Estadística aplicada

Estadística descriptiva

Entenem per estadística descriptiva la part de l'estadística que pretén presentar, organitzar i senyalar les diferents dades que es poden obtenir de l'estudi d'una característica de la població. Aquesta part de l'estadística no vol anar més enllà de la presentació dels resultats d'un treball de camp. No vol inferir característiques de la població. En aquells treballs que comportin aquesta feina s'ha de demanar un cert rigor en els càlculs i en la seva presentació, i un rigor encara més afinat en les representacions gràfiques, ja que la seves característiques intuïtives les fan especialment sensibles a una interpretació ràpida i convé evitar que aquesta sigi inexacta.

Estadística descriptiva. Càlculs

Una taula estadística resulta fonamental. La manera de fer-la dependrà de les variables que es volen estudiar. En variables discretes la notació de les diferents dades x_i i les freqüències f_i absolutes és imprescindible. Si aquestes són variables contínues caldrà determinat intervals de manera inequívoca (fent atenció a si els extrems de cada interval s'inclouen o no dins seu), a la marca de l'interval (que es fa servir en tots els càlculs posteriors, i a les seves freqüències.

Una possible notació d'una taula de càlcul elemental en variables discretes pot ser

X_i	F_i	x_if_i	x_i^2f

En una de variables contínues cal determinar el tipus d'interval i la marca de l'interval

$[x_a,x_b)$	X_m	F_i	x_mf_i	$X_m^2f_i$

En aquest cas l'interval és tancat a esquerra i obert a dreta, x_a s'inclou en l'interval i x_b no s'inclou. La marca de l'interval s'indica per x_m i habitualment és el valor central de l'interval.

En el cas del primer o últim dels intervals considerats s'ha de determinar la marca corresponent. Per exemple si la variable són els diners que porta a la butxaca un alumne, l'interval és més de 1000 ptes., s'ha de determinar quina serà la marca de l'interval que representem $(1000,\infty)$.

Els paràmetres habituals, com ara la mitjana o els percentatges resulten imprescindibles. Cal insistir en la importància de valors com ara la mediana i el mode, i que aquests valors s'interpretin. Molts d'ells venen donats amb el càlcul previ de les freqüències relatives i les relatives acumulades que no han de mancar.

Les freqüències relatives poden calcular-se en tant per 1 o en tant per 100. No és important establir un o altre com maneres de càlcul preferides. De totes maneres un tant per 1 sembla més proper a una futura interpretació de les probabilitats dels esdeveniments.

Convé que el nombre de decimals que figurin en els càlculs sigui la convenient en cada cas. A més a més cal vigilar que la quantitat de decimals sigui la mateixa en tots els càlculs d'un exercici o d'un treball . No sembla convenient que hi hagi més de quatre xifres decimals. Cal arrodonir la darrera. Només pot tenir un cert sentit l'augment de la quantitat de decimals quan aquests impliquin càlculs de probabilitat, de nivells de confiança, de coeficients de correlació,..

Menció especial mereixen el càlcul de les mesures de dispersió. Sovint aquest conjunt de paràmetres no es tenen en compte a l'hora d'estudiar la distribució d'unes variables estadístiques i són fonamentals. En l'estudi d'una variable és important el càlcul del valor central, la mitjana, però és també fonamental saber si les dades estan o no agrupades entre elles i aquesta informació només ens la donen els paràmetres de dispersió. D'ells els més coneguts són el recorregut i la desviació típica, però no són els únics.

El valor d'aquest paràmetre s'ha d'interpretar acuradament. La majoria dels càlculs d'inferència estadística es basen en ells.

Càlcul de valors centrals

El valor central fonamental és la mitjana, es defineix com la suma de tots els valors que assoleix la variable entre la quantitat de valors. S'acostuma representar amb $\bar{x}$ i es calcula

$$\bar{x} = \frac{\sum x_i}{N} = \frac{\sum x_i f_i}{N} = \frac{\sum m_i f_i}{N}$$

depenent de la manera de agrupar les dades. La primera de les expressions correspon a un conjunt de dades, una senzilla llista de valors. La segona pressuposa que hem calculat les freqüències de cada una de les dades que hi intervenen, i en la tercera hem format intervals i m_i és la marca de cada un.

La mediana és el valor que ocupa la posició central de la distribució. Si ordenem els valors serà aquell que en té tants abans com després. Correspon al valor d'una freqüència acumulada relativa de 0,5

La mediana coincideix amb un dels valors de la distribució si la quantitat de dades és imparella. Si la quantitat de dades és parella la mediana està entre dues d'aquestes. Si aquestes dues coincideixen es pren aquesta dada com mediana, si són diferents és pren la mitjana entre elles.

Com exemples

dades	ordenació	mediana
{4,7,3,5,2}	{2,3,4,5,7}	4
{1,2,3,5,1,3,2,3}	{1,1,2,2,3,3,3,5}	2,5 (quantitat parella de valors, els dos centrals diferents)
{1,2,3,5,1,3,2,2}	{1,1,2,2,2,3,3,5}	2 (quantitat parella de valors, els dos centrals iguals)

El mode és el valor de la distribució de màxima freqüència. És aquell valor que més es repeteix. Pot no ser únic; si les dades són {1,2,2,3,3} hi ha dos modes: 2 i 3, que tenen freqüència 2

Mesures de dispersió

La mesura de dispersió més simple és l'amplitud o recorregut de la distribució; és la diferència entre els valors superiors i inferiors de la mateixa. Respon a la pregunta de com és, d'extens, el conjunt dels possibles valors de la variable aleatòria. Si estudiem les alçades d'un grup d'alumnes i ens trobem amb una mínima alçada de 1,55 m i una màxima alçada de 1,90; el recorregut de la distribució d'alçades que estem estudiant és de 1,90-1,55=0,35 m

La mesura de dispersió més emprada és la variància que es defineix com la mitjana dels quadrats de la diferència dels valors respecte de la mitjana

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{N}$$

L'arrel quadrada de la variància és la desviació tipus o desviació estàndard, es representa habitualment amb la lletra grega sigma σ

Podem calcular la variància o la desviació tipus a partir de la definició, però és molt més pràctic distribuir les dades de la següent manera

x	**f**	**x.f**	**x^2**	**x^2.f**
3	1	3	9	9
4	2	8	16	32
5	4	20	25	100
6	6	36	36	216
7	5	35	49	245
8	5	40	64	320

9	2	18	81	162
	25	160		1084

En efecte es demostra que la variancia pot calcular-se formant la mitjana de la suma dels quadrats de les dades menys el quadrat de la mitjana segons:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{N} = \frac{\sum x_i^2 . f}{N} - \bar{x}^2$$

En l'exemple serà

$$s^2 = \frac{1084}{25} - \left(\frac{160}{25}\right)^2 = 2{,}4$$

S'ha de diferenciar de la desviació tipus corregida (o desviació tipus de la població) que es representa de la forma σ_{n-1} ,diferenciat de σ_n. La majoria de les calculadores científiques tenen aquestes dues tecles.

La desviació tipus corregida és

$$\sigma_{n-1} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}$$

observem que la diferència és el denominador N-1 en lloc de N. Aquest valor aproxima millor la desviació tipus de la població a partir de les dades d'una mostra. Serà el càlcul que convindrà fer en el cas d'una tasca inferencial: quan intentem calcular valors d'una població a partir de dades d'una mostra d'aquesta població.

Altres mesures de dispersió habituals són totes aquelles que pretenen fixar uns valors que reparteixen els valors de la distribució al llarg de l'amplitud del conjunt de dades. Si recordem la mediana aquest és el valor que divideix els termes de la distribució en dos conjunts del mateix nombre de dades. Si dividim en quatre conjunts del mateix nombre de dades parlarem dels quartils de la distribució. Si en fem 10 dels decils, si 100 dels centils,..

Resulta evident que el segon quartil, el cinquè decil, el centil 50 i la mediana han de coincidir.

La manera més convenient de calcular aquests valors és a partir de les freqüències relatives acumulades. Una freqüència relativa acumulada de 0,25 indicarà el primer quartil; una freqüència relativa acumulada de 0,9 el novè decil,..

Les qualificacions numèriques habituals de les proves escolars pretenen coincidir estadísticament amb els decils de les distribucions de notes.

Es coneix amb el nom de amplitud interquartíl·lica la diferència entre el primer i el tercer quartil. Aquest interval ha de contenir el 50% dels elements de la distribució.

Estandarització de dades

Si bé no és exactament una mesura habitual de dispersió de dades, la variable tipificada d'una dada x en una distribució de mitjana $\bar{x}$ i desviació σ és defineix com

$$z = \frac{x - \bar{x}}{\sigma}$$

i té una gran importància en les aproximacions de models estadístics, com ara la distribució normal. Es considera la mesura de la desviació respecte de la mitjana de cada una de les dades, però prenent com unitat de mesura la desviació tipus. Així, una z=1,5 indica una dada que està una desviació i mitja per sobre del valors de la mitjana.

Puntuacions T

Els valors que es poden obtenir fent servir les puntuacions z acostumen a interpretar-se amb dificultats sobre tot pel fet de donar valors negatius quan són inferiors a la mitjana. La correcció de les puntuacions T definida per:

$$T = 10z + 50$$

les transforma en un valor central 50 amb una desviació típica de 10. Una T=50 correspon exactament a un valor igual a la mitjana de la distribució

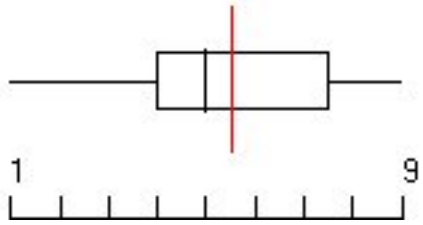
Teòricament les puntuacions T poden ser negatives. En la pràctica és molt difícil ja que haurien de desviar-se 5 desviacions típiques per sota de la mitjana i la probabilitat que això esdevingui és molt petita. De la mateixa manera una puntuació T de 100 o superior a 100 és també pràcticament impossible.

Comparació de mitjana i mediana

En situacions de distribucions estadístiques modèliques, la mitjana i la mediana coincideixen. Si no és així resulta significatiu observar que esperem de distribucions on la mitjana sigui superior a la mediana, o d'altres on aquesta situació s'inverteixi.

Observem que la mitjana és especialment sensible a valors molt grans o molt petits, i la mediana resulta poc influenciada si el valors extrems de la distribució són encara més extrems (majors, si són els superiors, menors si són els més petits). La cua de la distribució s'allargarà més cap al costat on hi hagi la mitjana de la distribució.

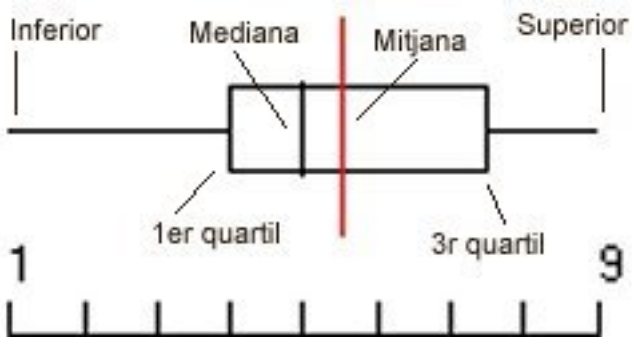
Els diagrames en caixa són molt convenients si es vol comparar els valors de mitjana i mediana en una distribució. És un gràfic com ara



En un diagrama de caixa es representa sobre una escala un rectangle d'extrems el primer i el tercer quartil, allargat per dos segments fins els valors superior i inferior de les dades de la distribució. La mediana divideix el rectangle en dues parts. S'acostuma representar també la mitjana com un segment diferenciat d'aquest esquema

Ens els diagrames de caixa es pot comparar la relació entre mitjana i mediana, la simetria de les dades i la possible acumulació en un o altre quartil. Les dades extremes són aquelles representades fora del rectangle, en els dos segments.

Pot fer-se un diagrama de caixa més elaborat si es prescindeix dels valors més extrems de la distribució. Aquest es consideren valors poc representatius.



Coeficient d'asimetria i curtosi

El coeficient d'asimetria es calcula

$$C_s = \frac{\sum z_i^3}{N} = \frac{\sum \left(\frac{x_i - \bar{x}}{\sigma} \right)^3}{N}$$

on z és el valor de la tipificació de cada dada. Un coeficient 0 indica una simetria perfecte, valors positius indiquen que en la distribució hi ha més valors superiors a la mitjana, serà el cas de distribucions esbiaixades a dreta. Valors negatius indiquen perfils amb cues a l'esquerra

El coeficient de curtosi K d'una distribució és calcula

$$K = \frac{\sum z_i^4}{N} - 3 = \frac{\sum \left(\frac{x_i - \bar{x}}{\sigma} \right)^4}{N} - 3$$

En una distribució normal la mitjana de les quartes potències de les dades tipificades és 3. Aleshores un valor de K=0 indica que la distribució té un grau d'apuntament semblant al de la distribució normal. Un valor de K>0 indica un apuntament superior (una llarga punxa en els valors centrals i unes cues molt planes, de freqüències baixes). Un valor de K<0 un menor apuntament: poca diferència entre les freqüències dels valors centrals i dels extrems, una gràfica de la distribució més plana.

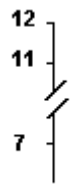
Estadística descriptiva. Gràfics

Una de les maneres més simples de representar les dades és fer un gràfic dels valors obtinguts. Es pot fer servir la manera habitual, ja sigui en diagrames de barres, de sectors, de línies, de punts,... Però es faci d'una o altra manera s'ha de tenir present evitar alguns aspectes que poden donar imatges falses de la realitat.

Diagrama de barres

Aquest tipus de representació gràfica consisteix en senyalar sobre l'eix OX els intervals de la distribució o els valors d'una distribució discreta, i en l'eix OY unes alçades que corresponen a les respectives freqüències de cada interval o de cada valor.

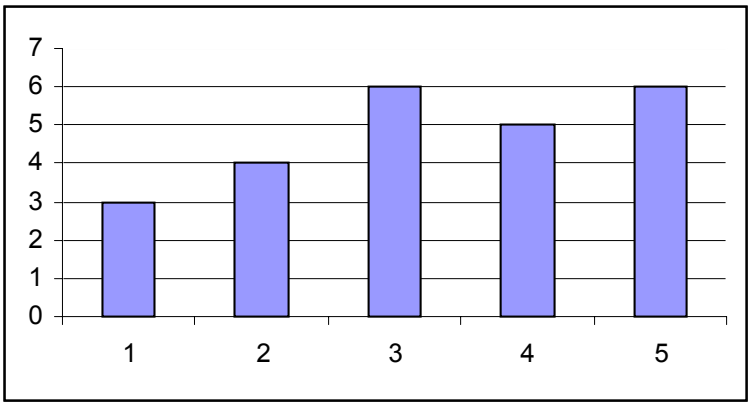
Les dades es representen fent servir rectangles (barres) de la mateixa base i altures proporcionals a les freqüències. Segons convingui els rectangles es poden orientar de manera que la base sigui proporcional a les freqüències i altures constants; aquest serà el cas que s'acostuma fer servir per representar piràmides de població.



