d'escollir abans, ja que el disseny del qüestionari ha de ser adequat per al mitjà escollit. Per a poder escollir el mitjà d'administració, cal valorar els següents aspectes:

- El temps de què es disposa per a fer l'enquesta i completar l'estudi.

- La naturalesa, el volum i la complexitat de la informació desitjada.

- La dispersió geogràfica dels individus que constitueixen l'objecte de l'estudi.

- El pressupost disponible.

La consideració d'aquests elements, tot tenint en compte els pros i contres que comporta cadascun dels mitjans d'administració, ens conduirà a la selecció d'un mitjà o un altre, segons quines siguin les circumstàncies específiques de cada cas.

1) L'enquesta en persona es caracteritza pel contacte personal amb la persona enquestada. Aquestes enquestes es poden fer al domicili particular, als punts turístics, als voltants dels monuments o recursos turístics, a l'aeroport, a les estacions de ferrocarrils i autobusos, etc. Com els altres mitjans, l'enquesta en persona presenta avantatges i desavantatges. Alguns dels avantatges de l'enquesta en persona són més flexibilitat respecte als recursos que cal fer servir per a la recollida de les dades, taxes de resposta generalment més elevades, control sobre la identitat de qui respon, no queda exclòs sistemàticament de l'estudi cap col·lectiu, l'enquestador controla l'ordre en el qual es plantegen les preguntes i, també, l'ordre en què l'enquestat emet les respostes, l'enquestador pot aclarir les respostes ambigües, exclou la consulta a terceres persones o les llargues reflexions per part de l'enquestat. Com a inconvenients, és el mètode més costós, l'enquestat que es mostra predisposat a donar respostes «socialment desitjables» pot introduir un biaix en sentir-se coaccionat per la presència de l'enquestador, la posició preeminent que ocupa l'enquestador pot fer que imposi d'alguna manera les seves pròpies opinions o actituds, i la planificació i el control del treball de camp són més complicats. Si no s'estableixen criteris d'aleatorietat alhora d'interpel·lar les persones enquestades, es pot produir un biaix propiciat per l'enquestador.

2) L'enquesta per telèfon es troba àmpliament difosa als països on la majoria de llars té telèfon fix. Hem de tenir present que en aquest cas l'enquestador ha de tenir uns coneixements i habilitats diferents dels requerits per a efectuar enquestes en persona. A més, si durant molt de temps la guia telefònica va proporcionar una bona aproximació de les llars d'una àrea geogràfica, la ràpida implantació de la telefonia mòbil ha transformat aquesta realitat. Avui en dia existeix un significatiu contingent de població que no té telèfon fix a casa i, per tant, cal tenir en compte aquesta circumstància. Avantatges de l'enquesta per telèfon: rapidesa (el treball de camp es pot fer molt més ràpidament, ja que no cal que els enquestadors es desplacin físicament); no calen enquestadors sobre el terreny; aquest mètode és més interessant com més dispersa i difícil de contactar mitjançant la visita de l'enquestador sigui la població; la utilització del telèfon per a administrar el qüestionari pot resultar menys costós que fer-ho en persona; és més fàcil supervisar el treball dels enquestadors, controlant les possibles fonts d'introducció de biaixos; les taxes de resposta que se solen registrar solen ser més elevades que les obtingudes amb les enquestes per correu; la utilització del telèfon permet una relació més anònima amb l'enquestador i eleva la taxa de resposta; permet efectuar diferents temptatives de contacte en diferents moments del dia o de la setmana, cosa que afavoreix la localització d'individus amb molta mobilitat; igual que en l'enquesta en persona, els enquestadors poden demanar aclariments o informació que complementi una resposta i és possible controlar l'ordre en el qual es plantegen les preguntes i es

produeixen les respostes. Alguns dels desavantatges de l'enquesta per telèfon són que no podem accedir a les llars que no tenen telèfon fix; és possible que la trucada telefònica sigui percebuda com un preludi a una acció comercial, tret que hagi estat precedida per una carta de presentació prèvia; l'enquestat pot penjar l'auricular en el moment que ho desitgi, cosa que resulta més fàcil que expulsar del domicili a l'enquestador; el qüestionari ha de ser breu. Una enquesta per telèfon d'uns minuts és fàcil de mantenir, per sobre de 15 minuts és convenient comprovar l'acceptabilitat de l'enquesta mitjançant proves prèvies; no s'estableix un contacte visual amb l'enquestat i els enquestadors poden introduir biaixos.

3) L'enquesta per correu postal es fa servir cada vegada menys. Avui en dia, als països on la població utilitza àmpliament el correu electrònic o similars, que vol dir una àmplia implantació d'internet a les llars, sorgeixen altres suports alternatius. Així, quan les característiques de la població que es vol estudiar ho permeten, ens podem plantejar la vehiculació del qüestionari per correu electrònic, Facebook, Twiter, o qualsevol altre instrument vinculat a la xarxa. En tots aquest casos, en el moment que es dissenya el qüestionari s'ha de preveure que aquest serà autoadministrat. Els avantatges de l'enquesta per correu són que és menys costosa que l'enquesta en persona o per telèfon, facilita l'accés a poblacions allunyades i disperses, ja que no intervenen enquestadors, qui respon percep una imatge de més confidencialitat, el marc temporal per emplenar el qüestionari ha de ser més flexible i s'exclouen els possibles biaixos que pot introduir l'enquestador. Pel que fa als inconvenients de l'enquesta per correu, les taxes de resposta que es registren són generalment més febles que les obtingudes amb els altres dos mitjans ja analitzats, no resulta adequada per als individus amb un baix nivell d'instrucció, una absència sistemàtica de resposta pot introduir un biaix, les interrupcions o col·lapses del servei postal o de la xarxa poden retardar el procés de recollida de dades, no es pot controlar l'ordre de presentació de les preguntes i l'ordre de les respostes, és més difícil per als enquestats sol·licitar detalls o aclariments i no es pot controlar la identitat de qui respon.

Després d'haver dissenyat el qüestionari, decidit el mitjà d'administració i dut a terme la resta de tasques, sempre cal fer un assaig d'utilització en condicions reals per tal d'avaluar-ne la validesa. D'això se'n diu avaluació del qüestionari i consisteix a dur a terme una prova pilot, que ha d'incloure: les preguntes, el seu format, la seva seqüència, les instruccions relacionades amb l'administració del qüestionari, l'adequació del temps que ha estat previst per a emplenar-lo, etc.

Per a avaluar el qüestionari cal considerar-ne diversos aspectes. Per exemple:

- Cadascuna de les preguntes mesura la dimensió que es vol mesurar?

- Es comprenen tots els termes que s'han utilitzat?

- S'han utilitzat termes ambigus?
- S'han inclòs preguntes supèrflues?
- S'ha deixat d'incloure alguna pregunta que es considera necessària?
- En les preguntes tancades, s'han previst totes les respostes possibles?
- La carta o frase de presentació crea un clima favorable que propiciï la cooperació?
- Hi ha alguna pregunta que comporti un esforç de memòria excessiu?
- Hi ha algun aspecte del qüestionari que condicioni les respostes en un sentit determinat?
- Resulta insuficient l'espai previst per a anotar les respostes a les preguntes obertes?
- Es pot administrar el qüestionari en el temps previst *a priori*?

Per a efectuar la prova pilot, es pot recórrer a tres tipus de persones:

- Altres especialistes que no estiguin implicats en el disseny de l'enquesta, que poden fer-ne una avaluació.
- Els usuaris potencials de la informació que generarà el qüestionari, als quals es pot demanar col·laboració.
- Finalment, cal fer un simulacre, és a dir, administrar l'enquesta a diversos membres del col·lectiu objecte de l'estudi. D'aquesta manera podrem obtenir un *feedback* verbal i no verbal sobre l'adequació de les instruccions i de les preguntes.

Un cop finalitzada la prova pilot caldrà fer-ne una avaluació i realitzar les modificacions que siguin necessàries.

Una vegada es disposi del qüestionari vàlid es passarà a fer el treball de camp, és a dir, la recollida de la informació. Aquest és un nou aspecte de notable rellevància: el disseny de la base de dades i l'enregistrament de la informació.

En aquest sentit, el primer que cal és escollir quin programari s'utilitzarà com a suport de la base de dades. Un cop escollit, caldrà configurar-ne la base de dades. Amb aquest fi, s'haurà de tenir en compte el nombre de variables, la seva naturalesa, els valors o modalitats que pot prendre cadascuna d'aquestes, l'etiquetatge dels valors codificats, etc.

Cal repetir que aquesta tasca es simplifica en gran manera si el qüestionari ha estat precodificat, és a dir, si s'ha assignat un codi a cadascuna de les possibles respostes que poden ser emeses per a un determinat ítem. Aquest procés serà més complex si el qüestionari recull diverses preguntes obertes.

El següent pas, òbviament, serà l'enregistrament de les dades. Aquest procés consisteix a introduir a la base de dades les respostes consignades al qüestionari. Aquesta tasca es pot fer manualment o bé mitjançant escàners òptics o marques sensorials. En cas que el qüestionari hagi estat construït en suport informàtic (mètodes CAPI i CATI), la base de dades es genera automàticament quan es registren les respostes.

Finalment, quan ja es disposa de la base de dades, es poden fer els càlculs estadístics que ja havien estat previstos anteriorment i que ens han de permetre obtenir la informació desitjada.

5. L'anàlisi estadística de dades

L'anàlisi estadística permet transformar les dades recollides en informació i coneixement posterior, amb el darrer propòsit d'acomplir els objectius establerts en la recerca i poder validar o refutar les hipòtesis plantejades en el disseny de la investigació.

Per a poder-ho portar a terme es disposa d'un seguit de tècniques estadístiques o càlculs matemàtics que permeten poder treballar les dades i obtenir-ne conclusions; que, si s'han fet correctament, permetran explicar el comportament de la població estudiada a partir de la mostra. És a dir, els resultats obtinguts es podran inferir a la resta de la població, tenint en consideració el marge d'error existent.

5.1. L'anàlisi descriptiva: tabulació, representació gràfica i estadístiques o mesures de síntesi

Un cop es disposa de la informació estadística que es vol treballar, el primer que cal fer és dur a terme una anàlisi descriptiva d'aquesta informació. Per a fer-ho disposem de les eines de l'estadística descriptiva. En aquest apartat veurem les tres eines principals de l'estadística descriptiva: les taules de freqüències, l'anàlisi gràfica i les mesures de síntesi.

5.1.1. Taules de freqüències

Quan procedim a fer una recerca estadística, acabem obtenint, com s'ha exposat, un fitxer o base de dades, amb tot un contingent d'informació. Aquesta informació, sense processar, no ens aporta gaire llum sobre el que estem investigant. És per aquest motiu que l'estadística proposa un seguit de tècniques per tal de resumir la informació, que ens sigui d'utilitat i que ens faciliti l'obtenció de resultats.

La primera d'aquestes eines són les anomenades taules de freqüències: són l'eina bàsica per a organitzar, sintetitzar i presentar les dades de manera informativa. Les taules ens faciliten la interpretació i els càlculs a partir de les dades obtingudes; el procés de construcció de les taules de freqüències el coneixem com a procés de tabulació de les dades.

Hi ha diversos tipus de taules, en funció de si mesuren la freqüència d'una o de més variables; o en funció del tipus de freqüència calculada. També podem tenir taules amb les dades sense agrupar, i les taules amb dades agrupades o en intervals. Començarem descrivint el funcionament de les taules més senzilles.

1) Taules per a dades unidimensionals no agrupades

Són les taules que recullen una sola característica o variable de les observades per al conjunt d'individus de la mostra; d'aquí el nom d'unidimensionals. D'altra banda, quan parlem de dades no agrupades fem referència al fet que agafarem els possibles valors que pot prendre la variable un a un. Això es pot fer sempre que les variables siguin quantitatives, amb valors enters (prenen valors discrets com ara 1, 2, 3..., però sense decimals; altrament les coneixem com a variables contínues), i la variable no prengui un nombre molt elevat de valors diferents.

Exemple

Quan preguntem pel nombre de vegades que s'ha visitat una determinada destinació, ens solen respondre 1, 2, 3..., com a molt podríem arribar fins a 10 (per damunt, en casos excepcionals). Tenim per tant una variable (nombre de vegades que s'ha fet una visita), que pot prendre un nombre relativament petit de valors de resposta, que a més són enters o discrets.

Per contra, si preguntem l'edat del visitant, ens podem trobar molts possibles valors, i a més, normalment no ens és rellevant conèixer l'edat exacta dels visitants, sinó saber en quins trams es mouen. És per això que acostumem a codificar aquesta variable en intervals: (14-17); (18-24); (25-34); (35-54); etc. Aquestes serien dades agrupades.

En general, una taula de freqüència és aquella en la qual consignem tots els possibles valors de la variable analitzada, i, per a cada valor fem correspondre la freqüència amb que aquest valor apareix en la nostra mostra.

Exemple

Imaginem que prenem una mostra de 10 turistes d'un parc natural als quals preguntem el nombre de vegades que han visitat el parc (inclosa l'actual). Les respostes obtingudes, per ordre d'enquesta, són les següents:

1 – 1 – 3 – 1 – 5 – 2 – 1 – 1 – 2 – 2

A partir d'aquesta informació podem establir que els valors de la variable «nombre de visites fetes al parc pels turistes enquestats» són 1, 2, 3, 4, 5 (fixem-nos que hem inclòs el 4 malgrat que no ha sortit en les respostes).

A continuació ja podem construir la taula de freqüències, o tabular la variable, tasca que òbviament avui dia s'encarreguen de dur a terme els programes de càlcul i estadístics. Abans, però, caldria definir els diferents tipus de freqüències que ens pot interessar conèixer:

- **Freqüència absoluta**: nombre de vegades que un valor es repeteix per a una determinada mostra. La seva suma ha de coincidir amb «n» raó per la qual totes les respostes obtingudes han d'haver estat incorporades a la taula. La notació per a la freqüència absoluta del valor i-èssim de la variable observada és n_i.

- **Freqüència relativa**: s'obté en ponderar el nombre de vegades que un valor es repeteix per a una determinada mostra sobre el nombre total d'observacions obtingudes. El seu sumatori ha de ser igual a la unitat. La

notació per a la freqüència relativa del valor i-èssim de la variable observada és f_i.

- **Freqüència absoluta acumulada**: és el resultat de sumar la freqüència absoluta corresponent a un determinat valor de la variable, a les freqüències absolutes associades a qualsevol valor anterior i inferior[3]. El seu darrer valor ha de coincidir amb el valor de n. La notació per a la freqüència absoluta acumulada del valor i-èssim de la variable observada és N_i.

[3]Tot i que no s'ha comentat encara, en tractar-se de variables quantitatives tenen caràcter ordinal, i per tant sempre seran introduïdes a la taula de manera ordenada, normalment, de menor a major valor.

- **Freqüència relativa acumulada**: és el resultat de sumar la freqüència relativa corresponent a un determinat valor de la variable, a les freqüències relatives associades a qualsevol valor anterior i inferior. El seu darrer valor ha de coincidir amb la unitat. La notació per a la freqüència absoluta acumulada del valor i-èssim de la variable observada és F_i.

Exemple

A partir de les dades exposades a l'exemple anterior, a continuació construirem la taula de freqüències corresponent, calculant els quatre tipus de freqüències assenyalats.

Valor: X_i	F. abs.: n_i	F. rel.: f_i	F. abs. ac.: N_i	F. rel. ac.: F_i
1	5	0,5	5	0,5
2	3	0,3	8	0,8
3	1	0,1	9	0,9
4	0	0	9	0,9
5	1	0,1	10	1

A la primera columna s'han consignat els 5 possibles valors de la variable (malgrat que el 4 no l'havia dit cap individu de la mostra, s'inclou perquè és un possible valor de la variable). En les columnes posteriors s'han calculat la freqüència absoluta, relativa, absoluta acumulada i relativa acumulada, a partir de les dades de l'exemple per a una mostra de n=10 observacions.

Si anem a la mostra obtinguda, veiem que 5 de les respostes ens deien que els individus havien visitat el parc una sola vegada: per això la freqüència absoluta del primer valor de X, 1, és igual a 5 ($n_1 = 5$). Així fem successivament. Com es pot observar, la suma de la columna relativa a les freqüències absolutes dóna com a resultat n, i per tant, 10. Pel que fa a les freqüències relatives, aquestes s'obtenen a partir de $\text{quocient}_i/n$, i la suma de la columna ens dóna 1. Finalment les freqüències acumulades són resultat de sumar les freqüències del valor estudiat més les de tots els valor inferiors a aquest.

A partir de la taula disponible, per tant, podrem afirmar coses com per exemple que el 50% dels visitants enquestats estan fent la seva primera visita, o que només un 20% han fet dues o més visites prèvies (1-0,8; 0,8 és la freqüència relativa acumulada corresponent al valor 2 - N_2. Per tant acumula els visitants que declaren que han fet només una visita al parc, l'actual, i els que declaren haver-ne fet dues, una de prèvia i l'actual).

Cal dir que malgrat que aquí l'explicació es fa per a variables de tipus quantitatiu (igual que els exemples que s'utilitzen), les taules de freqüències també es fan servir quan les dades són de naturalesa qualitativa.

2) Taules per a dades unidimensionals agrupades

Com s'ha esmentat anteriorment, quan el nombre de valors que pot prendre una variable es molt elevat, la millor manera per a resumir les dades és, en primer lloc, agrupant aquests valors en intervals. S'ha posat l'exemple de l'edat; una altra variable que tot sovint és treballada a partir dels intervals és la dels ingressos, i igualment la de despesa.

En el cas de treballar amb intervals, el primer que cal establir és el criteri que farem servir per a agrupar les dades. En aquest sentit, cal tenir en compte que podem fer servir intervals d'igual amplada, o intervals d'amplada irregular.

Exemple

Per a recollir la variable «edat» podem establir intervals que vagin, per exemple, de 10 en 10 anys ((14-23), (24-33), (34-43)...). Tanmateix, segons quins siguin la intenció i l'objectiu de l'estudi, una agrupació d'aquest tipus pot resultar-nos poc informativa, i en canvi, ens pot interessar establir intervals d'amplitud diferent o irregular, que s'adiguin a segments de demanda que ens interessa analitzar. Així per exemple agruparíem d'una banda els menors d'edat (14-17); un altre segment major d'edat però força jove (18-24); i així successivament. Com veiem, el primer interval conté només 4 possibles valors (14-15-16-17), mentre que el segon en conté 7 (18-19-20-21-22-23-24)[4].

[(4)]Atenció de no confondre valors de la variable, és a dir, les respostes que ens poden donar, i la freqüència amb què obtenim aquestes respostes. Quan diem que un interval conté per exemple 4 valors, com l'interval (14-17), significa que hi ha 4 possibles valors de les variables que conté. Això no té res a veure amb la freqüència amb què ens respondran si pertanyen o no a aquest interval. Aquesta freqüència només la coneixerem un cop duta a terme l'enquesta.

Cal dir que sempre mirarem de treballar amb un nombre d'intervals no gaire gran, ja que precisament es tracta de resumir la informació; però caldrà tenir en compte aquells que ens és d'interès recollir. Pensem que un cop les dades han estat processades en format d'intervals, es produeix una pèrdua d'informació. Perdem exactitud a favor de la capacitat de síntesi. Sabrem, per exemple, quants individus es troben en la franja d'entre 14 i 17 anys, però no sabrem quants tenen exactament 15 anys, o bé quina és l'edat mitjana. Per això, és molt important reflexionar bé sobre els intervals que ens interessa construir en funció dels objectius de l'estudi.

Pel que fa a la notació dels intervals, anomenarem L_{i-1} al límit inferior de l'interval, i L_i al límit superior. A més, a_i serà el que anomenarem amplada, mentre que h_i serà la seva «alçada», i c_i el que coneixem com a marca de classe o punt mitjà de l'interval. Per tant:

$a_i = L_i - L_{i-1}$

$h_i = n_i / a_i$

$c_i = (L_i + L_{i-1}) / 2$

Exemple

Agafant l'exemple de les dades anteriors, farem una taula de freqüències, agrupant ara els valors en 2 intervals: els visitants que estan en la seva primera o segona visita; els que n'han fet tres o més.

$L_{i-1} - L_i$	c_i	a_i	F. abs.: n_i	h_i	F. rel.: f_i	F. abs. ac.: N_i	F. rel. ac.: F_i
(1 - 2)	1,5	1	8	8	0,8	8	0,8
(3 - 5)	4	2	2	1	0,2	10	1

3) Taules per a dades bidimensionals o de doble entrada

Tot sovint ens interessa analitzar el comportament de la resposta de dues variables alhora. En aquest cas, entrem en l'anàlisi bidimensional o bivariant de les dades. Com es veurà més endavant, aquesta anàlisi ens permetrà conèixer, entre altres coses, la relació existent entre les dues variables observades, és a dir, determinar si les variables són independents, o bé estan relacionades.

Es considerarà que les variables X i Y són independents quan els valors d'una d'aquestes no afecti la distribució de l'altra. Per a poder establir aquesta relació, a part de confeccionar la taula bidimensional, s'haurà del calcular l'estadístic de referència que permetrà mesurar si existeix o no aquesta relació, i en quin grau. Més endavant s'explicarà en detall.

A les taules bidimensionals o de doble entrada, cal tenir en compte que els valors d'una variable (que anomenarem X) aniran a les files, mentre que els valors de l'altra variable (que anomenarem Y) aniran a les columnes. En les cel·les d'intersecció per a cada parell de valors hi consignarem la freqüència amb què hem obtingut tots dos valors simultàniament, és a dir, el que coneixem com a freqüència conjunta. Aquesta freqüència conjunta pot ser absoluta, relativa o acumulada, i es calcula igual que s'ha exposat anteriorment.

Exemple

Suposem que a l'entrevista duta a terme a l'entrada del parc a una mostra de n=10 visitants, se'ls ha preguntat simultàniament pel nombre de vegades que han visitat el parc (inclosa la visita actual) així com el nombre de persones que constitueixen el grup que fa la visita juntament amb l'enquestat. S'obtenen els següents resultats:

X_i	1	1	3	1	5	2	1	1	2	2
Y_j	3	3	2	2	2	1	1	2	1	2

El primer que podem observar és que podem extreure dues taules de freqüències unidimensionals d'una de bidimensional; però no a la inversa. Si les dues preguntes s'haguessin fet de manera no simultània, tindríem dues dades unidimensionals (d'una banda, la relativa al nombre de visites, i de l'altra, la relativa a la composició del grup de visitants), però no podríem establir parells de dades com tenim aquí, on el primer enquestat ens ha respost (1 visita, 3 persones), el segon enquestat (1 visita, 3 persones), el tercer enquestat (3 visites, 2 persones), etc. Per tant, en aquest cas podríem tractar cada variable per separat si fos del nostre interès, i obtindríem dues taules de freqüències:

Valor: X_i	F. abs.: n_i	F. rel.: f_i	F. abs. ac.: N_i	F. rel. ac.: F_i
1	5	0,5	5	0,5
2	3	0,3	8	0,8
3	1	0,1	9	0,9
4	0	0	9	0,9
5	1	0,1	10	1

Valor: Y_j	F. abs.: n_j	F. rel.: f_j	F. abs. ac.: N_j	F. rel. ac.: F_j
1	3	0,3	3	0,3
2	5	0,5	8	0,8
3	2	0,2	10	1

Més endavant explicarem que aquestes dues taules, en aquest cas –provenint d'una informació bidimensional–, s'anomenen taules de freqüències marginals.

Si ara construïm la taula de doble entrada, procedirem com s'ha assenyalat, és a dir, posant les freqüències conjuntes a les cel·les d'intersecció entre els dos valors de les variables pertinents. La freqüència absoluta conjunta serà n_{ij} mentre que la relativa serà f_{ij}. El fet d'incloure dos subíndexs ens indica que estem treballant freqüències conjuntes relatives a parells de valors de dues variables. El primer subíndex (i) farà referència als valors de la variable que hem identificat com a variable X, mentre que el subíndex j farà referència als valors de Y.

A la taula següent hi consignarem les freqüències absolutes conjuntes; la suma d'aquestes freqüències haurà de ser, com sempre, igual a n (en aquest cas igual a 10).

$X_i \setminus Y_j$	1	2	3
1	1	2	2
2	2	1	0
3	0	1	0
4	0	0	0
5	0	1	0

Aquesta taula ens informa, per exemple, que hi ha 1 resposta en la qual es compleix simultàniament que es tracta de primera visita, i la dóna una sola persona ($n_{11} = 1$), o que hi ha 2 respostes per a la combinació de primera visita i grup de 2 persones ($n_{12} = 2$).

Comprovem que es compleix que $\Sigma\Sigma\, n_{ij} = 10$: si procedim per files,

$1 + 2 + 2 + 2 + 1 + 0 + 0 + 1 + 0 + 0 + 0 + 0 + 0 + 1 + 0 = 10$.

D'altra banda, si afegim una fila i una columna més a la taula, i hi consignem els respectius sumatoris per a cada valor, obtindrem les taules de freqüències per a cada variable (tal i com ens havien sortit anteriorment). Aquestes distribucions de freqüències univariants, però que sorgeixen de dades bidimensionals, les anomenem freqüències marginals, ja que ens informen de la freqüència dels valors d'una sola variable, independentment del valor que prengui l'altra variable. La notació que s'utilitza és la següent: per a les freqüències absolutes marginals de X indicarem n_i. (substituïm la j per un punt assenyalant que no tenim en compte els valors de la variable Y), i per a les freqüències absolutes marginals de Y indicarem $n_{\cdot j}$ (ara substituïm la i per un punt, assenyalant que no tenim en compte els valors de la variable X). Al seu torn, cadascuna de les distribucions de freqüències absolutes marginals hauran de sumar n. Com sempre, si el que calculem són freqüències relatives marginals, la notació serà f_i. per a la X i $f_{\cdot j}$ per a la Y; i cadascuna de les distribucions haurà de sumar 1.

A continuació introduïm a la taula les distribucions marginals, com es deia, amb una columna i una fila addicionals, de fet a la darrera columna i la darrera fila, respectivament. Com s'observa, les distribucions per a X i Y són les mateixes obtingudes anteriorment, si s'hagués treballat el resultat de l'enquesta de l'exemple com a dues variables unidimensionals separades.

$X_i \setminus Y_j$	1	2	3	$n_{i\cdot}$
1	1	2	2	5

$X_i \setminus Y_j$	1	2	3	$n_{i.}$
2	2	1	0	3
3	0	1	0	1
4	0	0	0	0
5	0	1	0	1
$n_{j.}$	3	5	2	$\sum\sum n_{ij}=10$

Tot seguit reproduïm la taula de freqüències relatives conjuntes que quedaria (dividint cada freqüència absoluta per n), amb les freqüències relatives marginals per a les variables X i Y a la darrera columna i la darrera fila, respectivament.

$X_i \setminus Y_j$	1	2	3	$f_{i.}$
1	0,1	0,2	0,2	0,5
2	0,2	0,1	0	0,3
3	0	0,1	0	0,1
4	0	0	0	0
5	0	0,1	0	0,1
$f_{j.}$	0,3	0,5	0,2	$\sum\sum f_{ij}=1$

5.1.2. Representació gràfica

Una altra eina d'elevada utilitat i ús molt estès, per al resum de les dades obtingudes i la seva anàlisi i descripció, és la representació gràfica de la informació. Els gràfics permeten mostrar informació sintètica i rellevant sobre el comportament conjunt de les dades, de manera visual, i per tant a partir de la simple observació. La representació gràfica de les dades implica fonamentalment la plasmació del comportament d'aquestes dades sobre uns eixos o coordenades determinats, a partir d'unes regles i normes estandarditzades, que poden adoptar diferents formats i versions. Per tant, cal conèixer quines són les opcions més emprades en la representació gràfica de les dades a la fi de saber quina farem servir per a cada cas i poder donar una interpretació correcta del gràfic obtingut. En aquest sentit, cada tipus de variable acostuma a tenir un tipus de representació més escaient, tot i que de vegades aquest fet també depèn de la característica que volem conèixer i de l'objectiu perseguit.

Exemple

Quan volem conèixer distribucions percentuals –o pes relatiu– del conjunt d'una mostra en les diferents categories o valors d'una variable, normalment farem servir els gràfics de sectors. En canvi, si el que volem és tenir una informació visual de la possible relació entre dues variables, farem servir un diagrama de dispersió. En el cas que la variable sigui quantitativa amb dades agrupades, tindrà sentit fer un histograma, però si les dades són qualitatives, aquest tipus de representació no es pot fer. Més endavant s'expliquen les característiques i diferències d'aquests tipus de representacions.

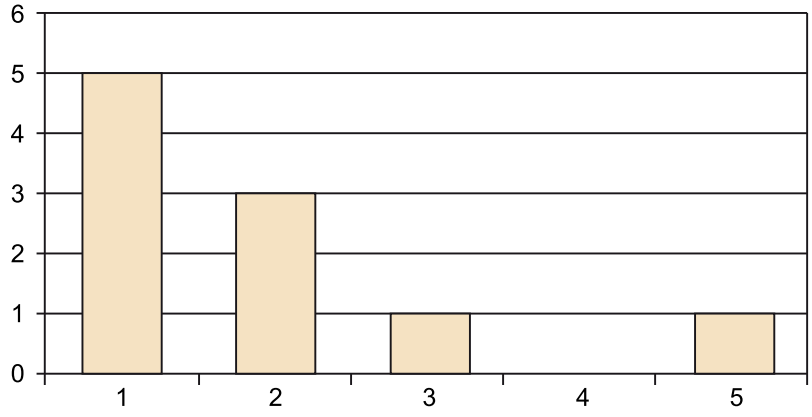
A continuació s'expliquen els principals tipus de gràfics que s'acostumen a utilitzar en l'anàlisi de dades estadística.

1) Diagrama de barres

S'utilitza per a variables quantitatives discretes i no agrupades. Es representa sobre un doble eix: un, el de les abscisses, serveix per a representar els diferents valors que pot prendre la variable; a l'altre, el de les ordenades, és on es representa el valor de les freqüències corresponents a cada valor. Normalment es treballa amb freqüències absolutes.

Exemple

A partir de les dades de la variable X (nombre de vegades que s'ha visitat el parc, inclosa l'actual), feta servir als anteriors exemples, a continuació reproduïm el que seria el seu diagrama de barres a partir de les freqüències absolutes.