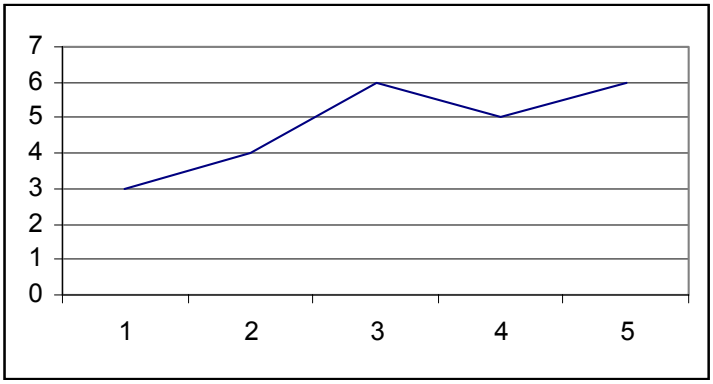
L'eix que és reserva per representar les freqüències, habitualment OY, no ha de canviar l'escala. Si en alguns valors convé canviar-la s'ha de fer constar explícitament. La manera habitual és representar l'eix de forma que es vegi el salt d'escala, per exemple fent servir dues línies paral·leles. En l'eix OX es representen habitualment les variables estadístiques. En aquest cas s'ha de diferenciar variables discretes de contínues, En el cas d'intervals no és aconsellable representar les freqüències de cada interval en forma de barres sense separació entre elles.

Si el nombre de barres és gran es pot representar com un diagrama de punts, en ell la variable x són cada un dels possibles valors i la variable y la freqüència

Un cas típic de diagrama en barres pot ser

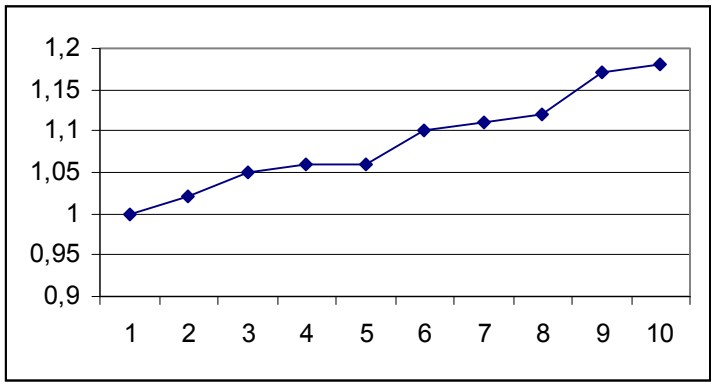


L'equivalent en diagrama de línies és



Escales

L'escala de representació de les freqüències o l'escala dels índexs, és molt sensible a manipulacions que donen imatges molt deformades de la realitat. Cal tenir molt present el valor de l'escala i interpretar les dades en conseqüència. Els dos gràfics que segueixen representen la mateixa evolució d'un determinat índex, però sobre escales diferents. El primer fent servir una escala que vol magnificar l'augment:



El segon canviant l'escala i fent que el comportament de la variable sigui molt més suau, es pot dir que pràcticament no canvia

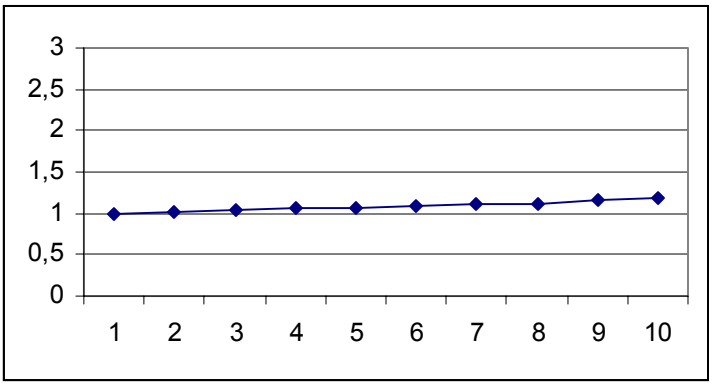
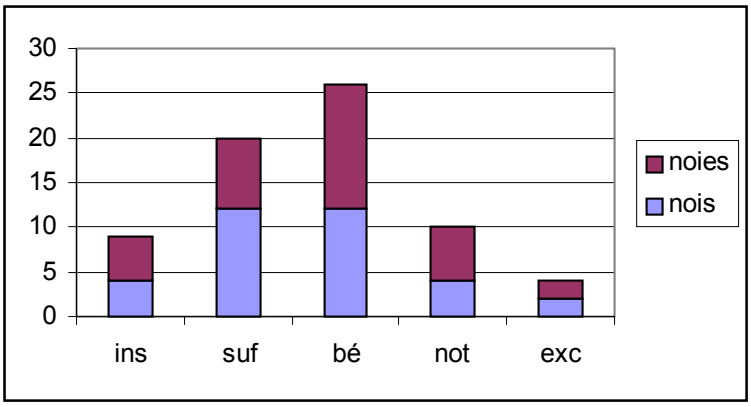


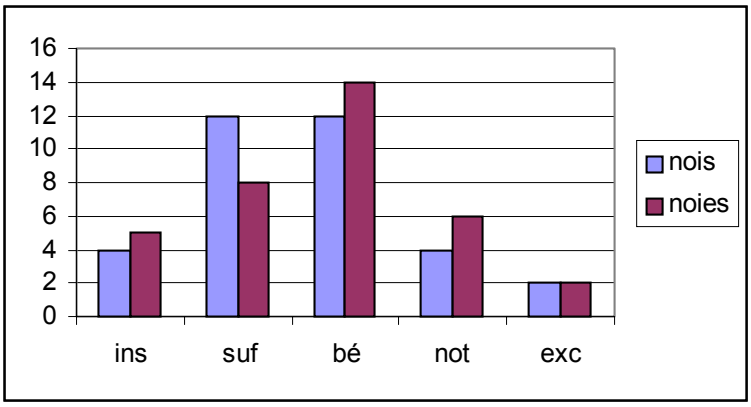
Diagrama de barres combinades

Si es vol diferenciar, en cada una de les dades, les freqüències que obté una de dues possibles categories, podem representar un doble gràfic de barres, cada un d'ells responsable d'una de les categories. Podem pensar, com exemple, en qualificacions d'una prova diferenciant nois i noies. El gràfic pot ser en barres combinades o formant barres apilades. La primera de les opcions presenta més clarament les diferències entre una i altra categoria. La segona opció presenta millor el total de cada dada (en aquest cas de cada una de les qualificacions) sense diferenciar tant bé com l'anterior les categories en cada una.

Barres apilades



Barres combinades



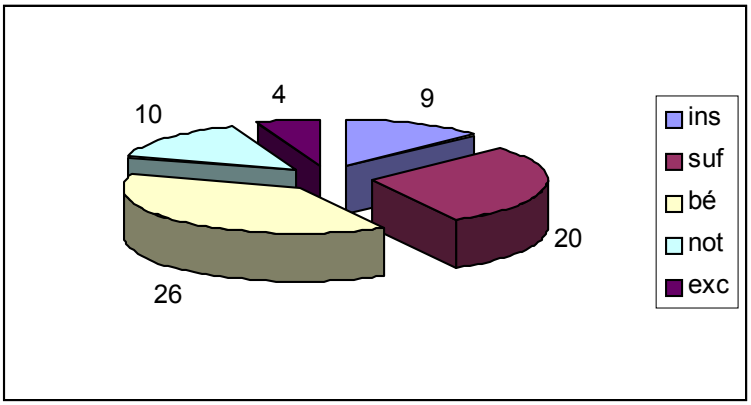
Existeixen altres opcions més o menys sofisticades de diagrames de barres. Una, independent de les freqüències de cada valor, és el que presenta només percentatges del total.

En aquest gràfic perdem informació. No tenim la freqüència de cada una de les dades: no sabem els alumnes que han obtingut una qualificació d'insuficient, per exemple.

Diagrames de sectors

Permeten veure clarament de quina manera les dades d'una variable es distribueixen entre els elements d'una població. Són especialment útils per mostrar repartiments i estem acostumats a veure'ls representant percentatges de vots o d'escons en eleccions.

Es poden representar sobre cercles o semicercles. Es poden separar o no els sectors per veure millor les



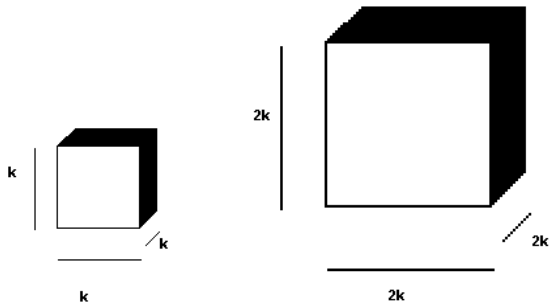
diferències o marcar un sector determinat, poden o no tenir un efecte de profunditat. Com exemple les dades anteriors sobre el total dels alumnes es representen

Pictogrames

D'una manera molt general el nom de pictogrames es reserva a gràfics que mostren un dibuix, una imatge, un objecte,... que canvia les dimensions segons les freqüències. Podem representar l'augment del consum de llet en una població en diferents anys dibuixant un got de llet més gran, o l'augment del parc automobilístic dibuixant cotxes cada vegada més grans. Podem representar la disminució de les vendes d'un determinat article de consum dibuixant aquest article cada vegada més petit.

Un error freqüent en la representació de dades estadístiques en forma de pictogrames és no considerar que en imatges que representin una superfície o un volum les dimensions lineals de les mateixes no són aquelles que s'han de considerar per representar proporcionalment les freqüències, sinó que aquestes han de ser proporcionals a la superfície o el volum que es vol representar.

D'una manera més precisa un augment en un factor lineal k fa augmentar en un factor k^2 la superfície i en un factor k^3 el volum. Quadrats d'aresta doble tenen superfície quatre vegades més gran. Esferes de radi triple tenen un volum 27 vegades més gran



Si un quadrat té un costat doble que un altre, la seva superfície serà quatre vegades la del més petit (i no dues) i si es forma un cub amb aresta doble que un cub més petit, el seu volum serà vuit vegades (i no dues) més gran que el volum del cub precedent.

Segons això, si volem representar en un cub una freqüència doble, s'haurà de dibuixar un cub d'aresta k vegades l'aresta inicial, on $k = \sqrt[3]{2}$. Si es vol representar un quadrat de superfície doble de la d'un quadrat inicial, l'aresta del quadrat gran ha de ser $\sqrt{2}$ vegades l'aresta inicial.

La distribució binomial

Correspon a un determinat esdeveniment, amb probabilitat d'èxit p, que considerem es repeteix de manera independent n vegades. Ens preguntem la probabilitat que existeix que en aquestes n repeticions l'esdeveniment tingui èxit k vegades. Evidentment $0 \leq k \leq n$.

Primer hem de veure que les k vegades que demanem que l'esdeveniment tingui lloc en les n repeticions poden estar en $\binom{n}{k}$ posicions diferents de la llista de n repeticions. D'aquesta manera s'originen els nombres combinatoris que es consideren en una distribució binomial. Si ara considerem només l'esdeveniment de probabilitat p, que ha de tenir èxit k vegades, la probabilitat és p^k. I si considerem que l'esdeveniment de probabilitat complementària (1-p) ha de succeir n-k vegades, la seva probabilitat serà $(1-p)^{n-k}$

En resum, donades n repeticions independents d'una experiència aleatòria que pot tenir èxit en cada repetició amb una probabilitat p, la probabilitat que tingui èxit en k de les n repeticions és

$$P(x=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

Per alguns valors de n, p i k els valors d'aquestes probabilitat estan calculades. En altres casos es poden determinar fàcilment.

Exemples

Considerem el sexe en el naixement una variable aleatòria independent amb probabilitat 0,5 de ser nen i 0,5 de ser nena. Una família de 5 fills tindrà exactament 2 nens i 3 nenes amb probabilitat determinada per una binomial de paràmetres n=5, p=0,5, q=0,5 i k=2 (si considerem "èxit" els nens) o k=3 (si considerem "èxit" les nenes) . En qualsevol cas la probabilitat serà

$$P(k=2)=\binom{5}{2}0{,}5^2 0{,}5^3 = P(k=3)=\binom{5}{3}0{,}5^3 0{,}5^2 = 10\cdot 0{,}5^5 = 0{,}3125$$

En aquest exemple hem de tenir present que els combinatoris $\binom{5}{2}$ i $\binom{5}{3}$ són iguals.

Sabem que una màquina fabrica el 5% de les peces amb algun tipus de defecte. Si aquestes es venen en capces de 6 peces la probabilitat que en una capsa hi hagi més d'una peça amb defectes és

$$P(x>1)=\sum_{x=2}^{6}\binom{6}{x}0{,}05^x\cdot 0{,}95^{6-x}=\binom{6}{1}0{,}05^1\cdot 0{,}95^5+\binom{6}{2}0{,}05^2\cdot 0{,}95^3+\ldots+\binom{6}{6}0{,}05^0\cdot 0{,}95^6$$

En la mateixa situació anterior, la probabilitat que en una capsa totes les peces siguin bones serà

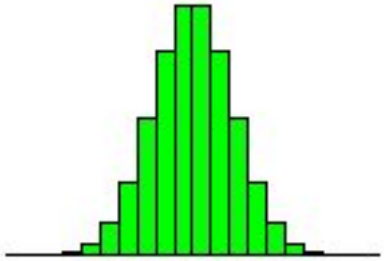
$$\binom{6}{6}0{,}05^0\cdot 0{,}95^6 = 0{,}7351$$

En la mateixa situació, el 26,5% de les capces fabricades d'aquesta manera tindran peces amb algun tipus de defecte. Observem que 1-0,7351=0,2649

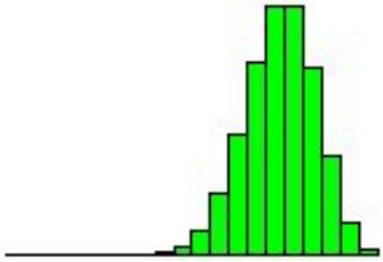
En un llançament d'un dau, la probabilitat d'obtenir un resultat 3 és de 1/6. La probabilitat de no obtenir cap 3 en el llançament de cinc daus és una binomial on n=5, k=0 i p=1/6. Podem calcular-ho

$$P(k=0)=\binom{5}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^5=\left(\frac{5}{6}\right)^5=\frac{3125}{7776}=0{,}4019$$

La distribució binomial serà simètrica només quan p=1-p=0,5, Observen el gràfic de probabilitats per a n=10 p=0,5 i q=0,5