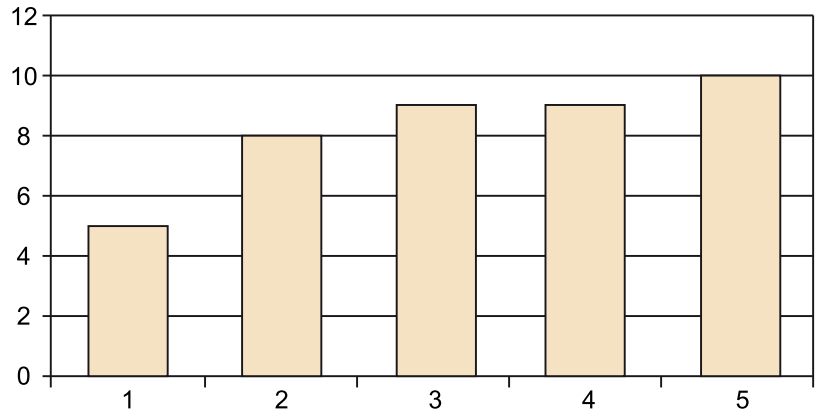
Com s'observa, la variable pot prendre valors des de l'1 fins al 5; i per a cada valor, la taula ens mostra la seva freqüència absoluta. Així, el valor 1 apareix 5 vegades entre ens enquestats, el valor 2 ho fa 3 vegades, etc.

2) Diagrama d'escala o freqüències acumulades

Es tracta d'un gràfic similar a l'anterior, però en comptes de representar la freqüència absoluta corresponent a cada valor, en representa la freqüència acumulada.

Exemple

Seguint amb l'exemple anterior, obtindríem el següent gràfic, on el valor 1 presenta una freqüència igual a 5, i a partir d'aquest es comencen a acumular freqüències: el valor 2 té una freqüència acumulada de 8, és a dir, hi ha 8 respostes, del total de n=10, que han contestat 2 o menys. Per descomptat la freqüència corresponent a la barra del darrer valor de la variable (en aquest cas, 5) ha de correspondre al valor de n (per tant, 10).

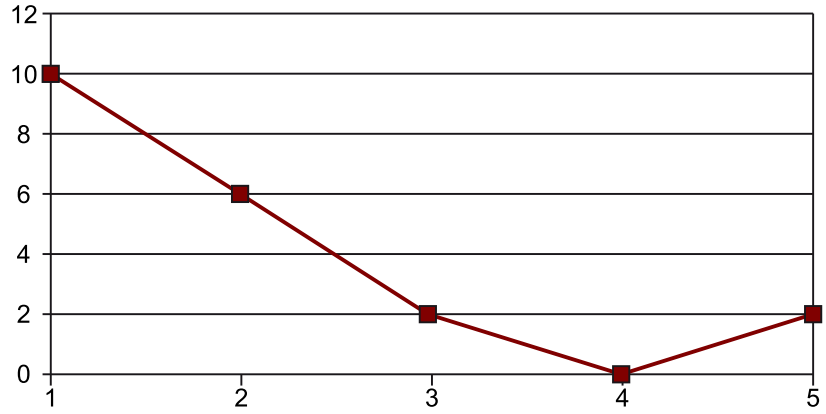


3) Polígon de línies

Consisteix en un diagrama que normalment s'utilitza per mostrar sèries temporals, i permet veure visualment les fluctuacions d'una determinada variable al llarg del temps.

Exemple

De nou a partir de les dades de la variable X, obtindríem el següent gràfic.



4) Histograma

S'utilitza per a la representació de dades quantitatives contínues o agrupades. També reprodueix d'alguna manera, en l'eix de les ordenades, la mesura de la freqüència absoluta de cada valor, però en aquest cas, en no tenir valors puntuals i tenir-los agrupats en intervals, en retorna una barra per cada interval. Tanmateix, en aquest cas, i atès que, com diem, cada barra no correspon a un únic valor, sinó a diversos, en el gràfic, com a mesura correcta no farem servir la freqüència absoluta sinó l'alçada (hi) de l'interval. Per què ho fem d'aquesta manera? Si disposéssim sempre d'intervals de la mateixa amplada, la base del rectangle que representaria cada barra seria constant; per tant, no hi hauria problema a fer servir la freqüència absoluta. Però com que l'amplada de

l'interval pot ser variable, ens trobem que, en representar a la gràfica les barres corresponents, la seva base també seria diferent. Si féssim servir aleshores la freqüència per a l'eix de les ordenades, la informació visual no seria la correcta.

Exemple

Il·lustrem a continuació la necessitat d'emprar l'alçada quan les dades estan agrupades. Seguim amb l'exemple de l'enquesta a n=10 visitants del parc. Ara els preguntem per l'edat i establim 5 intervals d'amplada diferent, i obtenim la següent taula de freqüències absolutes a partir de les respostes.

L_{i-1} - L_i	c_i	a_i	F. abs.: n_i	h_i	F. rel.: f_i	F. abs. ac.: N_i	F. rel. ac.: F_i
(14-18)	16	4	0	0	0	0	0
(19-24)	21,5	5	3	0,60	0,3	3	0,3
(25-34)	29,5	9	4	0,44	0,4	7	0,7
(35-54)	44,5	19	1	0,05	0,1	8	0,8
(55-99)	77	44	2	0,05	0,2	10	1

Si s'observa la taula, a la primera columna tenim els intervals amb els valors de la variable; a la segona la marca de classe o punt mitjà; a la tercera, l'amplada de l'interval; a la quarta, la freqüència absoluta de cada interval de valors; a la cinquena, l'alçada, és a dir, el quocient resultant de dividir la freqüència absoluta entre l'amplada; i la resta ens informen de les freqüències relativa, acumulada absoluta i acumulada relativa.

Efectivament, veiem que les amplades són molt dispars, en funció de grups de públic que ens interessa diferenciar. Per tant, quan haguem de representar els intervals a un gràfic, les barres tindran unes bases molt dispars. Pel que fa a n_i, o freqüència absoluta, veiem que l'interval en el qual hem obtingut una major freqüència de respostes és el de l'edat compresa entre 25 i 34 anys; en concret, 4 dels 10 enquestats es troben en aquesta franja de resposta. Podem dir, doncs, que, com a interval, aquest és el més freqüent.

Ara bé:

- si tenim en compte que aquest interval conté dins deu possibles valors concrets de resposta (amb una amplada igual a 9 = 34-25), que serien: 25, 26, 27, 28, 29, 30, 31, 32, 33, 34;

- i en canvi, l'interval anterior, (19-24), amb una freqüència absoluta igual a 3, només té una amplada de 5 (5 = 24-19) i per tant conté només 6 possibles valors de resposta: 19, 20, 21, 22, 23, 24,

- ens adonarem que, malgrat que l'interval més freqüent és, com s'ha dit, el de (24-35), tanmateix, en termes relatius, tenint en compte la quantitat de possibles respostes puntuals que cada interval recull, l'interval que conté els valors amb major freqüència, a nivell puntual o per cada valor inclòs en l'interval, és el de (19-24).

Com arribem a aquesta conclusió? Calculant l'alçada. Així:

- mentre que per a l'interval (25-34) l'alçada resultant és de 0,44 (0,44 = 4 / 9, és a dir, amb una amplada de 9, i per tant 10 possibles valors de resposta a dins de l'interval, s'han trobat 4 individus que s'inclouen en l'interval);

- per a l'interval (19-24), l'alçada obtinguda és de 0,6 (0,6 = 3 / 5; és a dir, amb una amplada de 5, i per tant 6 possibles valors de resposta a dins de l'interval, s'han trobat 3 individus que s'inclouen en l'interval).

Amb aquesta informació, si representem l'histograma, cada barra reflectirà visualment les àrees correctes, corresponents a la freqüència dels valors de la variable, ja que, amb bases irregulars, i per tant amplades canviants, el rectangle final que figurarà per a la barra de cada interval tindrà l'àrea real que li correspon. Recordem que l'àrea d'un rectangle es calcula com el producte de la base per l'alçada. En canvi, si la barra la construíssim fent servir la freqüència absoluta (com és el cas en el diagrama de barres per a dades no agrupades), les àrees finals serien incorrectes i visualment l'efecte ens distorsionaria la interpretació de la informació.

A continuació podeu veure la representació gràfica del que seria un gràfic de barres erroni i incorrecte, fet utilitzant les freqüències absolutes, i l'histograma correctament aplicat en aquest cas, emprant les alçades obtingudes per a cada interval.

5) Diagrama de dispersió

El diagrama de dispersió, o núvol de punts, es fa servir quan volem veure gràficament la possible relació entre dues variables quantitatives. En aquest cas, les dades obtingudes es representen en un gràfic de dues coordenades, a cadascuna de les quals representem una de les variables. Els punts dibuixats al gràfic corresponen, doncs, als parells de respostes obtingudes.

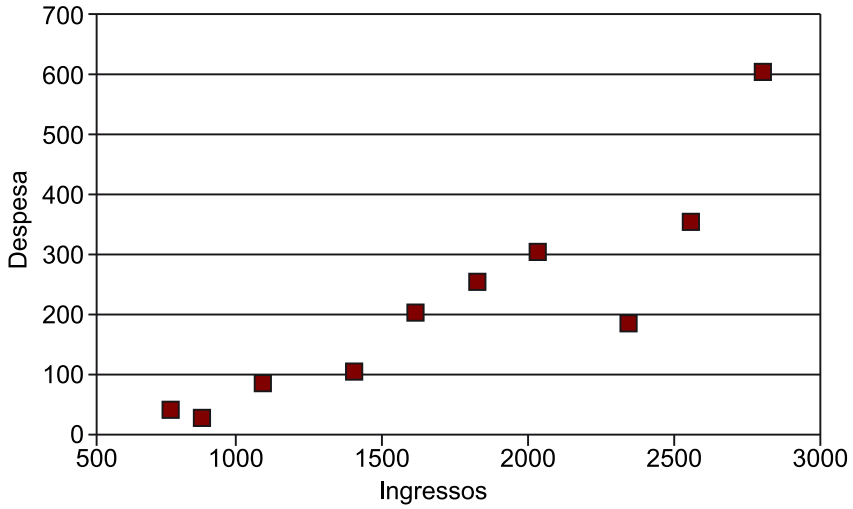
Exemple

Imaginem que, en l'enquesta feta als n=10 visitants del parc, els hem demanat també resposta al voltant de les següents variables: ING = ingressos mitjans mensuals (en euros) i DESP = despesa mitjana mensual en activitats de lleure, culturals, turístiques..., i n'hem obtingut les següents respostes (presentades per parells, és a dir, cada columna correspon a un mateix individu):

ING	2.500	1.400	1.100	2.300	800	1.600	900	2.000	3.200	1.800
DESP	350	100	80	180	30	200	20	300	600	250

La gràfica corresponent al diagrama de dispersió, o «XY» –que confronta variables dues a dues–, seria la que es reprodueix a continuació: com s'observa, per exemple, per a l'individu que ha formulat la cinquena resposta (Ingressos-X=800 i Despesa-Y=30), el gràfic ens retorna el primer punt que apareix representat començant de dreta a esquerra, i precisament on es creuarien els valors ING=800 i DESP=30. Per tant, el conjunt representat és el conegut com a núvol de punts. Observant-lo es pot concloure que efectivament, sembla haver-hi una relació entre ambdues variables estudiades, ja que quan una creix, l'altra també ho fa. El núvol de punts és clarament ascendent. En aquest cas, parlaríem

d'una relació de tipus directa o positiva. Si fos descendent, es tractaria d'una relació inversa o negativa, com es veurà més endavant.



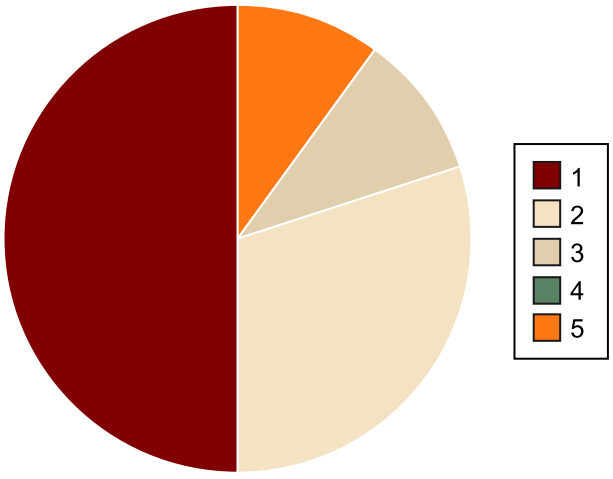
6) Diagrama de sectors

El diagrama de sectors, conegut també col·loquialment com a pastís o formatge. És un gràfic que s'utilitza per a dibuixar sectors a partir d'un cercle, sectors que corresponen a percentatge, freqüència relativa o pes relatiu de cada valor de la variable sobre el total. Es fa servir tant per a variables quantitatives com qualitatives, ja que, com s'ha dit, representa percentatges o freqüències relatives.

D'aquesta manera, els 360º del cercle són dividits en diversos sectors (tants com valors tingui la variable), de manera correlativa al pes de cada valor sobre el total.

Exemple

Reprenent un cop més les dades de la variable definida des de l'inici, X=nombre de visites fetes al parc, inclosa l'actual, per a una mostra de n=10, el diagrama de sectors corresponent seria el següent:



Com veiem, les persones que es troben a la seva primera visita representen la meitat del cercle, és a dir, el 50% (si recordem, 5 persones d'un total de 10 havien contestat aquesta opció). Així, el diagrama de sectors ens ofereix una interpretació visual força ràpida, simple i intuïtiva en relació amb el pes de cada valor sobre el total de les respostes obtingudes per a la variable analitzada.

Aquests gràfics que s'han explicat són els més extensament utilitzats. Tanmateix, se'n poden trobar molts d'altres que també poden ser útils per a poder explicar la informació desitjada. Els programes estadístics actuals disposen d'assistents que ofereixen una gran gamma de gràfics diferents i formats que ajuden a poder trobar la forma més adient de representar gràficament les dades.

5.1.3. Mesures de síntesi

Les mesures de síntesi són valors numèric que calculem a partir de la informació mostral, i que ens resumeixen alguna característica de la mostra, cosa que permet avançar en la seva descripció. Dit d'altra manera, amb un sol nombre o valor, podem conèixer alguna propietat important de la mostra que estem estudiant. L'exemple més conegut és la mitjana aritmètica. Prenent tots els valors de la mostra, fem un càlcul i podem saber quin és el valor central al voltant del qual se situen tots els valors observats.

A continuació s'exposaran les principals mesures de síntesi emprades en la investigació. Normalment es treballa amb quatre tipus de mesures: mesures de tendència central o centralitat, mesures de dispersió, mesures de posició i mesures de forma. En aquest cas ens centrarem en les mesures de tendència central i de dispersió, tot i que esmentarem les altres modalitats de mesures.

1) Mesures de tendència central

a) Mitjana o mitjana aritmètica

És el resultat de calcular la mitjana aritmètica del conjunt d'observacions, per tant, s'obté com a quocient entre el sumatori d'aquestes observacions entre el nombre total de dades observades (n).

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

Exemple

Per a la variable X, nombre de visites fetes al parc, on havíem obtingut les següents n=10 observacions: 1 – 1 – 3 – 1 – 5 – 2 – 1 – 1 – 2 – 2, es demana calcular la mitjana aritmètica X.