Però si canviem p a 0,6 la simetria es perd



La distribució de Poisson

Sigui un experiment que es repeteix una quantitat gran de vegades i considerem el nombre d'esdeveniment en un interval de longitud t. La distribució de probabilitat discreta P(x) serà

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

i correspon a la probabilitat de x esdeveniments en aquest interval. El paràmetre λ correspon a la mitjana de les repeticions per unitat de longitud. Els valors d'aquesta distribució estan calculats en taules. (Veure taules)

La distribució de Poisson ajusta bé esdeveniments com ara la taxa de bacteris per volum, la quantitat de cotxes que circulen per una carretera, la quantitat de plaquetes en una mostra de sang,.. Cada una d'aquestes variables aleatòries es distribueix al llarg del temps o de l'espai, de manera que la probabilitat d'un d'aquests esdeveniment simples en un interval (de temps o d'espai) és proporcional a l'amplitud d'aquest interval

En general la distribució de Poisson de paràmetre λ s'ha de considerar com aproximació d'una binomial quan la probabilitat p d'èxit en una prova individual sigui petita i el nombre de repeticions n de l'esdeveniment augmenti, de manera que np=λ pot considerar-se constant

Exemples

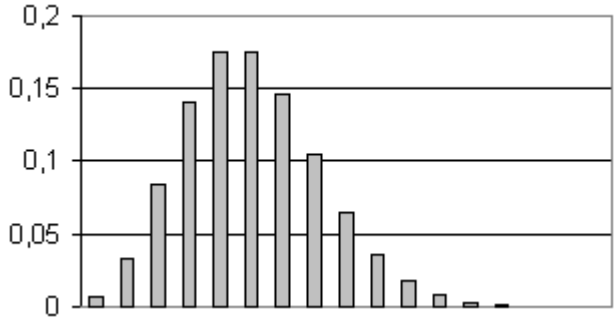
Si considerem que les trucades que es reben en una centraleta segueix una distribució de Poisson de 3 trucades per minut, la probabilitat que en un minut hi hagi exactament 4 trucades és

$$p(x=4) = \frac{3^4 e^{-3}}{4!} = 0{,}168$$

Si considerem que els cotxes que arriben en la cua d'una barrera de peatge cada minut segueix una distribució de Poisson de paràmetre 5 (en mitjana arriben 5 cotxes a la cua cada minut), les probabilitats que en un minut hi hagi 0,1,...4,.. són les que s'indiquen:

$$p(x=0) = \frac{5^0 e^{-5}}{0!}; \quad p(x=1) = \frac{5^1 e^{-5}}{1!}; \quad \ldots p(x=4) = \frac{5^4 e^{-5}}{4!}, \ldots$$

Quan els valors sobrepassen el valor de la mitjana, les probabilitats decreixen ràpidament. En l'exemple anterior el gràfic de les probabilitats des de 0 fins 13 és



En les hores d'oficina es reben una mitjana de 2,5 trucades de telèfon per minut. La probabilitat que en un minut es rebin 3 trucades és

$$p(x=3)=\frac{2{,}5^3 e^{-2{,}5}}{3!}=0{,}2138$$

Relacions de la distribució de Poisson amb la normal i la binomial

Quan n sigui gran i ni p ni la probabilitat complementària són properes a zero, una distribució binomial s'ajusta bé per una normal amb paràmetres $N(np;\sqrt{np(1-p)})$ Una distribució de Poisson s'ajusta bé a una normal de mitjana λ i desviació $\sqrt{\lambda}$. Una distribució binomial s'ajusta bé a una distribució de Poisson de $\lambda = np$ si p és propera a zero.

Les relacions que s'han mencionat es fan servir quan els càlculs en la distribució model són complicats. En especial quan aquests requereixen càlculs de diferents valors discrets, que en la distribució de Poisson o la binomila s'han de sumar per a cada valor de la variable discreta.

En general la distribució normal dóna molt bones aproximacions de la distribució binomial quan n és gran i p i q no són valors extrems. S'ha de establir una correcció de continuïtat en la normal transformada que acostuma ser suficient considerant els valors centrals de la distribució discreta. En una binomial de valors discrets k=0, 1, 2,... aproximada amb una normal corresponent, el valor de P[x>k] s'ha de transformar en P[x>k+0,5] ja que salta valors enters.

Exemple

En el llançament de 100 monedes perfectes, la probabilitat d'obtenir més de 55 cares correspon a la suma de

$$P(x>55)=\sum_{x=56}^{100}\binom{100}{x}0{,}5^x\cdot 0{,}5^{100-x}$$

Si calculem la suma de tots els termes (n'hi ha 45) el resultat és 0,13563

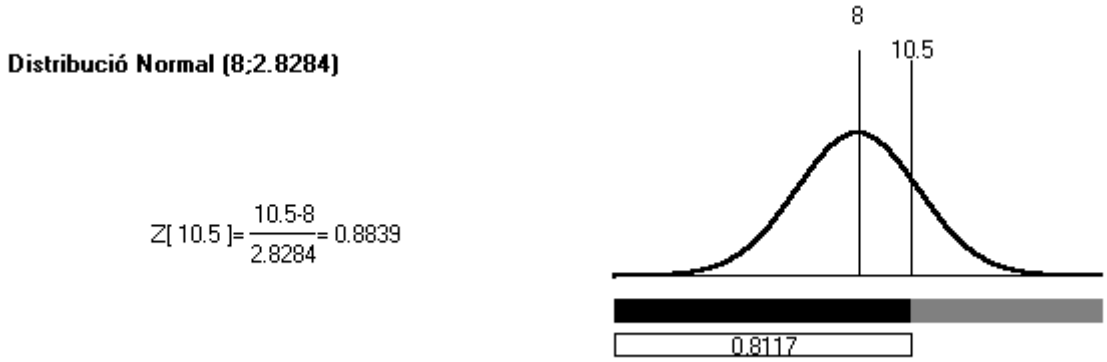
Intentem ara una aproximació fent servir una normal. La mitjana és np=100.0,5=50 i la desviació tipus és

$$\sigma=\sqrt{npq}=\sqrt{100.0.5.0{,}5}=5$$

Estem, aleshores, en una distribució N(50,5) que és una distribució contínua. Obtenir més de 55 cares és equivalent a sobrepassar l'enter 55 en la distribució normal. Considerem que obtenim més d'un valor 55 quan sobrepassem el valor mitjà entre aquest i el següent, en el nostre cas quan x>55,5

Veiem que en la binomial P[x>55] s'ha d'entendre P[x>55,5] en l'aproximació normal. La tipificació del valor 55,5 és z=1,1 i la probabilitat demanada de 1-0,86433=0,13567. Si ho comparem difereix molt poc amb el valor de la binomial.

Observem que la dificultat del càlcul en la binomial estava més en la suma de tots els termes discrets (resultat 56 cares, resultat 57 cares,...) que en el càlcul de cada un d'ells en particular.
Els clients que arriben a la cua d'una caixa d'un supermercat formen una distribució de Poisson de mitjana 8 clients cada minut. La probabilitat que en un minut arribin a la caixa més de 10 clients pot aproximar-se fent servir una distribució normal de paràmetres $(8, \sqrt{8})$. Amb la mateixa correcció de continuïtat de l'exemple anterior hem de calcular en aquesta normal P[x>10,5]



La probabilitat demanada és 1-0,8839=0,1161

Les probabilitat calculades fent servir una distribució de Poison d'aquest exemple donen unes probabilitats per a x=0,1,...,10 clients de:

x clients	P[x]
0	0,000335
1	0,002684
2	0,010735
3	0,028626
4	0,057252
5	0,091604
6	0,122138
7	0,139587
8	0,139587
9	0,124077
10	0,099262

La suma des de 0 fins 10 clients és 0,8158

Distribució hipergeomètrica

La distribució binomial es fonamenta en n proves independents on la probabilitat d'èxit és p, fixa en cada una de les repeticions. Sovint aquest model no es compleix, en particular si la mida de la població és petita comparada amb la quantitat de repeticions.

Si N és la mida de la població, on considerem una probabilitat p d'un determinat esdeveniment, i n la mida de la mostra que prenen en aquesta població, la probabilitat vindrà donada per

$$P(X=x)=\frac{\binom{Np}{x}\binom{N-Np}{n-x}}{\binom{N}{n}}$$

Pensem que Np són els elements de la població on es verifica la condició indicada amb probabilitat p, i n-x la resta dels elements de la població.

Quan n és només un petit percentatge de N, cal que N sigui petit per apreciar diferències entre els resultats d'una distribució binomial i una hipergeomètrica. Considerem, com exemple, que el 10% dels elements d'una població tenen una determinada característica. Si escollim 10 elements, la probabilitat que aquesta característica es verifiqui amb un màxim de 2 d'aquests 10 elements és

$$P(x \le 2) = \sum_{0}^{2} \binom{10}{x} p^x (1-p)^{10-x} = 0{,}93$$

si considerem una distribució binomial, i

$$P(x \le 2) = \sum_{0}^{2} \frac{\binom{10}{x}\binom{90}{10-x}}{\binom{100}{10}} = 0{,}94$$

si considerem una distribució hipergeomètrica.

La mitjana d'una distribució hipergeomètrica és np, però la variancia és

$$\sigma^2 = npq\frac{N-n}{N-1}$$

Com que la fracció $\frac{N-n}{N-1} < 1$, en la distribució hipergeomètrica les variables tenen una variancia menor que la corresponent binomial.

La distribució hipergeomètrica té una importància bàsica en la correcció de les probabilitats de mostres sense reemplaçament. Imaginem una població de mida N on P elements presenten una determinada característica. El primer element seleccionat té probabilitat $\frac{P}{N}$ de manifestar aquesta característica. Si escollim un segon element a l'atzar de la resta d'elements de la població, (sense reemplaçar el primer) la probabilitat que els dos presentin aquesta característica serà

$$\frac{P}{N} \cdot \frac{P-1}{N-1}$$

diferent del valor d'una binomial, on imaginem que el primer element pot ser seleccionat de nou.

Les distribucions hipergeomètriques s'utilitzen en el control de qualitat i les tècniques d'acceptació o rebuig segons els resultats de mostres.

Exemple

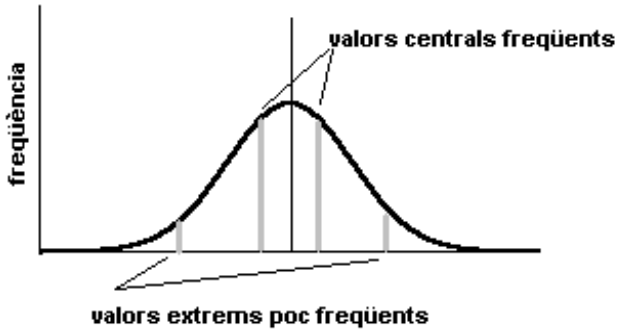
Una lot està format per 40 peces de les quals 2 tenen algun tipus de defecte. Decidim acceptar tot el lot si analitzem vuit peces i cap d'elles té defectes. En aquestes condicions la probabilitat d'acceptar el lot és

$$P = \frac{\binom{2}{0}\binom{38}{8}}{\binom{40}{8}} = 0{,}636$$

La distribució normal

Moltes de les tasques estadístiques fan referència a la distribució normal. No és aquesta la única distribució de probabilitat possible, però sí la més utilitzada. La correcta utilització de les operacions de tipificació d'una normal, així com el bon ús que de les taules d'una normal es pugui fer resulta del tot imprescindible en la majoria dels estudis estadístics.

Bàsicament la distribució normal respon a la suposició que valors centrals en la distribució són els més freqüents i que els valors extrems els més poc freqüents. A més a més hi ha una simetria entre la distribució dels valors dels extrems; s'esperen la mateixa quantitat de valors superiors que inferiors al que hom pot considerar habitual.

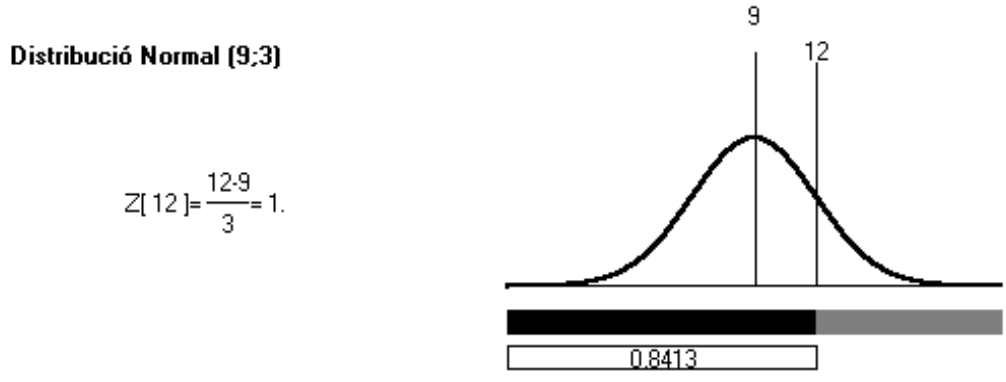


La funció que defineix una normal és

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

que depèn de x, com variable independent, i de dos paràmetres: la mitjana i la desviació típica. Fent la transformació definida per $z = \frac{x-\mu}{\sigma}$ es converteix en una normal de mitjana zero i desviació típica unitària, una normal de paràmetres N(0,1). Aquesta normal està tipificada; és a dir, els valors de les àrees sota la corba i l'eix OX estan calculats en taules. La posició central (diferència amb la mitjana) correspon ara al valor zero, i les "unitats" de mesura dels desplaçament relatiu a la mitjana són desviacions típiques. Allunyar-se una desviació típica de la mitjana correspon a una z=1 i això ens serveix quan el valor és 6 en una normal de mitjana 5 i desviació típica 1, però també serà el cas d'un valor 12 en una normal de mitjana 9 i desviació típica 3. En altres paraules, l'àrea sota la corba fins un punt z=1 d'una normal tipificada serà sempre la mateixa. I aquesta àrea és un valor de probabilitat.

El cas indicat pot correspondre a una situació com ara la que indica el gràfic



Per calcular el valor de probabilitat 0,8413 cal fer servir una taula N(0,1) i desplaçar-se fins el valor z=1.

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621

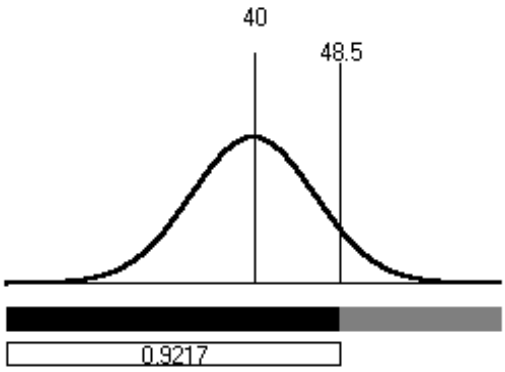
I això ens indica que si estem en una distribució normal, de mitjana 9 i desviació típica 3 podem esperar que el 84,13% dels valors siguin inferiors o iguals a 12.

Exemples

Si sabem que la distribució de les talles d'una determinada peça de roba segueix una distribució normal de mitjana la talla 40 i desviació típica 6, podem esperar que només el 8% de la població tingui una talla superior o igual a la talla 49

Distribució Normal (40;6)

$$Z[\,48.5\,]=\frac{48.5-40}{6}=1.4167$$



P[x<48,5]=0,9217 ; P[x>48,5]=0,0783

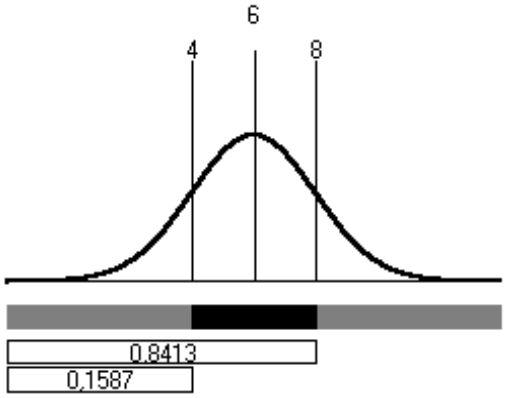
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319

Si les qualificacions d'una determinada prova són normals de mitjana 6 i desviació típica 2, el 68,26% dels alumnes tindran qualificacions entre 4 i 8 punts.

Distribució Normal (6;2)

$$Z[\,8\,]=\frac{8-6}{2}=1.$$

$$Z[\,4\,]=\frac{4-6}{2}=-1.$$



P[x<8]=0,8413 P[x<4]=0,1587 P [4 < x < 8] =0,6826

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621

Cues superiors i cues inferiors en la normal

Sovint es vol calcular valors que limiten una zona de la normal d'una determinada probabilitat. És el problema invers del que fins ara s'ha plantejat: Coneguda una probabilitat (un percentatge o una proporció) es vol determinar el valor de la distribució que fa que la probabilitat de valors superiors o inferiors sigui la indicada.

Coneguda la probabilitat, podem buscar en la taula N(0,1) el valor de z que limita aquesta àrea. Coneguda z (el valor tipificat de la distribució) i els valors de la mitjana i la desviació tipus podem aïllar el valor de x de l'equació

$$z = \frac{x - \bar{x}}{\sigma} \quad ; \quad x = \bar{x} + z\sigma$$

Exemples

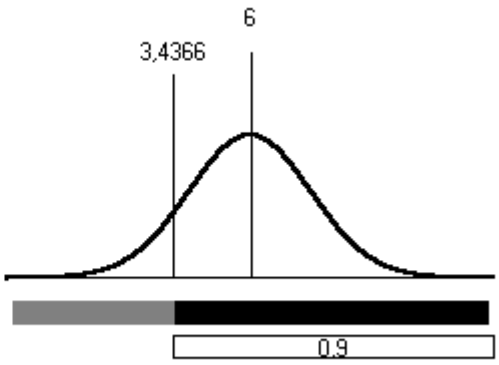
Una prova aplicada a uns estudiants té una mitjana de 6 i una desviació tipus de 2. Volem saber la qualificació que assoleixen almenys el 90% dels alumnes. Busquem a la taula N(0,1) un valor de probabilitat 0,9 que correspon a un valor tipificat de z entre 1,28 i 1,29. Podem interpolar a z=1,2817

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015

Aleshores apliquem

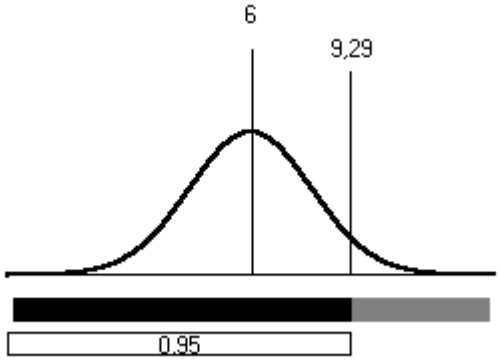
Distribució Normal (6;2)

Z [1,2817] = 0,9000
x = 6-(1,2817 . 2,)



Volem ara saber a partir de quina nota trobarem el 5% millors de la classe. Demanem un valor de k que faci que P[x>k]=0,05. Si tenim present que les probabilitat en la taula de la normal s'acumulen des de l'extrem inferior hem de buscar un valor 0,95 a la taula que correspon a z=1,645.

El valor que estem buscant serà



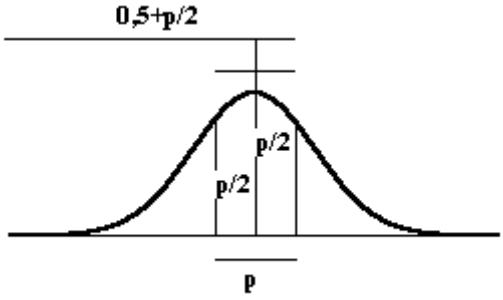
$$x = \bar{x} + z\sigma = 6 + 1{,}645 \cdot 2 = 9{,}29$$

Intervals centrats en la mitjana

En els exemples precedents hem calculat valors extrems d'un interval obert. Podem demanar si és possible calcular intervals que verifiquin una determinada probabilitat. No hi ha cap mena de limitació sobre cap dels extrems d'aquest interval que volem calcular, però el cas més habitual és demanar valors centrats en la mitjana de la distribució.

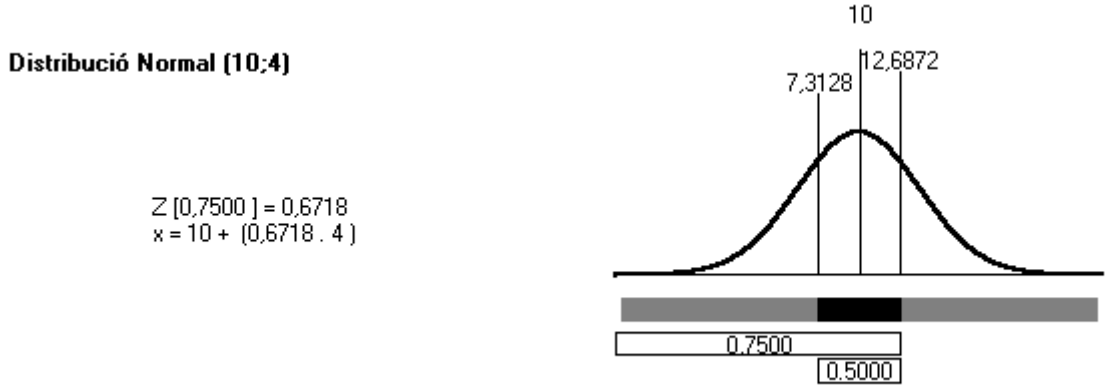
Volem saber els extrems d'un interval centrat en la mitjana que abasti una determinada probabilitat. Ja que és simètric serà de la forma $\bar{x} \pm k\sigma$ i el problema serà calcular el valor de k.

Per fer-ho hem d'observar la simetria de la normal. Un valor de probabilitat p centrat en la mitjana té p/2 a cada costat d'ella; aleshores la suma 1/2+p/2 serà el valor que haurem de buscar dins de la taula normal per trobar el valor de z que ho abasta.



Exemples

En una distribució N(10,4) un interval centrat en la mitjana que abasti el 50% dels casos serà



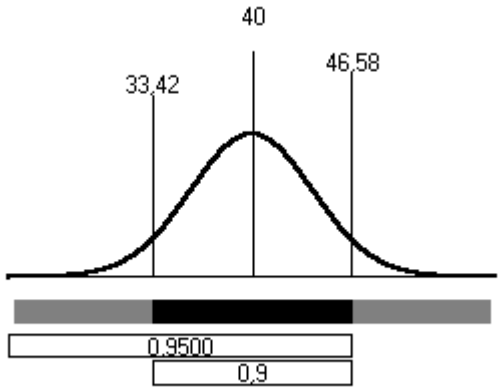
Cal observat que hem buscar a la taula el valor de probabilitat 0,75=0,5+0,25. La z corresponent està entre els valors 0,67 i 0,68. Si ho interpolem obtenim z=0,6718

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,6	0,7258	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7496	0,7518	0,7549

Si sabem que la duració d'un determinat trajecte segueix una distribució normal de mitjana 40 mn i desviació tipus 4 m, el 90% de les vegades que realitzem aquest trajecte tindrà una durada entre 33 i 47 mn.

Distribució Normal (40;4)

Z [0,9500] = 1,6450
x = 40 + (1,6450 . 4)



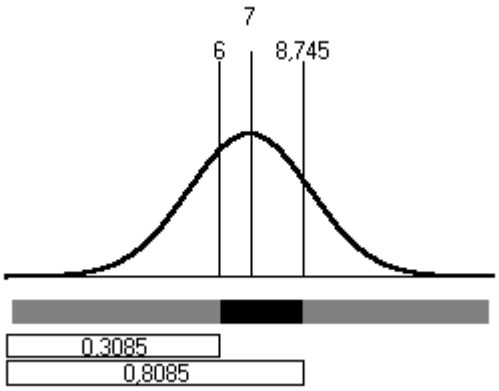
Si bé no és molt habitual, no hi ha cap impediment en calcular intervals de probabilitat fixat un valor de l'interval (el superior o l'inferior). En aquest cas es calcula la probabilitat de la mitjana fins l'extrem conegut i la resta de la probabilitat s'acumula a l'altre extrem.

Si considerem un interval del 50% de valor inferior 6 en una distribució normal de mitjana 7 i desviació tipus 2, els càlculs són:

Distribució Normal (7;2)

P [x<6]= 0,3085
0,3085 + 0,5000 = 0,8085

Z [0,8085] = 0,8725
sup = 7 + (0,8725 . 2,)



L'extrem superior és de 8,745, l'interval és (6 ; 8,745), que no està centrat en la mitjana.

Primer determinem la probabilitat de l'interval (6,7); del valor inferior a la mitjana. Z=-0,5 i correspon a una probabilitat de 0,1915=0,5-0,3085

Aleshores la zona superior a la mitjana ha de tenir una probabilitat de 0,3085; la zona acumulada és 0,8085 amb z=0,8725. Les consultes a la taula han estat

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,8	0,7881	0,7910	0,7939	0,7967	0,7996	0,8023	0,8051	0,8078	0,8106	0,8133

Distribució N(0,1)

	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91308	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997

Distribucions en mostres

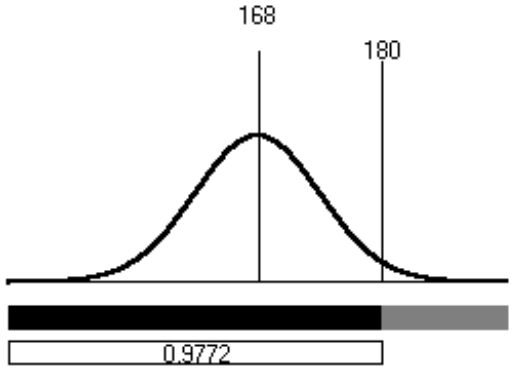
Considerada una població de N elements, es formen mostres de k<N elements. De totes aquestes mostres ens interessa saber la distribució d'alguns paràmetres. Els més habituals són:

Distribució de la mitjana en mostres

La mitjana de les mostres es distribueix amb mitjana igual a la mitjana de la població, i amb desviació típica $\sigma_x = \frac{\sigma}{\sqrt{k}}\sqrt{\frac{N-k}{N-1}}$. Quan la població es considera infinita o les mostres són amb reemplaçament podem considerar $\sigma_x = \frac{\sigma}{\sqrt{k}}$

Aquest valor és una desviació típica de les mitjanes, es coneix amb el nom d'error típic de la mitjana. Observem si k fos infinit (la mostra és tota la població) l'error seria zero i la mitjana de la mostra coincidiria amb la mitjana de la població. Segons augmenta k el valor de l'error típic de la mitjana es fa més petit

Considerem aquest exemple: Si la distribució de les altures d'una població segueix una distribució normal de paràmetres N(168,6), la probabilitat que una persona a l'atzar superi l'alçada 180 serà 0,0228



Però si en aquesta població prenem mostres de mida 100 i calculem la mitjana d'aquestes mostres, la probabilitat que una d'aquestes mostres tingui una mitjana superior a 180 es calcula

$$\sigma_x = \frac{\sigma}{\sqrt{k}} = \frac{6}{\sqrt{100}} = 0{,}6$$

sobre una distribució normal de mitjana 168 i desviació típica 0,6. Superar 180 en mitjana és pràcticament impossible ja que s'allunya 20 unitats tipificades de la mitjana

$$z = \frac{180-168}{0{,}6} = 20$$

Distribució de proporcions

Si en la població tenim una probabilitat p d'esdevenir una determinada característica en cada element, o determinem que hi ha una proporció p dels elements de la població amb una determinada característica, la

distribució d'aquesta proporció en les mostres que podem formar vindrà donada per una mitjana de p i una desviació típica $\sigma_x = \sqrt{\frac{p(1-p)}{k}}$

Distribució de sumes i diferències

Si ara considerem dues poblacions i dues mostres de mides k_1 i k_2 en cada població. Per cada una d'aquestes mostres considerem la distribució de les sumes o les diferències dels valors que s'obtenen. Les distribucions tindran de mitjana la suma o la diferència de les mitjanes, i com desviació típica l'expressió

$$\sigma = \sqrt{\frac{\sigma_1^2}{k_1} + \frac{\sigma_2^2}{k_2}}$$

Exemples

En una població de 3000 estudiants es coneix que les alçades es distribueixen normalment amb mitjana 170 i desviació típica 15. Si es pren una mostra de 25 estudiants la distribució de la mitjana de les alçades serà

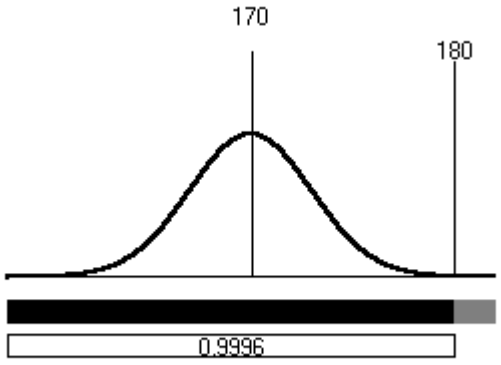
$$N(170; \frac{15}{\sqrt{25}}) = N(170;3)$$

Observem que aquesta és la distribució de la mitjana de les mostres, no de cada valor en particular. Podem considerar que un valor individual és un mostra de mida 1; en aquest cas la distribució d'aquest valor individual coincideix amb la distribució de la mitjana de les possibles mostres de mida 1.

Si volem calcular la probabilitat que la mitjana d'aquesta mostra de mida 25 superi 180 fem servir una tipificació de la normal i obtenim

$$Z[180] = \frac{180-170}{3} = 3.3333$$

1-0.9996=0.0004



Serà aquesta una probabilitat molt petita.