X = (1+1+3+1+5+2+1+1+2+2) / 10 = 19 / 10 = 1,9

Dues consideracions:

- Si les dades les tenim resumides ja en taules de freqüències, aleshores, en comptes de sumar una a una cada observació, directament multipliquen

cada valor possible (x_i) per la seva freqüència absoluta (n_i); fem la suma de tots els resultats obtinguts i aquest serà el numerador del quocient, que dividirem per la mida mostral o n.

Exemple

Si ara agafem la mateixa variable X, però agrupada per valors i amb les seves freqüències, obtindrem que:

X = [(1x5) + (2x3) + (3x1) + (4x0) + (5x1)] / 10 = 1,9

- Si les dades estan agrupades en intervals, per al càlcul de la mitjana es farà servir la marca de classe o c_i.

Exemple

Prendrem ara el cas de les dades relatives a la variable edat, agrupades en intervals, i farem el càlcul fent servir la marca de classe i la freqüència absoluta per a cadascun dels intervals presents.

Edat = [(16x0) + (21,5x3) + (29,5x4) + (44,5x1) + (77x2)] / 10 = 381 / 10 = 38,1

b) Moda

La moda és el valor de la variable que més vegades es repeteix a la mostra, és a dir, aquell amb una major freqüència. Cal dir que per a l'obtenció de la moda no cal fer cap càlcul, i que no ha d'haver-hi un únic valor per a cada distribució. És important saber que la moda no és aquesta freqüència màxima, sinó el valor de la variable que pren aquesta freqüència.

Exemple

Per a la variable X de sempre en l'exemple que estem utilitzant, la moda seria el valor Mo=1, ja que és el que més vegades es repeteix, amb una freqüència igual a 5.

Quan tinguem les dades agrupades en intervals, podrem saber quin és l'interval modal, és a dir, l'interval que pren una freqüència més elevada. Cal observar a més que aquesta és l'única mesura de síntesi que és també aplicable a les variables qualitatives.

c) Mediana

Aquesta mesura de síntesi ens informa de quin és el valor de la distribució que ens la divideix en dues parts iguals, és a dir, que deixa un 50% de les observacions per sota, i un 50% per sobre. Per a fer-ho, primer cal que les dades estiguin ordenades de menor a major. D'altra banda, si n és imparell, l'observació que deixa un 50% de les dades (no valors, dades) a cada banda, serà la mediana o Me. Per contra, si n es parell, la mediana serà la mitjana aritmètica de les dues observacions situades al centre.

Exemple

Tornant a les dades de la variable X de l'exemple: recordem que teníem aquests resultats per a una n=10:1 – 1 – 3 – 1 – 5 – 2 – 1 – 1 – 2 – 2. Primer de tot ordenarem de menor a major, per tant: 1 – 1 – 1 – 1 – 1 – 2 – 2 – 2 – 3- 5. Com que n és 10, i per tant, parell, cercarem les dues observacions centrals: en aquest cas, les que ocupen la cinquena i la

sisena posició, és a dir, un 1 i un 2: 1 – 1 – 1 – 1 – 1 – 2 – 2 – 2 – 3- 5. Fixem-nos que totes dues deixen la mateixa quantitat de dades, respostes o observacions a cada banda. Finalment, el valor de la mediana serà la mitjana entre aquests dos valors, per tant Me = 1,5.

En cas que tinguem les dades ja en format de taula de freqüències, no cal que tornem a desglossar-les una a una (i menys tenint en compte que habitualment farem servir mides mostrals molt elevades). En aquest cas, l'obtenció de la mediana es du a terme a través de l'observació de les freqüències relatives acumulades. Com que la F_i ens informa de quantes observacions tenim que siguin corresponents a un valor determinat o inferiors, en termes relatius, sempre arriba un punt en què hi ha un valor al qual correspon una F_i igual o superior al 0,5 (50%). Doncs bé, de nou, si n és imparell, agafarem el valor central; si n és parell, de nou farem la mitjana aritmètica dels dos valors centrals.

Exemple

Agafem les mateixes dades que abans, però ara a partir de la taula de freqüències:

Valor: X_i	F. abs.: n_i	F. rel.: f_i	F. abs. ac.: N_i	F. rel. ac.: F_i
1	5	0,5	5	0,5
2	3	0,3	8	0,8
3	1	0,1	9	0,9
4	0	0	9	0,9
5	1	0,1	10	1

Si mirem a la darrera columna, la de les freqüències relatives acumulades, veiem que ja el primer valor, és a dir, 1 (primera visita al parc) acumula el 50% del total d'observacions. Tanmateix com que n=10 i és parell, el valor que acumularia el 51% és el següent, és a dir, el 2. Així doncs, tornem a fer la mitjana aritmètica i obtenim de nou que Me=1,5.

Exemple

Prenem ara l'exemple amb la variable del nombre de persones que formen el grup que fa la visita amb l'enquestat (Y):

Valor: Y_j	F. abs.: n_j	F. rel.: f_j	F. abs. ac.: N_j	F. rel. ac.: F_j
1	3	0,3	3	0,3
2	5	0,5	8	0,8
3	2	0,2	10	1

Mirant la taula, observem que les persones que fan la visita soles representen el 30% del total; però si ens fixem en la darrera columna, on hi ha les freqüències relatives acumulades, veiem que quan pugem a 2 el nombre de persones que formen el grup, podem veure com aquí ja aquesta freqüència ha pujat al 80%: és a dir, que l'observació (en cas que n fos imparell) o les observacions (en cas

que n sigui parell, com en l'exemple emprat on n=10) que caldria prendre per a calcular la mediana s'inclouen en aquest valor. Així doncs, les dues observacions centrals serien 2 i 2; i fent la mitjana aritmètica, òbviament obtindríem que la Me=2.

2) Mesures de dispersió

Les mesures de dispersió ens informen de la variabilitat dels valors presents a la variable que estem estudiant. D'aquí el seu nom: ens diuen si les observacions es troben molt disperses, o bé si, al contrari, oscil·len a prop de determinats valors, sense allunyar-se gaire entre elles. Hem vist anteriorment algunes mesures de centralitat; tanmateix, si bé és cert que ens donen una informació valuosa, és tracta d'una informació força incompleta.

Exemple

Quan ens diuen que els ingressos mitjans d'una població són de 2.000 euros, aquesta dada ens aporta una certa informació. Tanmateix, aquesta mitjana pot haver sorgit d'una població en la qual els ingressos dels individus poden oscil·lar entre els 200 i els 6.000 euros, o d'una en la qual aquests ingressos es moguin entre els 1.000 i els 4.000 euros.

Les mesures de dispersió, per tant, ens completen de manera molt valuosa la informació relativa a allò que estem estudiant. De fet, les mesures de dispersió ens diuen com de representativa és la dada de tendència central.

Exemple

Seguint amb l'exemple anterior, la dada d'uns ingressos mitjans de 2.000 euros resulta més representativa del conjunt de la població (i valors) en el cas que les seves dades vagin entre 1.000 i 4.000 euros, que en l'altre cas posat com a exemple (de 200 a 6.000 euros).

Tot seguit veurem les principals mesures de dispersió i distingirem entre mesures de dispersió absolutes i mesures de dispersió relatives, és a dir, que permeten comparar entre poblacions o mostres, que és de fet amb el que acostumarem a treballar.

a) Mesures de dispersió absolutes

En veurem dues: la variància i la desviació estàndard o desviació típica. De fet, la segona no és més que l'arrel quadrada de la primera, per tant presentarem en primer lloc la variància.

La variància és una mesura de dispersió que ens informa de quina és la distància de cada observació respecte de la mitjana aritmètica en terme mitjà, i en termes quadràtics. Què volem dir amb això? La variància parteix del càlcul de la distància entre cada valor de la mostra i la mitjana aritmètica. Aquesta distància es calcula a partir de la resta. A partir d'aquí, per a obtenir la mitjana d'aquestes distàncies n'hi hauria prou de sumar-les i dividir-les per n (el nombre d'observacions). Però sorgeix un problema de plantejament fonamental: com que el valor de referència és la mitjana aritmètica, en calcular les distàncies trobarem valors que s'allunyen per dalt (excés) o per baix (defecte). Així

doncs, si sumem totes les distàncies sense prendre en consideració aquest fet, estarem incloent signes inversos; en l'extrem, en una mostra amb observacions molt disperses o allunyades, però que ho fossin simètricament, les diferències negatives i les positives s'anul·larien i ens donarien una variaciabilitat en els valors, o variància, de zero, que no seria pas real. És per aquest motiu que, per tal d'evitar que els signes es compensin entre ells, la variància parteix de la suma de totes aquestes diferències (de cada valor o observació respecte de la mitjana), però elevades al quadrat, és a dir, evitant el signe negatiu. El resultat no està mancat de problemes, ja que acabem tenint una mesura amb unitats al quadrat, que no té interpretació possible. Tanmateix és més el guany que l'inconvenient. Per això, la variància és, juntament amb la seva arrel quadrada –com s'explicarà de seguida– la mesura de dispersió òptima i més utilitzada.

La variància es representa com a S^2 i la seva fórmula és la següent:

$$s^2 = \frac{\sum_i (X_i - \overline{X})^2 n_i}{n}$$

També s'acostuma a fer servir, operativament, aquesta altra fórmula, obtinguda del desenvolupament de l'anterior:

$$s^2 = \frac{\sum_i X^2{}_i n_i}{n} - \bar{x}^2$$

S'aconsella fer servir aquesta segona fórmula, ja que a efectes de càlcul és molt més simple que la fórmula definitòria.

Exemple

Agafem la variable de nombre de visites al parc (X) i apliquem la fórmula proposada per a calcular-ne la variable.

X_i	n_i	$X_i * n_i$	$(X_i * ni)^2$
1	5	5	25
2	3	6	36
3	1	3	9
4	0	0	0
5	1	5	25
		Suma =	95

Per tant, si la suma de $(X_i * n_i)^2$ és igual a 95, i sabíem que la mitjana de X és igual a 1,9:

$S^2 = [(95 / 10) – (1,9)^2] = 5,89$ (en unitats al quadrat)

Activitat

Com a exercici es proposa calcular la variància de Y (nombre de persones que formen el grup) i verificar que és igual a 10,89 (de nou unitats al quadrat, la qual cosa no té cap sentit).

Parem atenció a un detall important: mentre que X té un recorregut que va des del valor 1 fins al 5, Y només oscil·la entre 1 i 3. Si observem la variància, ens pot semblar que X és menys dispersa (té menor variància) que Y; tanmateix, com s'acaba d'esmentar, X té més recorregut (per tant podria ser que presentés més dispersió o variabilitat en els seus valors) que Y. Tindríem aquí dos criteris contradictoris que no podríem clarificar. Cal, doncs, insistir que aquest tipus de conclusions no es poden extreure amb la variància, ja que es tracta d'una mesura de dispersió absoluta, en unitats al quadrat sense interpretació possible. Per a fer comparacions i saber quina variable és menys dispersa (i per tant la mitjana aritmètica és més representativa del conjunt de la població) caldrà fer servir una mesura de dispersió absoluta.

Finalment, definirem l'altra mesura de dispersió absoluta més utilitzada: la desviació estàndard. Com ja s'ha dit, aquesta es calcula a partir de l'arrel quadrada positiva de la variància, retornant doncs a les unitats originals.

$$s = \sqrt{s^2}$$

Exemple

La desviació estàndard de X seria igual a 2,4 mentre que la de Y seria de 3,3 (unitats).

b) Mesures de dispersió relativa

Les mesures de dispersió relativa són les que ens permeten establir comparacions entre una variable i una altra o fins i tot entre una mostra i altra.

L'indicador o estadístic fonamental pel que fa a la dispersió relativa és el coeficient de variació de Pearson (CV). Aquest coeficient es calcula dividint la desviació estàndard de cada cas per la seva mitjana. D'aquesta manera, el que ens està dient és quantes vegades està continguda la mitjana dins de la desviació estàndard i, per tant, com n'és de representativa. Com més gran sigui el CV, més dispersió i per tant menys representativitat de la mitjana.

$$CV = \frac{s}{|\overline{x}|}$$

Exemple

Sabem que la S_x = 2,4 i la S_y = 3,3; també sabem que X = 1,9 i, casualment prenen el mateix valor, Y = 1,9. Així doncs:

CV (X) = 2,4 / 1,9 = 1,26 i CV (Y) = 3,3 / 1,9 = 1,73

En conclusió, en aquest cas, i malgrat que Y té un menor recorregut de valors, es constata que Y té més dispersió que X en termes relatius i, per tant, la mitjana esdevé més representativa per explicar el conjunt de la mostra en el cas de X que en el cas de Y.

3) Altres mesures de síntesi: posició i forma

Tot i que no s'exposen amb detall, cal tenir en compte que per a una diagnosi descriptiva completa, els informes acostumen a incorporar també mesures de posició i forma.

Les mesures de posició ens permeten saber on es van situant els diversos valors de la distribució; de manera similar al concepte de la mediana, que divideix les observacions en dues parts iguals, deixant un 50% a cada banda, les mesures de posició ens diuen on se situen els valors que acumulen cada 10% de les observacions (10%, 20%, 30%, etc. Aquestes mesures s'anomenen decils, i de fet la mediana és el cinquè decil, ja que ens situa el 50%); cada 25% (25%, 50%, 75%, són els anomenats quartils, ja que divideixen el total d'observacions en quatre parts; de nou la mediana correspon al quartil segon); o cada 1% (1%, 2%, 3%..., serien els percentils, ja que divideixen les observacions en 100 parts; en aquest cas la mediana seria el percentil número 50). Així mateix, i igual que la mediana, són mesures que s'obtenen a partir de l'anàlisi de les freqüències relatives acumulades.

Finalment, les mesures de forma ens parlen de dos aspectes: d'una banda, la simetria en la distribució de les dades respecte de la mitjana, observant si es troben centrades, o per contra, si es prolonguen cap a una de les bandes (per sota o per sobre de la mitjana). D'altra banda, s'estudia la curtosi o apuntament: és a dir, si hi ha una elevada concentració en uns pocs valors centrals, amb la qual cosa tindríem un diagrama de barres o histograma molt apuntat, amb poques barres i altes; o si, per contra, tenim les observacions de la variable molt distribuïdes entre diferents valors, cosa que implicaria un diagrama de barres o histograma amb moltes barres (cadascuna per a cada valor observat) i sense un gran apuntament de cap d'aquestes barres, més aviat planes i constants.

5.2. Anàlisi de la relació entre variables

Una part important en la investigació és, no tan sols descriure què passa en la realitat d'allò que observem, sinó fer passes més enllà i mirar d'explicar per què passa, quines són les relacions que s'estableixen entre les variables analitzades i poder-se endinsar en les possibles causes o efectes. Així doncs, una vegada s'ha dut a terme una anàlisi conceptual, i una de descriptiva, tot sovint s'introdueixen altres tècniques per al tractament de la informació que permeten avançar en aquest coneixement.

De forma més simple podem trobar l'anàlisi bivariant que es limita a analitzar la relació entre dues variables, mentre que de forma més complexa podem trobar l'anàlisi multivariant, la qual permet introduir un nombre elevat de variables per a portar a terme l'anàlisi corresponent.