La proporció de cares en 120 llançaments d'una moneda serà una normal de mitjana p=0,5 i desviació típica $\sigma = \sqrt{\frac{p(1-p)}{k}} = \sqrt{\frac{0,25}{120}}$

Dos fabricants ofereixen el mateix producte que té una duració útil que segueix distribucions normals de paràmetres N(1400,200) i N(1200,100). Si prenem mostres aleatòries de 125 elements de cada un dels fabricants, la diferència en les durades mitjanes tindrà una distribució

$$N(\mu_1 - \mu_2; \sqrt{\frac{\delta_1^2}{k_1} + \frac{\delta_2^2}{k_2}}) = N(200; \sqrt{\frac{100^2}{125} + \frac{200^2}{125}})$$

Estimes de paràmetres de la població

Un dels problemes bàsic en l'estudi estadístic el genera la impossibilitat d'estudiar tots els elements de la població i haver de limitar l'estudi a un grup reduït d'elements; el que hom anomena mostra

Si es dona el cas d'abastar tots els elements de la població, el problema es redueix a un senzill anàlisi estadístic descriptiu. Caldrà només calcular els paràmetres de la població i treure'n les conclusions que d'ells es derivin. Aquest no és un problema que estudia la teoria de mostres.

Per fixar les idees, si la nostra població són els alumnes de l'Institut i volem saber el percentatge que presenta una determinada característica (són nois, tenen una idea política, treuen unes determinades qualificacions,...) podem considerar que està al nostre abast preguntar a tots i cada un dels 500 i escaig alumnes. En aquest cas no estem en un problema de teoria de mostres. Només ens cal recollir les observacions, quantificar-les i estudiar-les.

Però si pensem que no podem preguntar a tots els alumnes del Centre i que hem de limitar-nos a un subconjunt, ara sí que tenim un problema de teoria de mostres. La primera feina serà fixar els alumnes que ens cal estudiar i el nivell que volem que les seves respostes siguin representatives de la població.

Les estimes del paràmetres de la població no han de referir-se a un determinat valor, no han de ser puntuals. Afirmacions del tipus: "La mitjana és..." s'han de canviar per "La mitjana està entre els valors..." D'aquestes estimacions s'anomenen estimacions per intervals i s'han d'acompanyar d'un nivell de significació, d'una probabilitat d'encert

Mostres

La selecció dels elements que han de constituir la mostra és un dels moments més delicats. La llei dels grans nombres assegura que aquesta mostra serà representativa de la població quan els elements es determinin a l'atzar. Aquesta manera d'actuar fa que formació de mostres basades en respostes voluntàries de qüestionaris, enquestes telefòniques que puguin no ser contestades, o altres elements que facin que es trenqui aquesta hipòtesi d'esdeveniments escollits aleatòriament hagi de ser qüestionada.

En principi existeixen dues maneres inicials de formar les mostres: considerant que un element seleccionat pot ser escollit de nou (mostres amb reemplaçament) o bé evitant que aquest element ja seleccionat pugui tornar a considerar-se (sense reemplaçament). Des d'un punt de vista teòric les mostres amb reemplaçament són més fiables que les que es fan sense reemplaçament, si bé, a efectes pràctic, la majoria de mostres tenen lloc sense reemplaçament.

Les maneres habituals de formació de mostres volen garantir que tots els elements de la població tinguin la mateixa probabilitat de formar part de la mostra. Pot fer-se mitjançant mostres simples a l'atzar, on s'escull k de N elements de la població per alguna tècnica aleatòria, per exemple fent correspondre a cada element de la població un número i determinar aquells que formen la mostra a partir de taules de nombres aleatoris. (Veure taules). Un mètode de mostreig més elaborat és dividir la població en subconjunts (que poden existir de forma natural o no) anomenats estrats. Es determina el pes w_i de cada estrat com la

relació entre els elements de l'estrat i els elements de la població, i s'escull una mostra de cada estrat. Aquesta formació de mostres s'anomena estratificada.

Una variació del mostreig estratificat és el sistemàtic, quan els estrats tenen la mateixa mida. Si en la població hi ha k estrats de n elements podem formar les sèries $x_i, x_{i+n}, x_{i+2n}, ..$ un de cada un dels estrats.

Per últim cal considerat el mostreig bietàpic, quan existeixen uns elements primers que contenen elements segons (components dins de capses; cases en carrers,...) En aquest cas es consideren formació de mostres dels primers elements, per determinar les formacions de mostres dins d'aquests.

Mostreig estratificat

Anomenem estrats als diferents conjunts "naturals" que formen la població. Malgrat que l'adjectiu "natural" sigui poc precís es pot considerar que els estrats són aquelles divisions en la població que ja existeixen. Si considerem els alumnes d'un centre d'ensenyament secundari els tres possibles estrats estan formats per alumnes d'ESO, de Batxillerat i de CF. Si la població són els turistes d'una determinada localitat, els estrats poden diferenciar la procedència, nacional o no. Si la població és el parc de vehicles de motor d'una localitat, els estrats són els diferents tipus de vehicles: motocicletes, automòbils, camions, autocars,..

Amb tot la consideració dels diferents estrats pot ser encara poc precisa, en la situació concreta cal definir-los clarament i considerar el tipus de divisió òptima que s'ajusti més al nostre estudi.

Un cop determinats els estrats, s'han de fixar les mides de les diferents mostres que cal extreure. Determinar aquestes mides es coneix amb el nom de pes de l'estrat i es representa per W_i i ha de verificar que $\sum W_i = 1$

Un pes assignat de manera que $W_i = \frac{1}{K}$, on K són el nombre d'estrats de la població, s'anomena simple. Dóna el mateix pes a cada estrat, sense considerar els elements que els formen. És un mètode que podem aplicar quan es desconeix la quantitat d'elements que formen cada estrat.

Si es considera aquests n_i elements de cada estrat, on la seva suma ha de ser N, la mida de la població, podem assignar un pes proporcional a la mida de cada estrat, és a dir $W_i = \frac{n_i}{N}$. La majoria de mètodes de mostreig segueixen aquest criteri.

El mostreig òptim és aquell que assigna un pes tenint present la variabilitat de cada estrat, i no només la quantitat d'elements que el formen. Si σ_i és de desviació típica de cada estrat i σ la desviació de la població, una assignació òptima ha d'establir un pes $W_i = \frac{n_i \sigma_i}{N\sigma}$. Naturalment la dificultat en aplicar aquest mètode és el necessari coneixement d'aquests paràmetres de la població i dels estrats.

Exemples

Coneixem les dades que indica la taula sobre els treballadors d'una empresa i els seus guanys mensuals

		Sou	
	Treballadors	Mitjana	Desviació típica
categoria A	448	392	54
categoria B	296	611	93
categoria C	83	724	137

Si es decideix consultar a 60 d'aquests treballadors la distribució de la mostra ens els diferents estrats que formen les categories serà

Assignació simple.-
El pes de cada estrat és 1/3. Les mostres són de mida 20 en cada estrat

Assignació proporcional.-
En aquest cas el pes de cada estrat serà proporcional a la seva mida

categoria A	$W = \frac{448}{827} = 0{,}542$	$0{,}524 \cdot 20 = 10{,}48 \approx 10$
categoria B	$W = \frac{296}{827} = 0{,}378$	$0{,}378 \cdot 20 = 7{,}56 \approx 8$
categoria C	$W = \frac{83}{827} = 0{,}100$	$0{,}100 \cdot 20 = 2$

Assignació òptima.-
Primer calculem

$$\sum n_i \sigma_i = 448 \cdot 54 + 296 \cdot 93 + 83 \cdot 137 = 63.091$$

categoria A	$W = \frac{448 \cdot 54}{63091} = 0{,}383$	$0{,}383 \cdot 20 = 7{,}66 \approx 8$
categoria B	$W = \frac{296 \cdot 93}{63091} = 0{,}436$	$0{,}436 \cdot 20 = 8{,}72 \approx 9$
categoria C	$W = \frac{83 \cdot 137}{63091} = 0{,}180$	$0{,}180 \cdot 20 = 3{,}6 \approx 3$

Mitjana poblacional

Si tenim una mostra aleatòria de n observacions, la mitjana de la mostra es pren con estima de la mitjana de la població, però no podem saber, de moment, si és molt o poc propera a la mitjana poblacional que intentem calcular. Conegut el cas invers, de formació de mostres de mida n d'una determinada població, sabem que la mitjana de les mostres té una distribució $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Fent servir aquest resultat podem obtenir diferents intervals on la probabilitat d'abastar la mitjana poblacional dependrà del valor de una λ que calcularem en una taula N(0,1).

Alguns valors de λ habituals són els que figuren en la taula

0,90	λ=1,65	$\bar{x} \pm 1{,}65 \frac{\sigma}{\sqrt{n}}$
0,95	λ=1,96	$\bar{x} \pm 1{,}96 \frac{\sigma}{\sqrt{n}}$
0,99	λ=2,58	$\bar{x} \pm 2{,}58 \frac{\sigma}{\sqrt{n}}$

Observem un resultat notable: Si fixem σ i multipliquem per quatre la mida de la mostra obtenim un interval

$$\bar{x} \pm \lambda \frac{\sigma}{\sqrt{4n}} = \bar{x} \pm \lambda \frac{\sigma}{2\sqrt{n}}$$

que resulta ser meitat de l'anterior en el cas n.

Exemple

Una mostra de mida 100 d'una població normal amb σ=4 dóna una mitjana $\bar{x} = 6$. Amb una probabilitat del 95% la mitjana de la població estarà dins de l'interval

$$\bar{x} \pm 1{,}96\frac{\sigma}{\sqrt{n}} = 6 \pm 1{,}96\frac{4}{10} = 6 \pm 0{,}784$$

Si la mostra ara té 900 elements i obtenim les mateixes dades, l'interval per a la mitjana poblacional es redueix a

$$\bar{x} \pm 1{,}96\frac{\sigma}{\sqrt{n}} = 6 \pm 1{,}96\frac{4}{30} = 6 \pm 0{,}261$$

Estima de la variancia poblacional.

Una seriosa limitació a aquest càlcul és la presència del valor de la desviació típica poblacional. Habitualment aquesta no serà coneguda i hem de considerar com un possible estimador de la mateixa la desviació tipus corregida de la mostra

Una primera aproximació a la solució del problema es considerar com interval de confiança per a la desviació tipus poblacional l'interval

$$s \pm \lambda\frac{s}{\sqrt{2n}}$$

on λ té el mateix sentit que en el cas anterior.

Exemple

Si en una mostra de mida 200 la desviació tipus és 100, un interval al nivell del 95% per a la desviació tipus de la població és

$$100 \pm 1{,}96\frac{100}{\sqrt{400}} = 100 \pm 9{,}8$$

Estima de proporcions

Fent servir l'aproximació normal d'una distribució binomial $N(np, \sqrt{npq})$ podem calcular intervals de proporcions en la població a partir de proporcions en mostres fent

$$p \pm \lambda\sqrt{\frac{pq}{n}}$$

Com exemple, si una mostra de 100 elements dóna en ells una proporció p=0,46. La proporció que podem estimar en la població, amb probabilitat 0,95 és

$$0{,}46 \pm 1{,}96\sqrt{\frac{0.46.0.54}{100}} = 0{,}46 \pm 0{,}098$$

Petites mostres

Es considera una mostra petita quan n<30 En aquest cas estimar els valors de la població a partir dels valors d'aquesta petita mostra segueix un camí diferent.

Sovint les estadístiques de petites mostres es fan servir per calcular uns valors inicials que serveixen d'origen a càlculs més acurats fent servir mostres més amples.

Estimació de la mitjana de la població

En aquest problema es considera la distribució de l'estadístic t de Student

$$t = \frac{\bar{x} - \mu}{s}\sqrt{N-1}$$

d'un significat semblant a la tipificació z d'una distribució Normal. Els intervals de confiança a un nivell α per a la mitjana de la població vindran donats per

$$\bar{x} \pm t_\alpha \frac{s}{\sqrt{N-1}}$$

Els graus de llibertat de la distribució t és la constant que determina el nombre d'observacions independents de la mostra menys el nombre de paràmetres de la població que cal estimar. Si volem estimar la mitjana de la població (un paràmetres) els graus de llibertat coincidiran amb la mida de la mostra menys 1. Els valors de t_α es calculen en una taula d'una distribució t d'Student.

La forma de la distribució t recorda a la forma de la distribució normal. Segons els graus de llibertat aquesta corba té les cues de probabilitat superior a la corresponent normal.

Exemples

Si una mostra de mida 10 té de mitjana 0,053 i desviació típica 0,003, un interval per a la mitjana poblacional al 95% , t_α=2,26 amb 9 graus de llibertat

$$\bar{x} \pm t_\alpha \frac{s}{\sqrt{N-1}} = 0{,}053 \pm 2{,}26\frac{0{,}003}{\sqrt{9}} = 0{,}053 \pm 0{,}0002$$

Una mostra de 10 mesures dóna una mitjana de 4,38 i una desviació típica de 0,06. Els intervals per a la mitjana de la població, al 95% i al 99% venen donats per una t d'Student amb 9 graus de llibertat de valors

$t_{0,975}$=2,26
$t_{0,995}$=3,25

Els intervals seran

$$4{,}38 \pm 2{,}26\frac{0{,}06}{\sqrt{9}} = 4{,}38 \pm 0{,}045$$

i

$$4{,}38 \pm 3{,}25\frac{0{,}06}{\sqrt{9}} = 4{,}38 \pm 0{,}065$$

Comparació de mitjanes

La prova t s'utilitza també quan es volen comparar les mitjanes de dues mostres. En realitat és un contrast d'hipòtesi nul·la $\mu_1=\mu_2$ d'igualtat entre les dues mitjanes.

La diferència de mitjanes en mostres de mides n_1 i n_2 tindrà n_1+n_2-2 graus de llibertat, ja que s'estimen dos paràmetres, un en cada una de les mostres. Resulta útil si es vol comparar si dues mostres s'originen

de la mateixa població, que té mitjana única. A més a més la desviació típica de la població no ens cal si volem aplicar aquest contrast, podem utilitzar la desviació típica (corregida o no) de la mostra.

L'estadístic que es forma és

$$t = (\bar{x} - \bar{y})\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{(n_1 + n_2)(n_1 s_1^2 + n_2 s_2^2)}}$$

Exemples

Dues mostres de mides respectives 7 i 5 donen de mitjanes 7,07 i 9. Les variances respectives són 1,32 i 0,5. Si es vol determinar si existeixen diferències significatives en les dues mitjanes calculem

$$t = (\bar{x} - \bar{y})\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{(n_1 + n_2)(n_1 s_1^2 + n_2 s_2^2)}} = (7{,}07 - 9)\sqrt{\frac{7 \cdot 5 \cdot (7 + 5 - 2)}{(7 + 5)(7 \cdot 1.32 + 5 \cdot 0.5)}} = -3.04$$

si comparem aquest valor amb una distribució t amb 10 graus de llibertat al nivell 0,1 obtenim un valor de 1,812. S'ha d'acceptar que hi ha diferències significatives en les mitjanes de les dues mostres.

La producció d'uns arbres fruiters de la mateixa espècie en dues parcel·les diferents és

Parcel·la 1: 30 - 37 - 43 - 30 - 27 - 38 - 39
Parcel·la 2: 40 - 37 - 47 - 45 - 42 - 39 - 45 - 44 - 42

Les dues mostres tenen mides respectives de 7 i 9 elements, les mitjanes i les desviacions típiques són, respectivament

Parcel·la	Mida	Mitjana	Desv. típica
1	7	34,857	5.436
2	9	42,333	3.055

Estem en un cas de mostres de mides diferents. L'estadístic de contrast és

$$t = (\bar{x} - \bar{y})\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{(n_1 + n_2)(n_1 s_1^2 + n_2 s_2^2)}} = 7{,}476\sqrt{\frac{7 \cdot 9 \cdot 14}{16(7 \cdot 5{,}436^2 + 9 \cdot 3{,}055^2)}} = 0{,}8698$$

La distribució té 14 graus de llibertat. Si fixem un nivell del 95% el valor que dóna la taula t d'Student és 2,145. Hem d'acceptar que no hi ha diferències significatives en les mitjanes de la producció de les dues parcel·les.

Considerem el mateix exemple anterior on les dades de les mostres presenten una dispersió menor. Imaginem les mateixes mitjanes però desviacions típiques unitàries en les dues mostres. L'estadístic de contrast seria

$$t = (\bar{x} - \bar{y})\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{(n_1 + n_2)(n_1 s_1^2 + n_2 s_2^2)}} = 7{,}476\sqrt{\frac{7 \cdot 9 \cdot 14}{16(7 \cdot 1^2 + 9 \cdot 1^2)}} = 13{,}874$$

S'hauria d'acceptar, fins i tot a un nivell del 99,9%, que les mitjanes de producció són diferents. En aquest cas podem observar la sensibilitat de la prova t a les variances de les mostres.

Parells de mostres

En el cas de dues variables aleatòries definides sobre la mateixa població és freqüent estudiar un conjunt de parells de dades, cada una del mateix element de la població, però de valors cada una de les variables aleatòries. El contrast de la diferència de mitjanes nul·la en les variables correspon ara a una distribució de mitjana 0 i de graus de llibertat n-1, on n són els parells de dades.

L'estadístic de contrast serà

$$t = \frac{\bar{d}}{s}\sqrt{n-1}$$

indicant per $\bar{d}$ la mitjana de les diferències i s la variança

Exemples

La taula indica els resultats de 10 alumnes en dues proves diferents. La diferència dels resultats i la mitjana i variança de les diferències és

57	49	60	55	57	48	50	61	52	56
55	48	58	56	54	48	52	56	50	58
2	1	2	-1	3	0	-2	5	2	-2

$$\bar{d} = 1 \quad s = 2{,}144$$

El càlcul de t és

$$t = \frac{\bar{d}}{s}\sqrt{n-1} = \frac{1}{2{,}144}\sqrt{9} = 1{,}4$$

A un nivell del 0,95 t=2,262, s'accepta que no hi ha diferències significatives en les mitjanes.

La mesura d'un paràmetre fisiològic en vuit pacients abans i després d'un tractament mèdic és

56	56	147	58	121	57	49	118
47	63	125	26	99	36	34	90

Formem les diferències i calculem la mitjana i la desviació típica

$$\bar{d} = -17{,}75 \quad s = 11{,}465$$

El càlcul de t és

$$t = \frac{\bar{d}}{s}\sqrt{n-1} = \frac{-17{,}75}{11{,}465}\sqrt{7} = -4{,}09$$

A un nivell del 95% la distribució t amb 7 graus de llibertat és el valor que dóna la taula per a $t_{0,975}$=2,36. Aquest valor cau fora de l'interval d'acceptació de la hipòtesi nul·la i hem d'afirmar que hi ha diferències significatives en les mitjanes.

Estimació de la desviació típica poblacional

En aquest cas s'ha de fer servir una distribució chi-quadrat amb les cues corresponents al nivell de confiança establert. La desviació típica poblacional estarà entre

$$\frac{s^2 N}{\chi^2_{\alpha}} < \sigma^2 < \frac{s^2 N}{\chi^2_{1-\alpha}}$$

on s és la desviació típica de la mostra

Exemples

Una mostra de mida 16 dóna una desviació típica de 2,4. L'interval al 95% per a la desviació típica poblacional vindrà donat per una chi-quadrat

$\aleph^2_{0,975}=27,5$
$\aleph^2_{0,025}=6,25$

amb 15 graus de llibertat. Observem que els extrems centrats a 0,95 abasten des de 0,025 fins 0,975. La variança de la mostra és $2,4^2=5,76$

L'interval de la variança tindrà d'extrems

$$\frac{s^2 N}{\chi^2_\alpha} < \sigma^2 < \frac{s^2 N}{\chi^2_{1-\alpha}} \quad ; \frac{5,76 \cdot 16}{27,5} < \sigma^2 < \frac{5,76 \cdot 16}{6,25}$$

que formen l'interval ($3,35<\sigma^2<14,74$) de la variança i ($1,8<\sigma<3,8$) de la desviació típica.

Una mostra de mida 18 d'una població normal té una variança de $s^2=61$. Volem saber si podem acceptar, a un nivell del 0,1, que la variança poblacional és $s^2=50$.

Formem un interval amb els valors d'una chi-quadrat de 17 graus de llibertat, d'extrems 0,05 i 0,95

$$\aleph^2_{0,05} = 8,67 \quad \aleph^2_{0,95} = 27,6$$

L'interval d'acceptació de la desviació típica serà

$$\frac{s\sqrt{N}}{\chi_\alpha} < \sigma < \frac{s\sqrt{N}}{\chi_{1-\alpha}} \quad ; \frac{\sqrt{61}\sqrt{18}}{\sqrt{27,6}} < \sigma < \frac{\sqrt{61}\sqrt{18}}{\sqrt{8,67}}$$

$6,3<\sigma<11,3$

d'on la variança està en l'interval $39,7<\sigma^2<127,7$, que conté el valor 50. A aquest nivell es pot acceptar que la variança poblacional és 50

Mida de mostres

Si es vol conèixer un paràmetre poblacional, la mitjana, per exemple, i s'accepta que és impossible, o molt costós analitzar tots i cada uns dels membres de la població, sorgeix la necessitat de treure una mostra de la població i, segons els valors que d'ella s'obtenen, la mitjana en aquest cas, imaginar el valor de la mitjana de tota la població.

La qüestió clau en aquest mètode de mostreig és determinar la mida de la mostra. A quantes persones del nostre barri hem de preguntar la seva opinió per deduir quina és l'opinió de tots els habitants de barri? A quants alumnes del nostre Centre hem de preguntar abans de afirmar alguna cosa de tots els alumnes del Centre? De quantes bosses hem de recomptar els caramels que contenen abans d'imaginar el contingut mitjà de caramels de totes elles?

Hi ha una sèrie d'elements que s'han de determinar prèviament. El primer és la suposició d'un model de distribució que pot seguir la població. Habitualment ens va bé acceptar que la població es distribueix normalment, el segon element és l'error màxim que estem disposats a acceptar, és a dir, la diferència màxima amb el veritable valor. El tercer és el nivell de confiança, que és la probabilitat que tenim de no equivocar-nos en les nostres afirmacions. Lamentablement aquesta probabilitat mai pot ser 1, només en el

cas que estem disposats a analitzar tota la població, però aleshores ja no estem en un problema de determinació de la mida de la mostra .

Habitualment es consideren els nivells de confiança del 0,9 (o 90%) del 0,95 o del 0,99 segons si es vol una seguretat suficient, bona o molt bona. Per a cada un d'ells els valors que s'obtenen en una taula N(0,1) són de z 1,645; 1,96 i 2,58. D'aquesta manera es poden fitar valors de mitjanes, de proporcions,.. en la població a partir dels valors que s'obtenen en les mostres.

Existeix un quart factor, que depèn directament de les dades poblacionals, i aquest és la seva variancia. Si la població té una variancia gran, si els elements són molt diferents entre ells, caldrà una mostra de mida major. Si σ és la desviació típica de la població, la desviació típica d'una mostra de mida n és $\frac{\sigma}{\sqrt{n}}$ .Aquesta relació es pot veure en els exemples que segueixen

Exemples

Una mostra de 200 elements ofereix una mitjana de 0,824 i una desviació típica de 0,042. Si es vol determinar un interval que abasti la mitjana de la població a uns determinats nivells hem de considerar una distribució $N(\mu;\frac{\sigma}{\sqrt{k}}) = N(0{,}824;\frac{0{,}042}{\sqrt{200}})$. Els valors que obtenim són:

Nivell de confiança	interval
0,90	0,824±1,645.0,00297
0,95	0,824±1,96.0,00297
0,99	0,824±2.58.0,00297

Suposant coneguda la desviació típica d'una variable estadística en una població (per exemple 0,05) si volem saber la quantitat de mostres que s'hauran de prendre per assegurar, a un nivell del 0,95, que l'error no sigui superior a 0,01 hem de resoldre l'equació

$$\frac{1{,}96\cdot\sigma}{\sqrt{k}} \le 0{,}01 \quad \Rightarrow \sqrt{k} \ge \frac{1{,}96\cdot 0{,}05}{0{,}01} = 9{,}8 \quad \Rightarrow k \ge 96{,}04$$

Ens cal una mostra de 97 elements

Una mostra de 100 persones dóna que el 55% estan a favor d'una determinada opció. Si volem calcular un interval amb un nivell de significació del 99% que ens informi de la proporció de la població que escull aquesta opció hem de calcular

$$p \pm 2{,}58\sqrt{\frac{0{,}55\cdot 0{,}45}{100}} = p \pm 0{,}13 = 0{,}55 \pm 0{,}13$$

la població escull aquesta opció amb un percentatge que va des del 42% al 68% amb un nivell de confiança del 99%

Mida de mostres. Mitjanes

En una població que imaginem és normal volem calcular la seva mitjana amb un error E i un nivell de confiança α. Si coneixem la desviació típica de la població la mida de la mostra ve donada per

$$n = \frac{z^2\sigma^2}{E^2}$$

Observem que s'ha de conèixer la desviació típica de la població si volem aplicar el resultat anterior. Si no es coneix es pot estimar a partir de la variancia mostral. Una primera mostra, de mida petita, determina aquesta, en funció del resultat es calcula la mida de la mostra definitiva

Exemples

Amb un nivell de confiança del 95% volem estimar la mitjana de la població amb un error d'una unitat. La desviació típica de la població és 10,4. La mostra he de ser d'una mida de

$$n = \frac{z^2\sigma^2}{E^2} = \frac{1{,}96^2 \cdot 10{,}4^2}{1^2} = 415{,}5 \approx 416$$

Si mantenim el mateix nivell de confiança però volem un error més petit, mitja unitat, la mostra haurà de tenir una mida quatre vegades superior

$$n = \frac{z^2\sigma^2}{E^2} = \frac{1{,}96^2 \cdot 10{,}4^2}{(1/2)^2} = 1662$$

Si seguim amb les mateixes suposicions anteriors i podem acceptar fer un mostreig de només 200 elements, l'error que hem d'acceptar cometre és

$$E = \frac{z \cdot \sigma}{\sqrt{n}} = \frac{1{,}96 \cdot 10{,}4}{\sqrt{200}} = 1{,}44$$

Mida de mostres. Proporcions

Si es vol calcular una determinada proporció d'una població normal, fem servir que la desviació típica d'una distribució binomial de probabilitat d'èxit p és

$$\sigma = \sqrt{p(1-p)}$$

i la mida de la mostra es calcularà

$$n = \frac{z^2\sigma^2}{E^2} = \frac{z^2 p(1-p)}{E^2}$$

En el cas d'estima de proporcions resulta important conèixer que el màxim valor per a z i E fixades és quan p=1-p=0,5. En aquest cas és parla del supòsit de màxima indeterminació i és el que s'acostuma fer servir quan no hi ha cap mena de resultat conegut sobre la proporció de la població.

Exemples

Una empresa de vendes per correu coneix que el 80% dels seus clients són dones. Vol consultar si aquesta proporció ha canviat darrerament i està disposada a acceptar un error de 5% amb un nivell de confiança del 99%. La mostra ha de tenir una mida de

$$n = \frac{2{,}58^2 \cdot 0{,}8 \cdot 0{,}2}{0{,}05^2} = 426$$

en aquest cas z=2,58 ja que es vol un nivell de confiança del 99%, força alt.

En el mateix cas els canvis en els hàbits de consum fan que ja no sigui fiable la suposició del 80% de les dones com a clients. La mida de la mostra hauria de ser de

$$n = \frac{2{,}58^2 \cdot 0{,}5 \cdot 0{,}5}{0{,}05^2} = 665{,}6 \approx 666$$

Naturalment la mida de la mostra és superior. Estem en el cas de màxima indeterminació

Hipòtesis estadístiques

Quan s'enuncien suposicions sobre un determinat valor estadístic es fan hipòtesis estadístiques que cal comprovar analitzant valors de les mostres. Un vegada formulada una hipòtesis aquesta pot acceptar-se o rebutjar-se, i en aquests dos casos, es pot cometre un error.

Si anomenem H_0 a la hipòtesi formulada (hipòtesi nul·la) tota altra que difereixi de la primera s'anomena hipòtesi alternativa i es representa per H_1. Aleshores podem estar en una de les quatre situacions que indica la taula

		H_0 és:	
		Certa	Falsa
Decisió	Acceptem H_0	Decisió correcta	Error tipus II
	Rebutgem H_0	Error tipus I	Decisió correcta

Una decisió correcta és acceptar H_0 quan sigui certa o rebutjar H_0 quan aquesta sigui falsa, però el problema és el nostre desconeixement de la situació real, no sabem si la hipòtesi que formulem és o no certa i, possiblement, no es podrà saber amb exactitud. Hem d'acceptar que podem cometre un error.

La consideració dels dos tipus d'error que podem cometre és molt important i està en funció del tipus de dada que analitzem. Per exemple si una hipòtesi nul·la és que una persona diu la veritat, un error tipus I serà creure que ens diu una mentida quan està dient una veritat, i un error tipus II creure que diu la veritat quan en realitat menteix. Si formulem com hipòtesi que una persona està malalta, un error tipus I serà considerar-la sana quan està malalta, i un error tipus II considera que està malalt quan realment està sa.