Com a element important cal recordar que trobem diversos tipus de variables en funció de la seva naturalesa qualitativa o quantitativa. L'interès pot estar en el fet de relacionar tant variables d'un tipus com d'un altre, però ateses les seves característiques no podrem fer el mateix tractament estadístic, de manera que s'haurà d'anar amb cura de seleccionar aquell mètode adequat a cada cas o objectiu d'anàlisi.

Iniciem aquest apartat centrant-nos en l'anàlisi bivariant, en la qual distingirem entre:

- la relació entre dues variables quantitatives,
- la relació entre dues variables qualitatives, i
- la relació entre una variable qualitativa i una variable quantitativa;

i posteriorment s'introduiran alguns dels mètodes més comuns en l'anàlisi multivariant.

5.2.1. Relació entre dues variables quantitatives

1) Mesures descriptives de la relació entre variables quantitatives: covariància i correlació

Igual que en l'apartat anterior s'ha exposat el concepte de variància, existeix una altra mesura de síntesi que implica dues variables alhora i, per tant, és el punt de partida per a l'anàlisi descriptiva de l'existència o no de la relació entre dues variables quantitatives: aquesta mesura és la covariància.

La covariància mesura les distància de les observacions de dues variables en relació amb la seva mitjana respectiva, de manera simultània. D'aquesta manera, l'indicador recull el fet de si les dues variables mantenen un comportament paral·lel, en el qual es pugui detectar algun tipus de relació, o no. El seu símbol, paral·lelament al de la variància, és S_{XY}; en comptes de tenir un quadrat i una sola variable, veiem com en té en compte dues. Per tant també s'acabarà expressant en unitats «quadràtiques», en el sentit que el seu valor serà la multiplicació de dues unitats diferents (les corresponents a X i a Y), probablement sense sentit en la seva interpretació, igual que succeïa en el cas de la variància. Les fórmules, definitòria i operativa, són al seu torn, les següents:

$$S_{xy} - \frac{\sum_{i=1}^{n}(X_i - \overline{X}) \cdot \left(Y_i - \overline{Y}\right)}{n}$$

$$S_{xy} - \frac{\sum_{i=1}^{n}(X_i - Y_i)}{n} - \overline{X} \cdot \overline{Y}$$

Exemple

Agafant les dues variables que hem anomenat X i Y en l'exemple de la mostra de n=10 visitants al parc (recordem: X = nombre de visites al parc; Y = nombre de persones que integren el grup de visita), a continuació il·lustrarem com es calcularia la seva covariància. Farem servir la fórmula operativa; a més, com que tenim les dades resumides ja en taules de freqüències (igual que es va explicar per al cas de la mitjana aritmètica i la resta de mesures de síntesi), en comptes de sumar cada parell de dades, al numerador de la fórmula directament hi calcularíem els globals amb $\sum X_i.Y_j.n_{ij}$. És a dir, aprofitant que sabem la freqüència absoluta conjunta, simplement agafaríem cada parell de valors i el multiplicaríem pel nombre de vegades que apareixen en les respostes.

$X_i \setminus Y_j$	1	2	3	$n_{i.}$
1	1	2	2	5
2	2	1	0	3
3	0	1	0	1
4	0	0	0	0
5	0	1	0	1
$n_{j.}$	3	5	2	$\sum\sum n_{ij}=10$

Així doncs, per exemple, per al valor simultani de Y = 1 i de X = 1, la freqüència absoluta conjunta és igual també a 1; igual que per al valor simultani de Y = 2 i de X = 4, la freqüència absoluta conjunta és igual també a 0; recollint totes les possibles combinacions, i recordant que tant la mitjana de X com la mitjana de Y en aquest cas són coincidents i igual a 1,9, obtenim:

SXY = [[(1x1x1) + (1x2x2) + (1x3x0) + (1x4x0) + (1x5x0) + (2x1x2) + (2x2x1) + (2x3x1) + (2x4x0) + (2x5x1) + (3x1x2) + (3x2x0) + (3x3x0) + (3x4x0) + (3x5x0)] / 10] - 1,9 x 1,9 = [35 / 10] – 3,61 = -0,11

Pel que fa a la interpretació de la covariància, igual que passava amb la variància, cal tenir present per començar que és una mesura expressada en unes unitats que no tenen interpretació, sense sentit. A més, només el fet d'estar expressada en unitats ja implica que podem tenir problemes d'interpretació en funció de la magnitud de la variable. Dit d'altra manera: aquest -0,11 obtingut per l'exemple exposat, no podem saber si implica molta o poca relació entre les variables. En canvi, conclusions que sí que podem extreure'n:

- la covariància és diferent a zero: una $S_{XY} = 0$ implica independència estadística entre les dues variables. Per tant, en l'exemple anterior no podem concloure si la relació és intensa o feble, però sí que podem dir que no hi ha una independència, si més no, una independència absoluta;

- en cas de verificar-se una relació mínimament sòlida (no podem interpretar-ho encara, com ja s'ha dit), aquesta relació seria de tipus invers o negatiu, a causa del signe de la covariància obtinguda, en aquest cas menor a zero o negatiu. Això significaria que quan una de les variables creix, l'altre decreix, i viceversa.

Finalment, com ressolem doncs els problemes d'interpretació de la covariància? La resposta és fent servir el coeficient de correlació de Pearson, r_{XY}.

El coeficient de correlació és una mesura de síntesi que recull la relació lineal (recordem el supòsit fet al principi: estrictament de caire lineal) existent entre dues variables quantitatives, amb uns valors que oscil·len entre -1 i +1. La interpretació del seu resultat és la següent:

1) $r_{XY} = +1$: correlació lineal directa o positiva perfecta. Les dues variables estan perfectament correlacionades (fixem-nos que encara no hem entrat en l'estudi de causalitat, no sabem quina és causa i quina és efecte; només ens centrem en si existeix relació i com és), i a més ho fan de manera positiva: ambdues evolucionen en el mateix sentit.

2) $r_{XY} = 0$: implica independència lineal. No hi ha relació lineal entre ambdues variables.

3) $r_{XY} = -1$: correlació lineal inversa o negativa perfecta. Les dues variables estan perfectament correlacionades, però en aquest cas evolucionen en sentit invers.

Val a dir que una relació de +1 o -1 és igual de forta i perfecta; l'únic que canvia és el signe de l'evolució conjunta dels valors de les variables considerades.

Així doncs, el coeficient de correlació lineal, en oferir-nos un valor sostre, ens permet una interpretació respecte de si la relació entre les variables és forta o feble. En aquest sentit, com més a prop estigui el nostre resultat per a r_{XY} de la unitat, en valors absoluts, més forta serà la relació; en cas de ser proper al zero, parlarem d'una relació feble.

El càlcul de r_{XY} és el resultat de dividir la covariància entre el producte de les desviacions estàndard de X i de Y, de manera que les unitats de numerador i denominador queden, efectivament, anul·lades:

$$r = \frac{S_{XY}}{S_X S_Y}$$

Exemple

Seguint amb les dades obtingudes anteriorment, per a X i Y, sabem que $S_{XY} = -0,11$; $S_X = 2,4$; $S_Y = 3,3$; per tant r_{XY} serà:

$r_{XY} = -0,11 / (2,4 \times 3,3) = -0.01$

Fixem-nos en primer lloc que, atès que les desviacions estàndard són les arrels quadrades positives de la variància, el denominador sempre resultarà positiu. Com a conseqüència, el coeficient de correlació mantindrà el mateix signe que la covariància (al numerador).

En relació amb el resultat obtingut, ara sí que observem que ens trobem molt a prop del zero; per tant, amb un resultat de r_{XY} = -0,01 pràcticament es pot dir que no hi ha relació entre ambdues variables. Aquesta conclusió no la podíem extreure a partir de la covariància.

2) La recta de regressió lineal. Estimació per MQO

La metodologia per excel·lència per a tractar la relació de causalitat sota el supòsit de linealitat és la de l'estimació de la recta de regressió lineal entre dues variables quantitatives.

Abans, però, introduïm dos nous conceptes: el de variables explicades o endògenes i el de variables explicatives o exògenes.

- **Variable explicada o endògena (dependent)**: és la variable que volem estudiar, aquella que volem conèixer, tot explorant-ne l'evolució o comportament, i mirant de determinar quin o quins són els factors que condicionen o expliquen aquest comportament;

- **Variable explicativa o exògena (independent)**: és o són les variables que haurem d'estudiar per tal de determinar quin grau d'influència tenen sobre la variable explicada o endògena.

Exemple

Imaginem que l'Administració local del terme on se situa el parc de l'exemple que estem fent servir ens demana una investigació sobre el nombre de turistes que pernocten a la zona, durant un període de deu anys consecutius. La nostra variable endògena serà, doncs, aquesta mateixa. Suposem que en un dels anys del període detectem una davallada en la xifra i volem conèixer quins són els factors que poden ser-ne la causa. Després d'estudiar el context, arribem a la conclusió que alguns factors possibles serien: un increment en els preus de la destinació; una disminució en el creixement econòmic global; un canvi en les polítiques de conservació i protecció del parc, i l'estancament en el nombre de connexions i vols de l'aeroport més proper a la destinació. Doncs bé, aquestes quatre variables (evolució dels preus en la destinació, taxa de creixement econòmic global, canvi en les polítiques de conservació del parc, nombre de connexions/vols a la destinació) serien quatre possibles variables explicatives o exògenes, sobre les quals caldria centrar la recerca.

Aquesta tècnica consisteix, en definitiva, a obtenir una recta que expliqui aquesta relació, partint d'un model o funció matemàtic, lineal, en el qual una variable (la independent i exògena: X), explicaria la variable objectiu (dependent o endògena: Y). Per tant, a partir de la formulació, donat un determinat valor de X, podríem aproximar el valor esperat de Y. L'expressió de la funció seria:

$$Y = a + b\,X + (e)$$

a on:

- a seria el valor de la constant o ordenada en l'origen (és a dir, el valor de Y quan x = 0);

- b seria el valor del pendent de la recta, recollint així la relació entre les dues variables (positiva o negativa);

- (e) seria un residu o error degut a altres factors no introduïts en la funció.

Així doncs, donat un parell de variables X i Y, n'hi haurà prou de conèixer el valor de a i b per tal d'aproximar la seva recta de regressió, a partir de la seva relació de causalitat, i per tant poder definir el comportament de Y en funció dels diversos escenaris de X.

Tanmateix, com calculem aquests valors de a i b? L'estadística demostra que una bona resposta és l'aplicació de l'estimació d'aquests valors a través del mètode dels mínims quadrats ordinaris (MQO). No n'exposarem el desenvolupament, però sí que és important saber que les fórmules que s'explicaran a continuació procedeixen d'aquesta metodologia, ja que garanteix que els valors obtinguts seran òptims. Per tant, a partir dels desenvolupaments corresponents a l'estimació MQO, arribem a saber que:

$$a = \overline{Y} - b\overline{X}$$

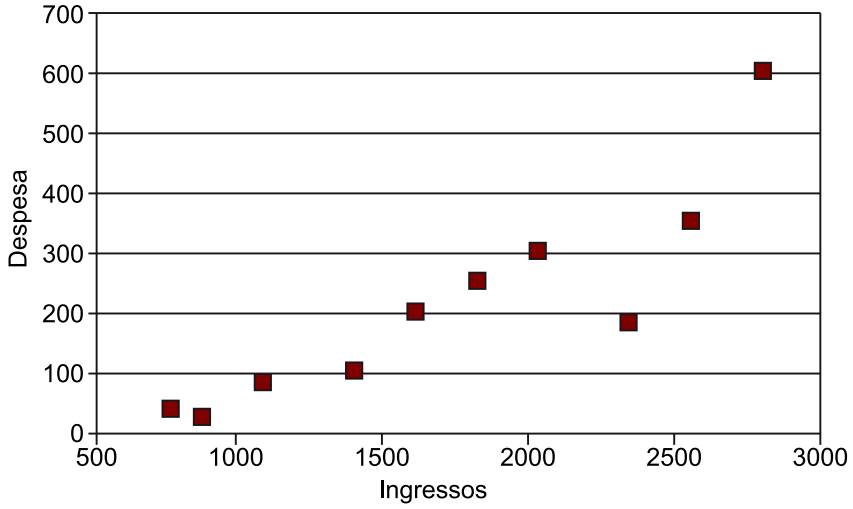
$$b = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum X_i Y_i - n\overline{X}\,\overline{Y}}{\sum X_i^2 - \overline{X}}$$

Exemple

Treballarem ara amb les altres dues variables ja introduïdes en exemples anteriors a partir de la mateixa enquesta a n=10 visitants al parc: ingressos mitjans i despesa mitjana en lleure, cultura, turisme, etc.

Recordem que les dades i el núvol de punts obtinguts eres els següents:

ING	2.500	1.400	1.100	2.300	800	1.600	900	2.000	3.200	1.800
DESP	350	100	80	180	30	200	20	300	600	250



El que ara cerquem és una recta que resumeixi la relació entre les dues variables, amb un sentit de causalitat. Els ingressos serien en aquest cas la nostra X (variable exògena) i la despesa, la nostra Y (variable endògena).

Volem, doncs, conèixer el valor de a i b que compleixin de millor manera que:

DESP = a + b * ING + (e)

Evidentment, les rectes que poden resumir aquesta relació són infinites, però només una és l'òptima. Per a obtenir-la, aplicarem les fórmules exposades abans.

Començarem pel càlcul de b; per a això, procedirem a fer alguns càlculs intermedis, necessaris per a aplicar la fórmula:

Observació	ING (X)	DESP (Y)	X*Y	X^2
1	2.500	350	875.000	6.250.000
2	1.400	100	140.000	1.960.000
3	1.100	80	88.000	1.210.000
4	2.300	180	414.000	5.290.000
5	800	30	24.000	640.000
6	1.600	200	320.000	2.560.000
7	900	20	18.000	810.000
8	2.000	300	600.000	4.000.000
9	2.800	600	1.680.000	7.840.000
10	1.800	250	450.000	3.240.000
Suma	17.200	2.110	4.609.000	33.800.000

Per tant:

X = 1.720
Y = 211

Aplicant doncs la fórmula:

b = [(4.609.000 – 10 x 1.720 x 211) / (33.800.000 – 1.7202)] = 979.800 / 30.841.600 = 0,03177

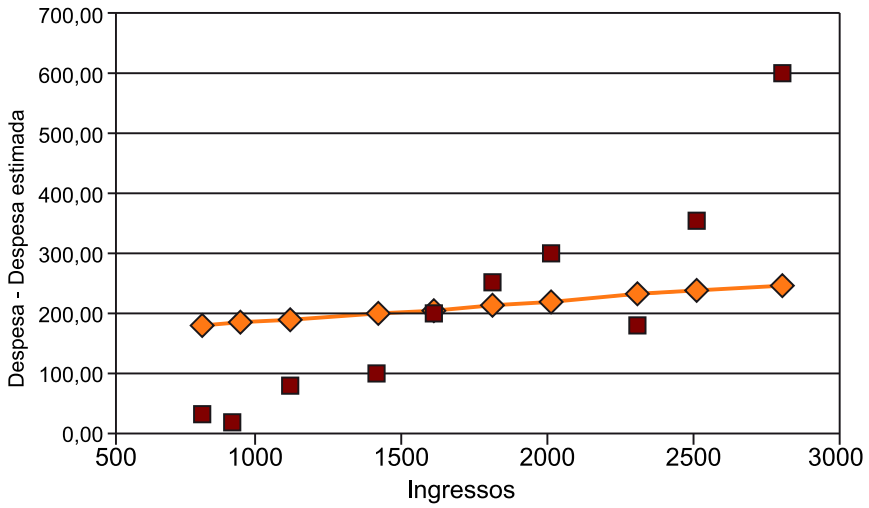
Veiem que s'obté un pendent baix, poc pronunciat, però positiu. A continuació ja podem també calcular a:

a = 211 – 0,03177 x 1.720 = 156,35

En conclusió, obtenim que la millor recta per a explicar la relació de causalitat entre X i Y és DESP = 156,35 + 0,032 x ING. Si ara agafem les dades originals dels ingressos (ING o X) i les substituïm a la fórmula, obtindrem els valors de la recta:

Observació	ING (X)	DESP (Y)	X*Y	X^2	DESP_Est (Y)
1	2.500	350	875.000	6.250.000	235,78
2	1.400	100	140.000	1.960.000	200,83
3	1.100	80	88.000	1.210.000	191,30
4	2.300	180	414.000	5.290.000	229,43
5	800	30	24.000	640.000	181,77
6	1.600	200	320.000	2.560.000	207,19
7	900	20	18.000	810.000	184,95
8	2.000	300	600.000	4.000.000	219,90
9	2.800	600	1.680.000	7.840.000	245,31
10	1.800	250	450.000	3.240.000	213,54
Suma	17.200	2.110	4.609.000	33.800.000	

Evidentment aquests valors no estan exempts d'errors, com es deia al principi, i es recull a la formulació teòrica, amb el terme d'error o pertorbació (e). Tanmateix, esdevé la millor forma d'ajust que podem proporcionar. Ara podem representar gràficament els valors originals, i els resultants de la recta, tal com apareix al següent gràfic.