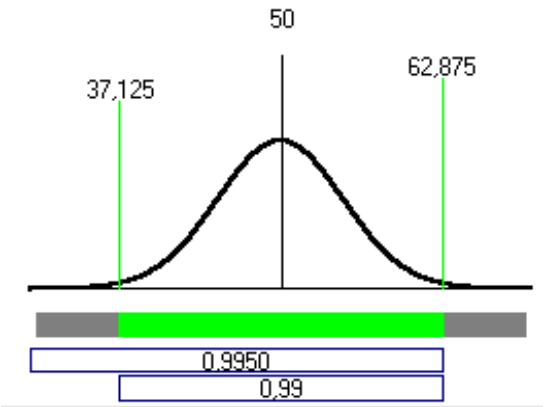
S'anomena nivell de significació a la probabilitat de cometre un error tipus I, es representa per α, i valors habituals són 0,1; 0,05 i 0,01. Si fixem un nivell de significació 0,05 indiquem que acceptem equivocar-nos rebutjant la hipòtesi cinc vegades de cada cent. Quan disminueix l'error tipus I augmenta l'error tipus II. La potència del contrast és la probabilitat complementària de la d'un error tipus II i es vol sigui un valor alt

Hipòtesis unilaterals i bilaterals

La hipòtesi nul·la pot ser rebutjada per excés i per defecte, o bé pot ser rebutjada només per excés o només per defecte. Si formulem com hipòtesis que una determinada proporció és p ens equivoquem quan en realitat aquesta sigui superior o inferior a p. Però si formulem com hipòtesis que una determinada proporció és superior a p només ens podem equivocar quan p sigui inferior al valor que ens hem imaginat. Aquestes són, respectivament, hipòtesis que poden contrastar-se bilateralment (o de dues cues) i unilateralment (una cua)

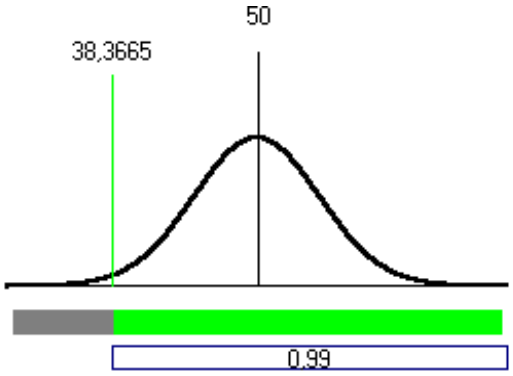
Exemples

Formulem com hipòtesi que una moneda és perfecta. Fem 100 llançaments i volem saber entre quins intervals ha d'estar la proporció de cares per acceptar aquesta hipòtesi a un nivell de significació prou baix, per exemple 0,01. La distribució del nombre de cares seguirà una normal de paràmetres 50 de mitjana i 5 de desviació típica. El valor que dóna una taula normal tipificada és 2,575.



Haurem d'acceptar que el nombre de cares estigui entre 37 i 63 si volem acceptar aquesta hipòtesi. És una hipòtesi bilateral. Podem equivocar-nos quan la moneda presenti menys o més cares.

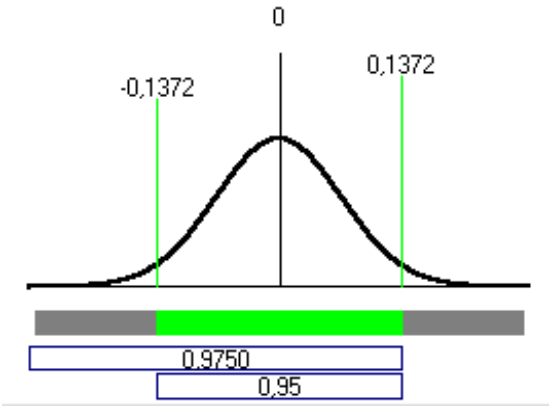
Sospitem ara que una moneda presenta més vegades cares que creus. En els mateixos 100 llançaments i al mateix nivell de significació tindrem una hipòtesi d'una cua que haurem d'acceptar si s'obtenen més de 38 cares



Si es vol considerar ara com hipòtesi nul·la que no hi ha diferències entre les proporcions p_1 i p_2 d'una característica en dues mostres de mides k_1 i k_2 es contrasta fent servir una distribució normal de mitjana 0 (la diferència de proporcions) i de desviació típica

$$\sigma = \sqrt{p(1-p)\left(\frac{1}{k_1}+\frac{1}{k_2}\right)}$$

És especialment interessant considerar el cas de màxima indeterminació, quan les proporcions són desconegudes. En aquest cas el màxim valor té lloc quan p=1-p=0,5. A un nivell del 0,05 i per una mida de mostres de 100 obtindrem $\sigma = \sqrt{0{,}5 \cdot 0{,}5 \cdot \frac{2}{100}} = 0{,}07$ i s'acceptarà quan la diferència de proporcions sigui menor del 13,7%

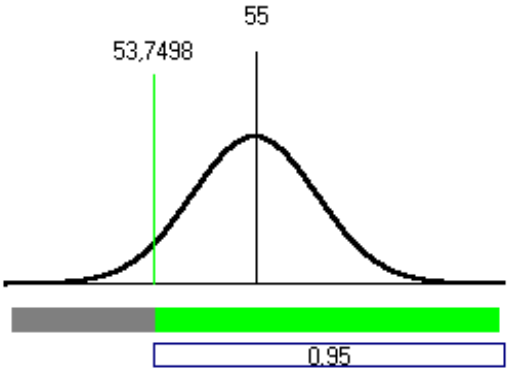


Sabem, de dades anteriors, que la mitjana d'edats dels clients d'una empresa és de 55 anys amb una desviació típica de 12 anys. Sembla que l'edat dels clients disminueix i volem saber si aquesta afirmació és certa. Per contrastar aquesta hipòtesi prenem una mostra de 250 clients que dóna una mitjana d'edat de 53 anys

Aquesta és una hipòtesi unilateral, que formulem com H_0:μ<55. La mostra ha de ser normal de mitjana 55 i desviació

$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{250}} = 0{,}76$$

Amb un nivell de significació del 0.05 podem acceptar mitjanes d'edat en la mostra fins



El valor de la mitjana de la mostra és inferior. Hem de rebutjar H_0 i acceptar que l'edat mitjana dels clients és inferior.

Assaig de diferència de mitjanes

Sovint es presenta el cas de decidir si dues mostres presenten o no diferències significatives en les seves mitjanes. Resulta útil plantejar una hipòtesi d'igualtat de mitjanes fent servir que la diferència de mitjanes segueix una distribució normal de mitjana zero i desviació

$$\sigma_{x_1-x_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

Considerem, com exemple, que una prova aplicada a dues classes de 40 i 50 alumnes dóna com resultats mitjanes de 74 i 78 amb desviacions tipus de 8 i 7. Volem saber si hi ha o no diferències significatives entre els resultats de les dues classes. Observem que la hipòtesi nul·la (igualtat de mitjanes) ha de contrastar-se de manera bilateral.

La desviació serà

$$\sigma = \sqrt{\frac{8^2}{40} + \frac{7^2}{50}} = 1{,}606$$

i la diferència de mitjanes tipificada

$$z = \frac{74 - 78}{1{,}606} = -2{,}49$$

A un nivell del 0,05 hem d'acceptar la hipòtesi d'igualtat de mitjanes quan |z|<1,96. Rebutgem la hipòtesi d'igualtat de mitjanes

Però a un nivell del 0,01ha de ser |z|<2,59. No podem rebutjar la hipòtesi a aquest nivell

Aquest és també el mètode aplicable en el cas de interessar si dues poblacions poden considerar-se equivalents en mitjanes a partir del coneixement que donen l'estudi de dues mostres

Assaig de diferències de proporcions

El cas semblant s'aplica quan comparem proporcions de mostres. La desviació tipus de la diferència de proporcions és

$$\sigma_{x_1 - x_2} = \sqrt{pq\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

Com exemple considerem dos grups de 100 malalts. Un grup es tracta amb una determinada substància i l'altre no. Les curacions en cada grup han estat de 75 i 65. Volem saber si la substància administrada és o no efectiva.

Contrastem una igualtat de proporcions amb una desviació tipus de

$$\sigma = \sqrt{0{,}75.0{,}65\left(\frac{2}{100}\right)} = 0{,}0648$$

La diferència de proporcions, tipificada, és

$$z = \frac{0{,}75 - 0{,}65}{0{,}0648} = 1{,}54$$

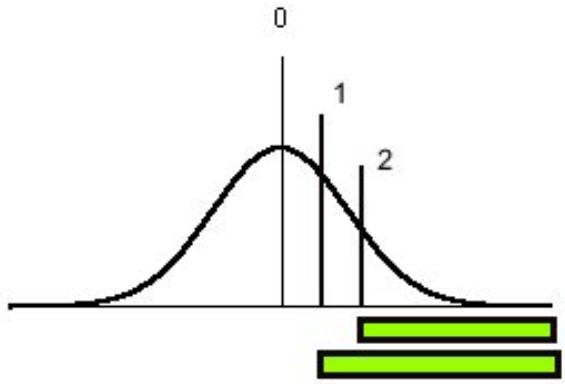
No es pot rebutjar la hipòtesi d'igualtat a un nivell del 0,01 ni del 0,05, però sí a un nivell del 0,1

Error tipus II

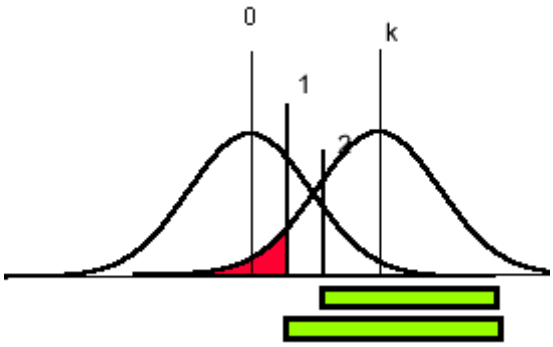
L'error tipus I pot determinar-se arbitràriament. L'error tipus II, o la probabilitat d'acceptar una hipòtesi falsa, no pot determinar-se si no es coneix el veritable valor del paràmetre de la població. Si formulem com hipòtesi que la mitjana d'una població és m, acceptem aquesta hipòtesi i resulta que ens equivoquem, serà perquè la mitjana de la població no és m, però, aleshores, quin és el seu valor?.

No podem contestar a aquesta pregunta. De fet no és una pregunta vàlida. Però sí que podem imaginar diferents valors de la mitjana poblacional i, per a cada un d'ells, calcular la probabilitat d'un error tipus II. Quan determinem un valor d'error tipus I estem també definint el valor de l'error tipus II per a una situació concreta, que no coneixem, però que podem estudiar.

Posem com exemple una hipòtesi d'una cua i considerem en ella dos valors crítics, x=1 i x=2. Per a cada un d'ells el valor de α serà la part marcada sota la corba de la normal, la probabilitat de rebutjar aquesta hipòtesi quan sigui certa.



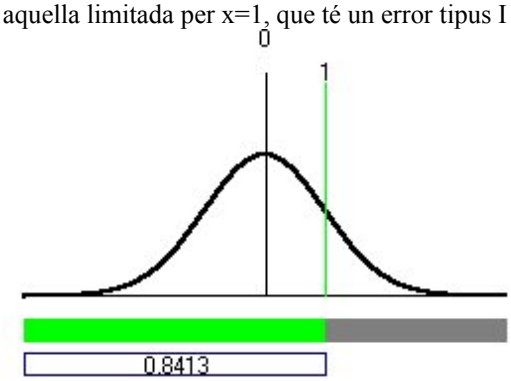
El valor crític 2 dóna un error tipus I menor que el valor 1. Si aquesta hipòtesi és ara falsa i l'acceptem serà perquè la veritable població tindrà una distribució de mitjana k desplaçada respecte de la primera. Si dibuixem les dues observem que l'error tipus II que correspon al valor crític 1 (la zona pintada sota la segona corba) és més petita que la zona corresponent al valor crític 2. En aquest cas, si l'error tipus I és el menor, l'error tipus II augmenta.



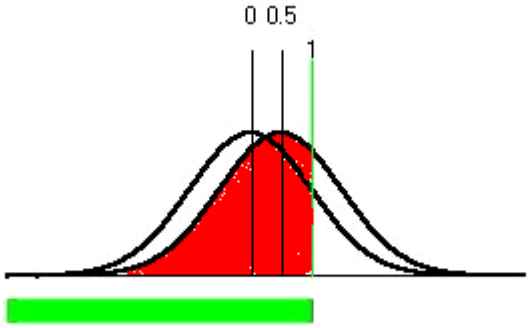
Corba de potència

La funció de potència dóna, per a unes determinades suposicions sobre la població, les probabilitats d'un error tipus II. Aquesta funció o corba de potència és pot calcular imaginant errors tipus II per a diferents valors de la mitjana de la població.

Considerem una hipòtesi H_0:μ=0 contra una hipòtesi alternativa H_1:μ>0. Considerem com regió crítica aquella limitada per x=1, que té un error tipus I de 0,1585.



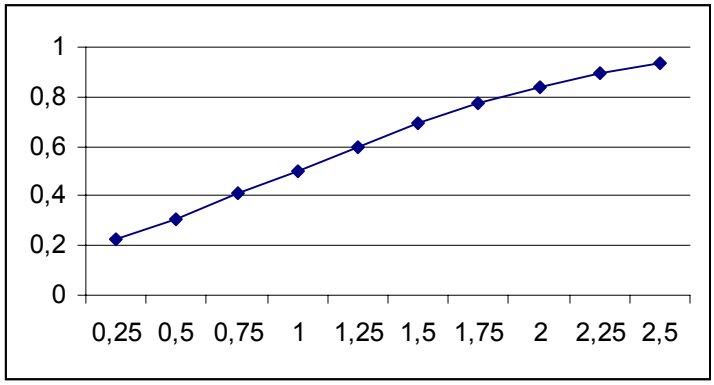
Imaginem ara diferents situacions de la veritable mitjana. Per exemple que és 0,5. L'error tipus II en aquesta regió crítica serà la indicada en el gràfic; el valor d'aquest error tipus II és 0,6915



Podem calcular el valor de l'error tipus II per a diferents valors de la mitjana, i els valors complementaris, anomenats potència de la prova

μ	0.25	0.5	0.75	1	1.25	1.5	1.75	2	2.25	2.5
error tipus II	0.7734	0.6915	0.5887	0.5	0.4013	0.3085	0.2266	0.1587	0.1056	0.0668
1-β	0.2266	0.3085	0.4113	0.5	0.5987	0.6915	0.7734	0.8413	0.8944	0.9332

La corba definida per 1-β és la corba de potència de la prova



Teoria d'ajust de distribucions

La prova chi-quadrat permet també estudiar si la diferència entre les freqüències teòriques i les freqüències observades pot considerar-se a l'atzar o si les diferències són prou significatives per rebutjar el model teòric de distribució. Si f_t i f_o són, respectivament, aquestes dues freqüències, s'ha de calcular la suma dels quadrats de les diferències sobre les freqüències teòriques. Aquesta suma es distribueix com una chi-quadrat amb una quantitat de graus de llibertat com en l'apartat anterior.