La interpretació dels resultats obtinguts seria la següent: quan els ingressos són igual a zero, la despesa se situa en 156,35 euros; d'altra banda, per cada euro més d'ingrés, la despesa pujaria 0,032 euros.

Tanmateix, no sempre aquesta recta és vàlida o explicativa: és a dir, malgrat que la metodologia dels MQO ens assegura l'obtenció de la recta òptima, això no significa que la recta faci un bon ajust, simplement pel fet que pot ser que

no hi hagi la relació causal que nosaltres estem cercant o que suposem sobre la seva existència. Com mesurem doncs si aquest ajust és bo o no? Com sabrem si la relació de causalitat existeix i és prou forta per a poder fer servir l'ajust mitjançant la recta de regressió? Per a donar resposta a aquestes preguntes, calculem el coeficient R^2, o de bondat de l'ajust.

Existeixen diverses fórmules per a calcular R^2, però com que ja coneixem com es calcula r_{XY}, és a partir d'aquest indicador que calcularem la R^2. Així, R^2 no és més que el quadrat del coeficient de correlació lineal r_{XY}, és a dir:

$$R^2 = (r_{XY})^2$$

Per tant, R^2 només podrà prendre valors positius, entre 0 i 1. Com més proper a 0, més dolent serà l'ajust; per contra, com més proper a 1, millor serà l'ajust obtingut. Com a orientació, cal tenir en compte que els estudis acadèmics no acostumen a acceptar ajustos per sota del 0,8 (80% d'ajust), i de fet, la recerca normalment prossegueix amb l'ànim de no treballar amb ajustos menors a 90-95%, tot i que no sempre és possible.

Exemple

Per a l'ajust fet entre les variables d'ingressos i despesa, podem calcular la bondat de l'ajust, obtenint prèviament el coeficient de correlació lineal r_{XY}. Per això haurem de calcular la S_{XY}, la S_X i la S_Y. A continuació plantegem com a exercici aquest càlcul, tot oferint-ne directament el resultat:

$S_{XY} = 97.980$; $S_X = 649,31$; $S_Y = 167,3$. Per tant $r_{XY} = 0,90$. I finalment, caldrà tan sols elevar aquest resultat al quadrat, i s'obtindrà que $R^2 = 0,81$. Es pot dir, en conclusió, que la recta obtinguda ofereix un ajust bo. No seria un ajust excel·lent, però sí que aconsegueix explicar el 81% del comportament de la variable que estem estudiant, en aquest cas, la de la despesa a partir de la de l'ingrés.

5.2.2. Relació entre dues variables qualitatives

La relació entre variables també es pot produir quan aquestes són de naturalesa qualitativa. Tanmateix, no es pot valorar l'existència de relació o no entre aquestes variables de la mateixa manera que en el cas de les variables quantitatives, i caldrà veure quines són les mesures d'associació que haurem d'utilitzar.

A fi de poder observar el resultat de les dues variables a estudiar, en primer lloc, caldrà elaborar una taula de doble entrada o bidimensional en la qual es podrà veure el nombre d'observacions que corresponen a cada combinació de valors, tal com ja hem vist abans. Per a poder fer la interpretació més fàcilment és habitual disposar de la taula en termes relatius, i la interpretació es pot fer per files o columnes, en funció de quina és la variable que volem explicar.

Amb la taula de doble entrada només podem observar el resultat de cadascuna de les variables en ser creuada amb l'altra, però no podem determinar quin és el grau d'associació en termes estadístics entre ambdues variables. Per a fer això necessitem el càlcul de dos estadístics que ens proporcionaran informació diferent, però complementària sobre la relació entre les variables.

1) Test de khi-quadrat (χ^2) – inferència estadística

El test de khi-quadrat (χ^2) consisteix a calcular la suma de la diferència de les freqüències observades (O_i) i esperades (E_i) al quadrat, dividides per les esperades (E_i). Les freqüències esperades són aquelles que s'haurien produït en el cas que la hipòtesi d'independència fos certa. La fórmula d'aquesta prova és la següent:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)}{E_i}$$

El resultat d'aquest estadístic ens donarà informació de si les dues variables són independents o, per contra, estan relacionades en la població, entenent que uns dels objectius de l'estadística és poder inferir els resultats a la resta de la població. Així doncs, quan el resultat sigui 0 ens indicarà que no hi ha diferència entre les freqüències observades i esperades i que, per tant, es tractarà de variables independents.

Per contra, com més diferència hi hagi entre les freqüències observades i esperades, més gran serà el valor que adquirirà χ^2, i, per tant, es podrà concloure una major relació entre ambdues variables. No obstant això, aquest estadístic no té un límit màxim, de manera que és difícil poder determinar si la dependència és total o no.

Exemple

Agafant dues variables que anomenem X i Y, en la qual X = tipus d'allotjament i Y = tipus de reserva, vegem com es calcularia el test del χ^2.

Xi \ Yj	Internet	Directa	ni.
Hotel	30	20	50
Càmping	10	30	40
nj.	40	50	∑∑ nij=90

En primer lloc, caldrà procedir al càlcul de les freqüències esperades, per a les quals caldrà calcular el total de cada columna (n_j) multiplicat pel total de cada fila (n_i) i dividit per la suma total (∑∑ nij): E_1 = [(40x50)/90]; E_2 = [(40x40)/90]; E_3 = [(50x50)/90]; E_4 = [(50x40)/90]. Una vegada es disposa dels valors esperats ja es pot procedir al càlcul de l'estadístic χ^2, aplicant la fórmula que s'ha vist abans. El resultat serà d'11,025 i ens indica que efectivament les dues variables estan relacionades, ja que el valor és superior a 0. Però, com bé hem dit abans, no podem dir si hi ha molta o poca relació.

χ^2=[(30-22,2)2/22,2]+ [(20-27,8)2/27,8]+ [(10-17,8)2/17,8]+ [(30-22,2)2/22,2]= 11,025

2) L'estadístic V de Cramer

Per a poder solucionar el problema de manca de precisió en el grau de dependència de les dues variables analitzades podem utilitzar l'estadístic V de Cramer, que ens proporciona el grau de relació entre les dues variables en la mostra, el qual es deriva del càlcul del test khi-quadrat (χ^2), i es calcula de la manera següent:

$$v = \sqrt{\frac{x^2}{N(q-1)}}$$

El resultat d'aquest estadístic es troba entre 0 i 1, de manera que quan prengui valors de 0 o propers a aquest indicarà independència de les dues variables (les variables no estan relacionades), i, per contra, quan el valor sigui d'1 o proper a aquests indicarà que hi ha relació entre ambdues variables analitzades.

Exemple

Seguint amb l'exemple anterior, l'aplicació de la fórmula en la qual N és la mida de la mostra i *q* la diferència entre el nombre de files i columnes (*v*=11,025/90[(2-2)-1]) ens donaria un resultat de 0,350, la qual cosa indicaria que efectivament hi ha relació entre ambdues variables, malgrat que aquesta dependència no és gaire gran.

5.2.3. Relació entre una variable qualitativa i una variable quantitativa

1) Anàlisi de la variància

Igual que en els casos anteriors, també pot interessar constatar si hi ha o no relació entre una variable quantitativa i una variable qualitativa. En aquest cas l'ANOVA (anàlisi de la variància) permet fer una comparació entre les mitjanes de la variable quantitativa i les categories de la variable qualitativa, la qual dona com a resultat la variància intragrups (dins de cada grup) i intergrups (entre grups).

És a dir, es procedeix a calcular la variància intragrups (recordem que *s* és la desviació típica):

$$V_{intra} = \left(n_1 - 1\right)s_1^2 + \ldots + \left(n_{1t} - 1\right)s_t^2$$

I la variància intergrups:

$$V_{inter} = n_1\left(\bar{x}_{-1} - \overline{x}\right)^2 + \ldots + n_t\left(\bar{x}_t - \overline{x}\right)^2$$

De manera que es permet calcular la variació total:

$$V_{total} = V_{intra} + V_{inter}$$

2) Estadístic eta (η)

En aquest cas la mesura d'associació que utilitzarem és l'estadístic eta-quadrat (η), el qual mesura quin efecte té la variable qualitativa (independent) sobre la variable quantitativa (dependent) en la mostra. Per exemple, determina si una persona que utilitza hotel (variable independent) és més probable que faci una estada curta o llarga (variable dependent). Aquest estadístic pren valors entre 0 i 1, on 0 significa que no hi ha relació i 1 que hi ha una plena relació. És a dir, quan el valor és 0 indica que els individus d'un mateix grup tots són diferents entre ells; en canvi, quan el valor és 1, indica que els individus d'un mateix grup tenen un patró de conducta similar.

$$\eta = \sqrt{\frac{V_{nter}}{V_{inter} + V_{intra}}}$$

Exemple

Agafant dues variables que anomenem X i Y, en la qual X = tipus d'allotjament i Y = nombre de nits, vegem com es calcularia l'estadístic η

Xi \ Yj	$x\bar{}_j$	n_j	s_j	$(n_j-1)s_j^2$	$n_j(x\#_j - x\#)^2$
Hotel	1,94	50	0,712	25,347	59,405
Càmping	4,43	40	1,107	49,018	78,4
Turisme Rural	2,90	10	0,316	0,998	0,169
Total	3,03	100		75,364	137,974

Per tant, la variació total serà de 213,34 (75,364 + 137,974). Així doncs, aplicant la fórmula que hem vist abans l'estadístic η tindrà un valor de 0,5943, el qual indica l'existència de relació entre ambdues variables.

η = √[75,364/(76.364+137.974)

5.2.4. Anàlisi multivariant

Els mètodes d'anàlisi multivariant, com el seu nom indica, impliquen l'estudi simultani de més de dues variables.

No exposarem aquí totes les tècniques multivariant que existeixen per a tractar la informació disponible; de fet constantment apareixen eines i metodologies noves i més sofisticades. Seria del tot impossible fer-ne un repàs exhaustiu. Tanmateix, sí que hi ha un seguit de tècniques que, per la seva especificitat pel que fa a objectius a assolir, i per la freqüència amb que són emprades en la recerca aplicada, cal conèixer i entendre.

És important tenir en compte que en funció de quin sigui el nostre objectiu d'investigació, serà interessant aplicar una eina o una altra.

1) Model de regresió lineal múltiple

Cal dir, en primer lloc, que l'anàlisi de regressió, el model de regressió, vist en apartats anteriors (anàlisi causal i recta de regressió per anàlisi temporal univariant), té la seva ampliació també en termes multivariants. Així, podem dir que un dels usos freqüents de l'anàlisi multivariant és també l'anàlisi causal, a través de les relacions que dues o més variables tenen amb una principal, que és la que volem explicar (variables exògenes respecte de la variable endògena o explicada).

Exemple

Anteriorment s'ha exposat el model de regressió simple, amb l'exemple de la variable Ingressos dels visitants a un parc, com a única variable explicativa de la Despesa en lleure, consum, turisme, etc. d'aquests mateixos visitants. Tanmateix, tots sabem que no sempre persones amb els mateixos ingressos tenen una despesa igual, i que caldria incloure altres qüestions, com per exemple l'edat, el nombre de membres de la unitat familiar, el percentatge dels ingressos que cal destinar a habitatge, etc.

Una anàlisi causal de tipus multivariant, per tant, tindria com a objectiu quantificar aquestes relacions de manera simultània, a partir d'una formulació ampliada del model de regressió, com s'exposa en la següent expressió:

DESP = a + b1xING + b2xEDAT + b3xMEMBRES_FAMIL + b4xDESP_HABITATGE

2) L'anàlisi factorial: la reducció de les dimensions

D'altra banda, trobem tot un conjunt de tècniques multivariants, la finalitat de les quals és la de poder treballar amb moltes variables i informació, mitjançant la seva reducció o simplificació en noves variables. Aquestes noves variables serien menys en nombre, però caldria que recollissin al màxim el comportament del conjunt de variables originals. Es tractaria, en definitiva, de poder reduir les dimensions per a incorporar el màxim d'informació, però sense que aquesta esdevingui tan complexa i diversa que n'impossibiliti el tractament. Aquestes eines són molt utilitzades en la recerca del turisme, quan, per exemple, volem aglutinar molta informació de diverses variables (diferents tipologies d'allotjament, diferents tipologies de recursos tractors, agents que participen en el procés turístic, aspectes relatius a l'entorn i infraestructures, variables de mesura d'impactes ambientals...), i mirar de resumir-la en un màxim de 3, 4 o 5 vectors que ens permetin treballar d'una manera còmoda i àgil.

Una d'aquestes tècniques és l'anàlisi factorial. Es tracta d'una metodologia estadística de reducció de dades que, a partir de l'estudi de les correlacions entre les variables observades, extreu un nombre menor i més reduït de factors, que recullen el màxim d'informació continguda a les variables originals.

Per tant, partint d'un contingent ampli de variables cadascuna de les quals aporta una informació diferent però que es troben correlacionades, es combinen de manera que s'obtenen en el seu lloc un nombre menor de factors, que serà amb els que es treballarà finalment, i sempre mirant de minimitzar la pèrdua d'informació.

Exemple

Seguint sempre amb l'exemple del parc natural, imaginem que volem analitzar-ne la competitivitat i comparar-la amb la d'altres parcs naturals per a saber-ne el posicionament.

Per a dur a terme aquesta recerca prendrem un conjunt de n parcs naturals i definiríem les variables que volem incorporar. Aquestes variables poden incloure la descripció de la demanda (visitants que pernocten a la zona, excursionistes, despesa feta, repetició de la visita, durada de l'estada...); la descripció de l'oferta turística de la zona (allotjament, restauració, oficines de turisme, accessos...); la descripció dels atractius i oferta del parc (serveis, senyalització, recursos destacables, quilòmetres d'extensió...); l'altra oferta d'interès (museus, centres d'interpretació, llocs històrics i monuments, festes populars i tradicions...), etc. Com veiem, tindrem un contingent molt considerable de variables, fins i tot sovint tindrem més variables per incorporar com a explicatives que la mateixa mida n d'individus que volem explicar, la qual cosa genera una inconsistència greu que du a models irresolubles.

Després de l'aplicació de l'anàlisi factorial i a través de mètodes com els components principals, el conjunt de variables seleccionades es veuria reduït a uns quants factors, que recollirien aspectes d'interpretació comuna, com podrien ser els relatius a l'atractivitat, l'accessibilitat (en el sentit ampli del terme), o l'interès natural i cultural de cadascun dels parcs. Això implica que ja disposaríem d'unes poques i noves variables (factors) que, ara sí, permetrien fer una anàlisi d'altre tipus, ja amb un nombre raonable de variables caracteritzant cada individu, o, en aquest cas, cada parc. Cal dir, finalment, que aquestes metodologies es treballen, com es pot intuir, amb programes específics. En aquest sentit, dins de la informació que ens ha de ser retornada quan executem aquest tipus de tècnica, a banda evidentment dels resultats principals (factors que expliquen i resumeixen les variables originals), se'ns informa també del total de la variabilitat conjunta que contenen les variables inicials que ha estat retinguda (i que per tant queda explicada) pels factors finalment triats. D'aquesta manera, sabem quanta informació hem perdut en el procés i si ens interessa el resultat o bé cal mirar de retenir més informació total –normalment, augmentant doncs el nombre de factors–.

3) L'anàlisi clúster: l'agrupació d'individus

L'anàlisi clúster és un conjunt de tècniques multivariants utilitzades per a classificar un conjunt d'individus en grups el màxim d'homogenis. S'inclou en el conjunt de tècniques que té per objectiu l'agrupació i classificació dels individus en una sèrie de grups, que no són predeterminats *a priori* sinó que precisament sorgeixen del mateix procés, de l'aplicació de la tècnica.

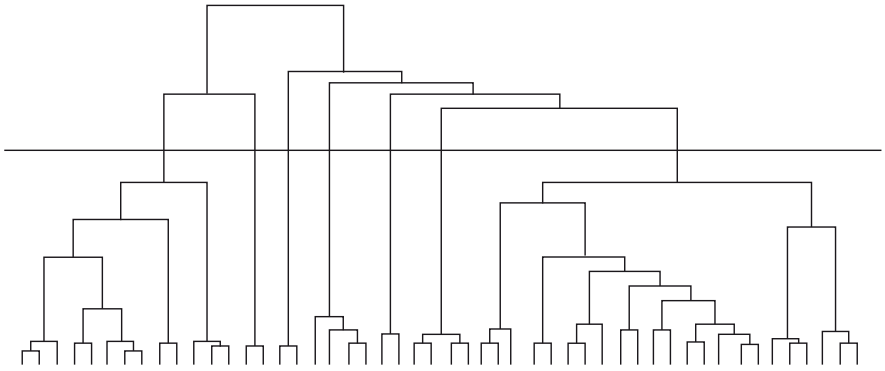
Per a l'anàlisi clúster es parteix d'un conjunt d'individus (mida = n), dels quals es coneix tot un seguit de característiques; aquestes característiques, com sempre, configuren unes variables. Doncs bé, l'objectiu d'aquesta metodologia és classificar aquests individus a partir de les variables recollides, de manera que

de forma lògica els individus més similars entre ells acabin conformant un grup o conglomerat, i s'obtinguin com a resultat grups al més uniformes possible, a nivell intern, i diferenciats entre ells.

Evidentment, la configuració dels grups vindrà donada en funció de les variables que fem servir: per tant, és fonamental que aquestes variables responguin correctament als objectius de la recerca que estem duent a terme. Un altre aspecte important és el fet que el clúster, en el seu nivell de màxima desagregació, parteix de la situació en la qual cada individu seria un grup bàsic. A partir d'aquí, i a partir constarà d'un algoritme de classificació, els individus es van unint i classificant de manera progressiva i per fases, podríem dir, de forma que si no aturem la iteració, al final obtindríem tot el contrari que a l'inici, és a dir, un únic grup contenint tots els individus. Per tant, caldrà sempre establir criteris per aturar l'agrupació en un determinat nivell.

Els resultats de l'anàlisi clúster es visualitzen d'una manera molt àgil i còmoda amb la que és la seva representació gràfica específica: el dendrograma. Amb la forma d'arbre invertit, el dendrograma il·lustra com, començant a partir de n individus o grups elementals, es van produint les agrupacions per la via de l'aplicació de l'algoritme de classificació. En el dendrograma es parteix d'aquests n individus inicials, i clou a la part alta (si és vertical) o a la part dreta (si és horitzontal) amb un únic grup que els conté a tots. Es pot apreciar a més en quin nivell d'iteració del procés una determinada observació o individu s'ha afegit a un grup determinat; i es pot també prendre la decisió d'on aturar el procés de classificació. La conclusió és que el dendrograma resulta de gran utilitat per a la correcta interpretació dels resultats obtinguts.

A la següent imatge es reprodueix un dendrograma, en aquest cas en sentit vertical. Com es pot veure, per a aquest cas, disposem de 50 individus, que conformarien els 50 punts inicials de la base, d'on comencen a sortir les línies d'agrupació cap amunt.



Si ens situem en el primer nivell d'agrupació (les agrupacions es produeixen quan dos individus o grups ja formats en una fase anterior s'uneixen a través d'una línia horitzontal), i mirant el gràfic començant des de la part inferior, i d'esquerra a dreta, podem observar que l'observació 1 i la 2 s'uneixen ja en

aquest primer estadi, mentre que l'observació 3 queda a banda; no serà fins a una propera iteració (més amunt) que l'observació 3 s'unirà al grup creat en primera instància per les observacions 1 i 2. Fent aquesta mateix lectura cap amunt, arribem, com es veu, a l'extrem superior, on efectivament totes les observacions inicials, corresponents als 50 individus, conformen un únic grup. Com que, com es deia, la finalitat és aconseguir definir alguna classificació, no interessa arribar a aquest extrem. Aturar-nos en el penúltim esgraó (de nou ara llegint el gràfic per la banda superior) tampoc sembla òptim, ja que potser dos grups per a explicar 50 observacions implica una simplificació excessiva. És per això, doncs, que, com es pot veure en el dendrograma, s'ha dibuixat una línia horitzontal que ens marcaria el punt de tall, el moment on decidim aturar el procés de classificació. Llegint doncs el gràfic a partir d'aquesta línia divisòria cap avall, veiem que, en el cas que s'il·lustra, s'ha optat per una classificació en 7 grups, que concretem a continuació:

- Grup 1. Conté les observacions de la 1 fins a la 13
- Grup 2. Conté les observacions 14 i 15
- Grup 3. Conté les observacions 16 i 17
- Grup 4. Conté les observacions de la 18 fins a la 21
- Grup 5. Conté les observacions 22 i 23
- Grup 6. Conté les observacions des de la 24 fins a la 27
- Grup 7. Conté les observacions des de la 28 fins a la 50

5.3. Anàlisi descriptiu de sèries temporals

Parlem d'anàlisi univariant quan només estudiem una variable i obtenim el màxim d'informació a partir d'aquesta única variable. A l'apartat d'estadística descriptiva ja s'ha explicat com l'estadística elemental ens ajuda a resumir i sintetitzar la informació de cadascuna de les variables una per una. Però fins ara hem parlat de variables que eren recollides en un mateix moment del temps, per a diferents individus: a aquest tipus d'estudi l'anomenem estudi transversal o *cross-section*. Amb l'anàlisi univariant s'introdueix una visió diferent, ja que aquesta tècnica –o conjunt de tècniques– s'associa amb l'estudi de les sèries temporals, o el que és el mateix, observem el valor que pren una mateixa variable, per a un mateix individu, però al llarg del temps, en diferents períodes. Aquesta anàlisi es coneix com a anàlisi de sèries temporal o *time-series*.

Exemple

Si analitzem el nombre de visitants que varen rebre el darrer any el conjunt de parcs naturals d'un territori (cada parc seria, doncs, un individu de la nostra investigació), estaríem fent una recerca transversal. En canvi, si agafem un únic parc i analitzem el nombre de visitants que ha rebut al llarg dels darrers 25 anys, estaríem parlant d'una recerca temporal o evolutiva.

Definim sèrie temporal, per tant, com el conjunt d'observacions d'una variable determinada per a diferents moments del temps. És important tenir en compte que aquestes observacions s'han de fer a intervals regulars de temps. Així, per a una variable Y, obtindríem que la seva sèrie temporal seria:

$Y = [Y_1, Y_2, Y_3,..., Y_t,..., Y_T]$

on el subíndex «t» indica el moment del temps al qual es refereix un valor, i T seria el nombre total de períodes disponibles.

En l'àmbit del turisme tenim accés a una gran quantitat de dades estadístiques relatives a variables que es troben mesurades en diferents moments del temps. Doncs bé, l'anàlisi univariant, i d'aquí el seu nom, es centra a estudiar una variable de manera aïllada, a partir únicament de l'anàlisi de la seva història passada, tant per a conèixer-ne el comportament com per a poder establir-ne escenaris i prediccions de futur.

L'anàlisi univariant o de sèrie temporal no és una sola metodologia sinó que incorpora un seguit de tècniques diverses, cadascuna de les quals resulta més adient per al tractament d'una o altra sèrie, en funció de les seves característiques. En tot cas un dels aspectes fonamentals en l'anàlisi univariant és la distinció dels components que conformen –o que poden conformar– una sèrie temporal, i que es descriuen tot seguit.

1) Tendència: representa l'evolució a llarg termini de la sèrie. Aquesta evolució pot ser creixent o decreixent. La tendència recull i ens revela aspectes com els relatius a canvis demogràfics de llarg recorregut, amb els consegüents increments de demanda i oferta de productes i serveis; les millores tecnològiques, els canvis i avenços en eficiència i innovació; o les transformacions estructurals en els hàbits dels consumidors. Per exemple, en el moment actual es detecta que hi ha un canvi de tendència important pel que fa a la sensibilitat envers la sostenibilitat ambiental i cultural.

2) Cicle: recull els moviments oscil·latoris per sobre o per sota de la tendència. La seva durada es mesura des d'un cim (part més alta del cicle) fins al següent cim, o bé des d'una vall (part més baixa del cicle) fins a la següent vall. La durada del cicle no és constant, però sempre serà superior a un any i es deu a canvis en l'activitat econòmica, generalment. Sovint és difícil separar la tendència del cicle.

3) Estacionalitat: recull les oscil·lacions d'una sèrie temporal que es completen dintre d'un any i que es repeteixen en anys successius. El període del factor estacional és inferior a l'any i presenta una forta estabilitat any rere any. Cal tenir present que perquè la informació estudiada pugui presentar un comportament estacional, les dades han de fer referència a períodes inferiors a l'any, és a dir, cal que es tracti de dades mensuals, trimestrals, semestrals, etc. Pel que fa a les causes de l'existència de factor estacional, en trobem fonamentalment

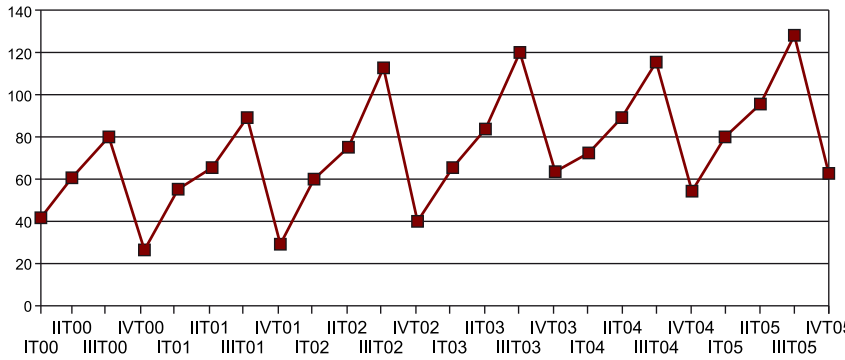
de dos tipus: factors físics naturals, com la meteorologia, els cicles biològics, etc., que afecten, per exemple, la producció agrícola de determinats productes, o la demanda de certs productes turístics (turisme d'hivern enfront de turisme d'estiu), i factors institucionals, com festes, vacances escolars i altres. En definitiva, es tracta de l'efecte del calendari. En el turisme aquest component és especialment rellevant.

4) Component irregular: són aquelles variacions d'una sèrie temporal que no estan recollides en els tres components anteriors i que tenen un caràcter residual. És altament aleatòria i per tant difícil de predir. Aquí s'inclouen per exemple factors climatològics inesperats com inundacions, riuades, onades de calor o altres aspectes propis de la natura (erupcions de volcans, tsunamis...), però també de caràcter o origen humà com els conflictes bèl·lics.

En definitiva, analitzant el passat i la història d'una sèrie, i observant-ne l'evolució, tot tenint en compte aquests factors, l'anàlisi univariant mira d'extreure conclusions pel que fa a la dinàmica i comportament de la variable estudiada. En aquest sentit, una de les principals eines que cal fer servir, sempre en primera instància, és l'anàlisi gràfica, a partir de dues coordenades, on a l'eix de les abscisses se situa el temps, i a les ordenades, els valors que va prenent la variable.

Exemple

A continuació es reprodueix un exemple de sèrie temporal, amb dades trimestrals (quatre observacions per cada any, per tant), des de l'any 2000 fins a l'any 2005 (en total, 24 observacions). Es tracta de la sèrie que recull el nombre de visitants d'un parc natural d'alta muntanya.



Observant el gràfic, es pot concloure que es tracta d'una sèrie amb tendència creixent i un marcat component estacional: els quarts trimestres de cada any mostren la dada més baixa: a partir del primer trimestre, la dada repunta, probablement per la temporada d'esquí d'estacions properes al parc; segueix creixent, tot i que amb menys força, al segon trimestre (Setmana Santa), i arriba al punt més àlgid a l'estiu (període vacacional escolar).

Pel que fa al tractament de les dades, com ja s'ha comentat les tècniques que es poden aplicar són diverses, però no les exposarem ja que calen desenvolupaments estadístics i economètrics abans de la seva aplicació. Però sí que introduirem, a mode d'exemple, una tècnica ja coneguda. Efectivament, la recta de regressió, exposada anteriorment, és una de les eines, simples i amb molts biaixos –per exemple, cal destacar que no recull les oscil·lacions de tipus esta-

cional–, però que en tot cas s'utilitza per tal d'aproximar la tendència a llarg termini de la sèrie. La formulació és la mateixa que la que s'ha explicat en l'apartat corresponent, amb la diferència que, si bé en aquell cas es parlava de treballar amb dues variables (X i Y), ara som dins de l'anàlisi univariant i per tant tenim una sola variable: Y. Així doncs, per a l'aplicació de les fórmules, el que es fa es confrontar la variable que s'ha d'explicar (Y) en funció del temps (t). Per tant n'hi ha prou de substituir en la formulació explicitada a l'apartat corresponent la X, ara inexistent, per «t». Val a dir que per a dur a terme els càlculs i obtenir l'estimació de recta de tendència, t no s'expressarà amb el nom o etiqueta que correspon a l'any o mes o trimestre, o període considerat, sinó que t serà construïda seqüencialment a partir de 1. Per tant t, a aquests efectes, sempre prendrà valors 1, 2, 3... T.

Exemple

Disposem de les dades fetes servir per al gràfic anterior; la següent taula recull aquesta informació. A més, s'ha introduït la columna t per tal de poder fer els càlculs esmentats.

Període	Y(t)	t
IT00	42	1
IIT00	60	2
IIIT00	81	3
IVT00	27	4
IT01	56	5
IIT01	65	6
IIIT01	89	7
IVT01	30	8
IT02	59	9
IIT02	75	10
IIIT02	112	11
IVT02	40	12
IT03	65	13
IIT03	83	14
IIIT03	120	15
IVT03	63	16
IT04	72	17
IIT04	89	18
IIIT04	115	19
IVT04	54	20

Període	Y(t)	t
IT05	80	21
IIT05	95	22
IIIT05	127	23
IVT05	63	24

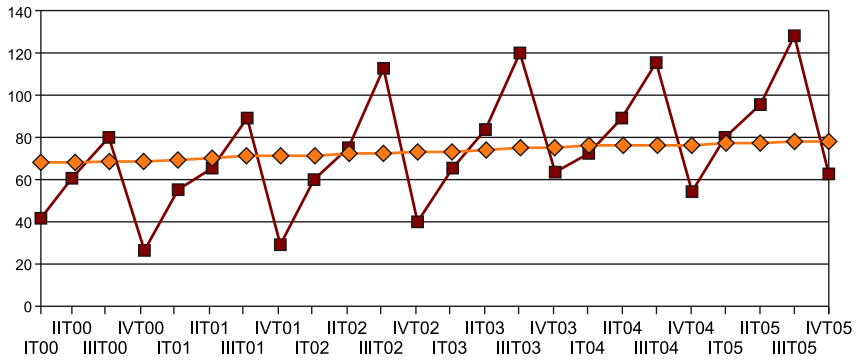
Aplicant el model de regressió lineal per MQO, acabem obtenint que:

b = 0,45; a = 67,8.

Per tant, l'estimació del pendent ens quedaria expressat com a:

Y = 67,8 + 0,45 x t

Finalment, aplicant-ho per als valors de t= 1... fins a 24, ens quedaria la següent representació gràfica.



5.4. El Compte Satèl·lit del Turisme

A causa de la seva transversalitat, la quantificació de l'impacte econòmic del turisme és una tasca estadísticament molt complexa. És per aquest motiu que a instàncies de l'UNWTO es va posar en marxa l'elaboració d'una metodologia específica que assolís aquest objectiu, i que rep el nom de Compte Satèl·lit del Turisme. Sense aprofundir en aspectes tècnics massa complexos, a continuació s'exposen alguns dels aspectes fonamentals d'aquesta metodologia.

La comptabilitat satèl·lit es basa en dues premisses:

1) Qualsevol sector econòmic pot ser proveïdor de turisme, en funció de la característica de visitant del consumidor.

2) Els sectors econòmics estan interrelacionats entre si, a través de les relacions entre client i proveïdor (els anomenats efectes intersectorials, mesurats amb una eina específica coneguda com a taula input-output).

En aquest sentit, cal distingir tres tipus d'impactes econòmics:

- **Efecte directe**: es tracta de l'impacte immediat que es genera arran d'una despesa. Per exemple, quan un turista contracta una nit d'hotel, aquesta despesa seria part de l'efecte directe.

- **Efecte indirecte**: es tracta de l'impacte econòmic que es produeix arran de les relacions intersectorials. Per exemple, un hoteler, per tal de poder oferir una nit d'hotel, abans haurà efectuat despeses en mobiliari, productes tèxtils, subministraments, etc. Aquestes despeses s'englobarien dins l'efecte indirecte del turisme.

- **Efecte induït**: es tracta de l'impacte econòmic que té lloc gràcies a les rendes generades per l'activitat. Seguint amb el mateix exemple, l'hoteler hauria contractat uns treballadors, els quals percebrien un sou, que al seu torn destinarien als seus propis consums de habitatge, subministraments, alimentació, etc. Aquest seria l'efecte induït.

Així doncs, la base d'aquesta tècnica es centra en el següent:

1) Analitzar quina part dels ingressos de cadascuna de les branques d'activitat d'una economia és generada pel turisme.

2) A través de l'estudi de les relacions intersectorials, mesurades per les taules input-output (taules que analitzen les relacions entre sectors, a partir de les compres i vendes que es produeixen entre ells, per explicar-ho d'una manera senzilla), saber quins efectes globals té l'activitat turística sobre l'economia d'una destinació.

El Compte Satèl·lit del Turisme és per tant:

- Un sistema d'informació macroeconòmic que permet conèixer i dimensionar la contribució de certes activitats no tradicionals en una economia, tot identificant les activitats productives que hi contribueixen i hi participen, així com les relacions amb la resta d'activitats i teixit productiu (input/output).

- Un conjunt de comptes i taules, basats en els principis de la comptabilitat nacional i l'equilibri general de l'economia (oferta i demanda), així com les relacions existents en el sistema productiu (I/O).

- Una eina necessària per a activitats no tradicionals, que disposa d'una metodologia pròpia i que permet detallar específicament les activitats o productes objecte del seu interès, incorporant informació no monetària i modificant de manera instrumental, i amb una finalitat funcional, la metodologia bàsica, sempre, però, salvaguardant les classificacions i definicions del sistema.

- La metodologia que ens informa sobre aspectes com ara consum, formació bruta de capital o inversió, comptes de producció i estructura de les empreses (oferta), pes i paper del sector públic, exportacions i importacions, llocs de treball generats, VAB, PIB, impacte fiscal, i per descomptat, en el cas del turisme, impacte de la demanda desagregat per sectors.

5.5. Les estadístiques turístiques com a font d'informació

A banda de la investigació que un mateix pot dur a terme, i dins de la primera fase de cerca documental, també cal tenir present que existeixen fonts secundàries d'informació (les fonts primàries són aquelles elaborades per l'investigador i les secundàries, les ja existents).

En el món actual l'elaboració d'estadístiques és una pràctica, no tan sols habitual, sinó fomentada i promoguda per organismes internacionals, com ara l'Organització per a les Nacions Unides. De fet, aquest organisme té una oficina específica que es dedica a això exclusivament: la UNSTATS. La divisió d'estadístiques de l'ONU té com a objectiu fonamental avançar cap a un sistema estadístic global, i entre altres tasques, s'encarrega d'establir, consensuar i publicar normes i estàndards comuns, donar suport per a la consolidació de sistemes estadístics nacionals que permetin l'agregació i comparabilitat de les dades i difondre la informació disponible.

El primer aspecte que cal tenir en compte quan volem analitzar l'activitat turística des d'un punt de vista estadístic és la definició clara dels objectius que perseguim, com s'ha comentat més extensament als apartats inicials. Aquesta qüestió, que en moltes altres disciplines resulta relativament fàcil, no és tan directa en el cas del turisme. Això és degut a la complexitat i transversalitat del fenomen, però cal dir que amb el pas de les dècades i el creixement imparable de l'activitat, s'ha dut a terme un important esforç per tal d'unificar criteris i definicions. En aquest apartat, doncs, el que et proposem és que coneguis els documents de referència així com les fonts estadístiques més rellevants.

El document de referència pel que fa a les definicions del turisme és de l'Organització Mundial del Turisme (antiga OMT, o WTO en les sigles en anglès, i actualment UNWTO en adscriure's com a organisme específic de Nacions Unides), posteriorment ratificat i subscrit per Nacions Unides i l'OCDE. En la publicació de l'any 2008 «Recomanacions internacionals per a les estadístiques del turisme», que actualitza el document de 1993, es recullen criteris fonamentals per tal d'endegar una recerca turística fonamentada en la definició correcta i comparable de les variables relatives a la temàtica.

Pel que fa a les activitats turístiques, en puritat es podria afirmar que no existeixen en si mateixes sinó que qualsevol activitat pot esdevenir turística mentre el consumidor compleixi els requisits de ser un visitant. Tanmateix, a les recomanacions de 2008 es defineixen i classifiquen les activitats considerades

«característiques» del turisme, a més dels productes característics. Val a dir que, per tal de seguir les normes de comparabilitat internacional, les classificacions segueixen, respectivament, les pautes de la Classificació industrial internacional unificada de les activitats econòmiques (ISIC) i les de la Classificació central de productes (CPC).

En tot cas, l'UNWTO és sens dubte una de les principals fonts d'informació estadística del turisme per a tot el món, tot i que l'organisme no tan sols ofereix dades i magnituds sinó que també s'encarrega d'elaborar i publicar recomanacions per a una millor i més comparable recollida d'informació.

La United Nations World Tourism Organization té un departament i un programa específic d'estadístiques, encarregats de generar dades i també publicacions amb l'objectiu de millorar l'estadística del turisme de tot el món.

«The United Nations recognizes the World Tourism Organization as the appropriate organization to collect, to analyse, to publish, to standardize and to improve the statistics of tourism, and to promote the integration of these statistics within the sphere of the United Nations system.»

Podríem dir que es tracta de l'estadística «oficial» del turisme a escala mundial, ja que compta amb el segell de l'ONU. L'UNWTO publica periòdicament les grans xifres del turisme mundial, presentades per grans regions, i amb un detall que arriba fins a més de 220 estats. A la web de l'organisme es poden consultar, entre altres temes:

- les notes de premsa
- el baròmetre de conjuntura
- el document de destacats
- monogràfics
- anuaris estadístics
- documents metodològics

Finalment, l'UNWTO treballa en l'elaboració d'un sistema d'estadístiques del turisme homogeneïtzat, per al qual ofereix suport i assistència tècnica als estats i membres que s'hi vulguin adherir.

De tota manera, no és l'únic organisme que publica dades sobre turisme a escala global; també n'hi ha altres que se n'ocupen.

L'OCDE (Organització per a la Cooperació i el Desenvolupament Econòmic) també té una unitat dedicada al turisme i ofereix estadístiques i estudis; en la seva pàgina web es poden consultar els diferents informes, enquestes i publicacions que ofereix, amb dades relatives a diversos països membres.

L'organisme també publica monografies sobre diferents temes d'interès, a més de tenir el seu propi comitè d'experts per a la conciliació de metodologies, inclosa la relativa als comptes satèl·lit, metodologia que per la seva rellevància i especificitat, més endavant explicarem.

Per la seva banda, l'EUROSTAT és l'organisme de la Comissió Europea encarregat de les estadístiques de la Unió Europea, i no cal dir que també té un apartat dedicat al turisme. Treballant en la mateixa direcció que els altres dos organismes, l'EUROSTAT també ofereix directrius per a la consecució d'una estadística comparable i homogènia per part dels estats membres. Pel que fa a la informació disponible, l'organisme fa anys que treballa per a oferir informació a escala regional.

També té una destacada importància l'establiment i difusió de la metodologia per al càlcul de l'impacte econòmic del turisme a través del Compte Satèl·lit, que també forma part de les seves línies de treball.

En el cas d'Espanya, al marge dels observatoris, instituts i altres entitats d'àmbit autonòmic o local, l'Estat compta amb dos instruments per a aquest fi: l'INE (Instituto Nacional de Estadística) i l'IET (Instituto de Estudios Turísticos), tot i que actualment s'està treballant per a la unificació en l'organisme proveïdor de dades.

L'INE és l'organisme encarregat de l'estadística oficial a l'Estat espanyol i treballa seguint les directrius internacionals, en concret, el Sistema Estadístic Europeu. Abasta, per tant, tots els àmbits i sectors d'activitat. Tanmateix, per les peculiaritats del turisme i per la seva transversalitat, l'INE només s'encarrega –o tradicionalment s'encarregava, ja que en el moment actual s'estan fusionant operacions que eren pròpies de l'IET– de l'anàlisi de l'activitat des d'un punt de vista de l'oferta. Les estadístiques més conegudes d'aquest organisme pel que fa a turisme són les enquestes mensuals d'ocupació.

D'altra banda, l'INE és també l'encarregat de l'elaboració del Compte Satèl·lit del Turisme a Espanya.

Pel que fa a les estadístiques de demanda, se n'encarrega l'IET. Encara que aquest organisme també fa altres operacions (com les relatives a visitants en determinats centres, companyies de baix cost o ocupació), sens dubte les tres grans estadístiques que proporciona l'IET són:

- Moviments turístics en frontera (FRONTUR), que recull l'entrada de viatgers internacionals.

- Moviments turístics dels espanyols (FAMILITUR), que estudia els viatges dels residents a l'Estat.

- Enquesta de despesa turística (EGATUR), que aprofundeix en l'anàlisi d'aquesta variable.

Bibliografia

Alegre, J.; Juaneda, C.; Cladera, M. (2003). *Análisis cuantitativo de la actividad turística*. Madrid: Ed. Pirámide.

Coenders Gallart, G.; Renart Vicens, G.; Vall-Llosera Casanovas, L.; Xabadia Palmada, À. (2009). *Tècniques d'anàlisi turística*. Girona: Documenta Universitaria.

Corral, J. A.; Cànoves, G. (2013). «La investigación turística publicada en revistas turísticas y no turísticas: análisis bibliométrico de la producción de las universidades catalanas». *Cuadernos de Turismo*. Núm. 31, pàg. 55-81. Universidad de Murcia.

Dwyer, L; Gill, A.; Seetaram, N. (eds) (2012). *Handbook of Research Methods in Tourism. Quantitative and Qualitative Approaches*. Northamton: Edwar Elgar.

García, M.; Calle, M. de la (2004). «La investigación geográfica del turismo en España». *Anales de Geografía*. Núm. 24, pàg. 257-277. Universidad Complutense de Madrid.

Instituto de Estudios Turísticos (1996). «La investigación turística en España». *Revista de Estudios Turísticos*. Núm. 129. Madrid: Instituto de Estudios Turísticos.

López Bonilla, J. M.; López Bonilla, L. M. (2015). *Manual de investigación de mercados turísticos*. Madrid: Ed. Pirámide.

Martín Pliego, F. J. (2004). *Introducción a la estadística económica y empresarial. Teoría y práctica* (3ed.). Madrid: Ed. Paraninfo.

McKercher, B.; Bruce Prideaux (2014). «Academic myths of tourism». *Annals of Tourism Research*. Núm. 46, pàg. 16-28. Elsevier Ltd.

Naciones Unidas (2010). *Cuenta satélite de turismo: Recomendaciones sobre el marco conceptual* (CST: RMC 2008). Luxemburg/Madrid/Nova York/París: Nacions Unides.

Naciones Unidas (2010). *International Recommendations for Tourism Statistics 2008* (IRTS 2008). Madrid/Nova York: Nacions Unides.

Organización Mundial del Turismo (2001). *«Apuntes de metodología de la investigación en turismo»*. Madrid: Organización Mundial del Turismo.

Pulido, J.I. *et al* (2006). «¿Está la investigación en turismo suficientemente reconocida y valorada en España?» *Revista de Análisis Turístico*. Núm 2. Madrid: AECIT.

Pulido, J.I. *et al* (2006). «Validez de las fuentes de información del turismo español». *Revista de análisis turístico*. Núm. 1. Madrid: AECIT.