Exemples

En 200 llançaments d'una moneda s'observaren 120 cares. Calculem

$$X^2 = \frac{(120-100)^2}{100} + \frac{(80-100)^2}{100} = 8$$

El valor d'una chi-quadrat amb un grau de llibertat i a un nivell de significació del 0,01 és 6,63. Aquest valor es supera i hem de rebutjar que la moneda està ben feta

S'han estudiat 320 famílies de cinc fills. En la taula figura la distribució de fills i filles i les freqüències teòriques que s'obtenen al suposar que els naixements de nen i nena tenen la mateixa probabilitat. Es vol contrastar a un nivell de significació del 0,05 i del 0,01 si es pot considerar que el naixement d'un nen i el d'una nena és igualment probable. El valor que dóna l'estadístic chi-quadrat és 11,96.

$$\chi^2 = \sum \frac{(f_t - f_o)^2}{f_t} = 11{,}96$$

S'ha de concloure que a un nivell del 0,05 es rebutja la hipòtesis d'igualtat de probabilitats, però no es pot rebutjar a un nivell del 0,01. Els valors que dóna una taula de la chi-quadrat són 11,1 i 15,1 amb 6-1=5 graus de llibertat.

nens	nenes	famílies	probabilitat	F teòr.	$\frac{(f_t - f_o)^2}{f_t}$
5	0	18	$1/32$	10	6,4
4	1	56	$5/32$	50	0,72
3	2	110	$10/32$	100	1
2	3	88	$10/32$	100	1,44
1	4	40	$5/32$	50	2
0	5	8	$1/32$	10	0,4

Si llencem un dau 120 vegades, els resultats 1,2,... s'obtenen 25, 17, 15, 23, 24 i 16 vegades respectivament. Volem veure si el model teòric de que cada resultat té una probabilitat 1/6 i una freqüència esperada de 20, és prou bo.

L'estadístic chi-quadrat dóna

$$\chi^2 = \frac{(25-20)^2}{20} + \frac{(17-20)^2}{20} + \ldots + \frac{(16-20)^2}{20} = 5$$

Tenim cinc graus de llibertat en la distribució; els valors de la distribució $\chi^2_{0,95} = 11{,}1$ S'ha d'acceptar que el dau es comporta segons el model

Taules de contingència

Considerem el problema de determinar si hi ha relació o no entre dues característiques de la població, on s'han fet divisions en categories. Volem veure si les característiques poden considerar-se independents, en el sentit de no alterar el compliment d'una, la probabilitat de compliment de l'altra. Pensem, com exemple, en els alumnes d'un Centre classificats en nois i noies, i classificats també segons el seu rendiment escolar. En un segon exemple busquem si hi ha o no relació entre els conductors d'automòbils, classificats per edats, i el nombre de sinistres que declaren a les companyies d'assegurances. Volem veure si el fet de fumar o no fumar té alguna relació amb l'existència de problemes cardíacs. En aquests exemples podem formar una taula de contingència, on les files i les columnes indiquen les divisions que s'han format de les dues característiques que volem estudiar.

Siguin dues característiques A i B, cada una d'elles dividida en categories A_1, A_2,...A_n i B_1, B_2,...B_m. En cada una de les cel·les indiquem les observacions $n_{i,j}$ de la característica i en la categoria A i la característica j en la categoria B. Si formen els totals de fila i de columna, i els totals d'observacions podem construir una taula com ara

		Característica B				
		1	2	...	n	Totals
Característica A	1	n_{11}	n_{12}	...	n_{1n}	$N_{1.}$
	2	n_{21}	n_{22}	...	n_{2n}	$N_{2.}$
	...	...	...	...	...	
	m	n_{m1}	n_{m2}	...	n_{mn}	$N_{m.}$
	Totals	$N_{.1}$	$N_{.2}$	...	$N_{.n}$	N

Observem que un punt en el subíndex indica que estem sumant els totals de fila o de columna, la suma de la segona columna s'indica $N_{.2}$ i la suma de la tercera fila serà $N_{3.}$

Una hipòtesi d'independència entre les dues característiques determinarà que les freqüències teòriques en cada una de les cel·les hagin de ser el producte dels totals de la fila i de la columna de cada cel·la dividit per N, la mida de la mostra. En aquest cas les freqüències teòriques són:

		Característica B				
		1	2	...	n	Totals
Característica A	1	$N_{1.}N_{.1}/N$	$N_{1.}\ N_{.2}/N$	...	$N_{1.}\ N_{.n}/N$	$N_{1.}$
	2	$N_{2.}\ N_{.1}/N$	$N_{2.}\ N_{.2}/N$	...	$N_{2.}\ N_{.n}/N$	$N_{2.}$
	...	...	...	...	...	
	m	$N_{m.}\ N_{.1}/N$	$N_{m.}\ N_{.2}/N$	...	$N_{m.}\ N_{.n}/N$	$N_{m.}$
	Totals	$N_{.1}$	$N_{.2}$	...	$N_{.n}$	N

La distribució de

$$\sum \frac{(f_o - f_t)^2}{f_t}$$

Ha de comparar-se amb una distribució $\aleph^2$ de (m-1).(n-1) graus de llibertat. Indiquem per f_o i per f_t les respectives freqüències observades i teòriques. El contrast és efectiu quan les freqüències esperades en cada una de les cel·les de la taula sigui superior o igual a 5.

Exemples

Els alumnes d'un curs s'han classificat en noies i nois i en aprovats i no aprovats. Les freqüències i els totals són:

		Alumnes		
		noies	nois	Totals
Qualificació	aprovats	20	11	31
	no aprovats	15	14	29
	Totals	35	25	60

Les freqüències teòriques serien

	Alumnes		
	noies	nois	Totals

Qualificació	aprovats	31.35/60=18	31.25/60=13	31
	no aprovats	29.35/60=17	29.25/60=12	29
	Totals	35	25	60

on els càlculs s'arrodoneixen a nombres enters. No sempre convé arrodonir els resultats a xifres enteres, podem conservar els decimals. L'estadístic de contrast és

$$\aleph^2 = \frac{(20-18)^2}{18} + \frac{(11-13)^2}{13} + \frac{(15-17)^2}{17} + \frac{(14-12)^2}{12} = 1{,}0985$$

Tenim (2-1).(2-1)=1 grau de llibertat. Si consultem la taula d'una distribució chi-quadrat amb un grau de llibertat i a un nivell 0.01 veurem que és de 6,6349. Donat que aquest valor no es supera, no hi ha cap raó per rebutjar la hipòtesi nul·la d'independència entre el sexe dels alumnes i les qualificacions.

Un centre comercial sospita que els pagaments en efectiu i els pagaments fent servir targetes de crèdit depenen de l'època de l'any. Per això elabora una enquesta on pregunta als clients si han pagat en efectiu o fent servir targetes en les quatre estacions de l'any. Els resultats són:

	Primavera	Estiu	Tardor	Hivern	Totals
Efectiu	7	3	4	1	15
Targetes	3	3	5	4	15
	10	6	9	5	30

Les freqüències teòriques són:

	Primavera	Estiu	Tardor	Hivern	Totals
Efectiu	5	3	4,5	2,5	15
Targetes	5	3	4,5	2,5	15
	10	6	9	5	30

L'estadístic de contrast és

$$\aleph^2 = \frac{(7-5)^2}{5} + \frac{(3-3)^2}{3} + \ldots + \frac{(4-2{,}5)^2}{2{,}5} = 3{,}51$$

La distribució té 3 graus de llibertat (4-1)(2-1). El valor que dóna una taula chi-quadrat a un nivell del 0,05 és 7,81. Aquest valor no es supera i s'ha d'acceptar que no hi ha relació entre les diferents estacions de l'any i la forma de pagament.

Les característiques de salari i antiguitat es classifiquen en tres nivells, baix, mitja i alt i poca, mitjana i molta respectivament. Es vol veure, a l'1%, si són independents aquestes característiques. Pes això es fa una taula

Salari/Antiguitat	poca	mitjana	molta
baix	62	10	2
mitjà	14	38	12
alt	2	9	51

A partir dels totals de fila i columna es construeix la taula de freqüències teòriques

Salari/Antiguitat	poca	mitjana	molta
baix	28.86	21.09	24.05
mitjà	24.96	18.24	20.80
alt	24.18	17.67	21.15

El valor de $\aleph^2$ és 167,876 i supera el valor de la taula amb (3-1).(3-1)=4 graus de llibertat. S'ha de rebutjar la independència entre salaris i antiguitat

Taula Chi-Quadrat

	0,99	0,98	0,95	0,9	0,8	0,7	0,5	0,3	0,2	0,1	0,05	0,02	0,01	0,001
1	0,0002	0,0006	0,0039	0,0158	0,0642	0,1485	0,4549	1,0742	1,6424	2,7055	3,8415	5,4119	6,6349	10,827
2	0,0201	0,0404	0,1026	0,2107	0,4463	0,7133	1,3863	2,4079	3,2189	4,6052	5,9915	7,8241	9,2104	13,815
3	0,1148	0,1848	0,3518	0,5844	1,0052	1,4237	2,366	3,6649	4,6416	6,2514	7,8147	9,8374	11,345	16,266
4	0,2971	0,4294	0,7107	1,0636	1,6488	2,1947	3,3567	4,8784	5,9886	7,7794	9,4877	11,668	13,277	18,466
5	0,5543	0,7519	1,1455	1,6103	2,3425	2,9999	4,3515	6,0644	7,2893	9,2363	11,07	13,388	15,086	20,515
6	0,8721	1,1344	1,6354	2,2041	3,0701	3,8276	5,3481	7,2311	8,5581	10,645	12,592	15,033	16,812	22,457
7	1,239	1,5643	2,1673	2,8331	3,8223	4,6713	6,3458	8,3834	9,8032	12,017	14,067	16,622	18,475	24,321
8	1,6465	2,0325	2,7326	3,4895	4,5936	5,5274	7,3441	9,5245	11,03	13,362	15,507	18,168	20,09	26,124
9	2,0879	2,5324	3,3251	4,1682	5,3801	6,3933	8,3428	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	2,5582	3,0591	3,9403	4,8652	6,1791	7,2672	9,3418	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	3,0535	3,6087	4,5748	5,5778	6,9887	8,1479	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	3,5706	4,1783	5,226	6,3038	7,8073	9,0343	11,34	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	4,1069	4,7654	5,8919	7,0415	8,6339	9,9257	12,34	15,119	16,985	19,812	22,362	25,471	27,688	34,527
14	4,6604	5,3682	6,5706	7,7895	9,4673	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,124
15	5,2294	5,9849	7,2609	8,5468	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,698
16	5,8122	6,6142	7,9616	9,3122	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32	39,252
17	6,4077	7,255	8,6718	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,791
18	7,0149	7,9062	9,3904	10,865	12,857	14,44	17,338	20,601	22,76	25,989	28,869	32,346	34,805	42,312
19	7,6327	8,567	10,117	11,651	13,716	15,352	18,338	21,689	23,9	27,204	30,144	33,687	36,191	43,819
20	8,2604	9,2367	10,851	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,41	35,02	37,566	45,314
21	8,8972	9,9145	11,591	13,24	15,445	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932	46,796
22	9,5425	10,6	12,338	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289	48,268
23	10,196	11,293	13,091	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	38,968	41,638	49,728
24	10,856	11,992	13,848	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	40,27	42,98	51,179
25	11,524	12,697	14,611	16,473	18,94	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314	52,619
26	12,198	13,409	15,379	17,292	19,82	21,792	25,336	29,246	31,795	35,563	38,885	42,856	45,642	54,051
27	12,878	14,125	16,151	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	44,14	46,963	55,475
28	13,565	14,847	16,928	18,939	21,588	23,647	27,336	31,391	34,027	37,916	41,337	45,419	48,278	56,892
29	14,256	15,574	17,708	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	46,693	49,588	58,301
30	14,953	16,306	18,493	20,599	23,364	25,508	29,336	33,53	36,25	40,256	43,773	47,962	50,892	59,702
31	15,655	17,042	19,281	21,434	24,255	26,44	30,336	34,598	37,359	41,422	44,985	49,226	52,191	61,098
32	16,362	17,783	20,072	22,271	25,148	27,373	31,336	35,665	38,466	42,585	46,194	50,487	53,486	62,487
33	17,073	18,527	20,867	23,11	26,042	28,307	32,336	36,731	39,572	43,745	47,4	51,743	54,775	63,869
34	17,789	19,275	21,664	23,952	26,938	29,242	33,336	37,795	40,676	44,903	48,602	52,995	56,061	65,247
35	18,509	20,027	22,465	24,797	27,836	30,178	34,336	38,859	41,778	46,059	49,802	54,244	57,342	66,619
36	19,233	20,783	23,269	25,643	28,735	31,115	35,336	39,922	42,879	47,212	50,998	55,489	58,619	67,985
37	19,96	21,542	24,075	26,492	29,635	32,053	36,336	40,984	43,978	48,363	52,192	56,73	59,893	69,348
38	20,691	22,304	24,884	27,343	30,537	32,992	37,335	42,045	45,076	49,513	53,384	57,969	61,162	70,704
39	21,426	23,069	25,695	28,196	31,441	33,932	38,335	43,105	46,173	50,66	54,572	59,204	62,428	72,055
40	22,164	23,838	26,509	29,051	32,345	34,872	39,335	44,165	47,269	51,805	55,758	60,436	63,691	73,403

Anàlisi de la variancia

Les tècniques de comprovació de les diferents hipòtesi estadístiques, en particular les que pressuposen igualtat de mitjanes o de variances, s'han vist en el cas que hi hagi una comparació a fer entre dos mostres, o una mostra i una població. Algunes vegades ens interessa contrastar hipòtesis d'igualtat entre diferents mostres (més de dues), en aquest cas podem formular H_0 com
$H_0=\mu_0=\mu_1=\mu_2=...=\mu_n$
on 1,2,..n fan referència a les diferents mostres de la població.

En aquest cas les tècniques habituals reben el nom d'ANOVA (sigles d'anàlisi de la variancia) entre els diferents grups que podem formar.

ANOVA d'un factor

Considerem un factor en t nivells. Dels elements del nivell i en prenem una mostra de mida n_i; podem agrupar els elements de la forma

nivell	mostra
1	$x_{1,1}, x_{1,2},....x_{1,n}$
2	$x_{2,1}, x_{2,2},....x_{2,n}$
..	...

En el nivell 1 tenim n_1 elements, en el nivell 2 n_2,... En total N elements

Calculem les diferents mitjanes per nivells $\bar{x}_i = \frac{\sum_j x_{i,j}}{n_i}$ i mitjana general $\bar{x} = \frac{\sum_{i,j} x_{i,j}}{N}$

La quasi-variancia total ve definida per

$$S_T^2 = \frac{\sum_{ij} (x_{ij} - \bar{x})^2}{N-1}$$

amb N-1 graus de llibertat. La intervariacia (o variancia entre nivells) és

$$S_b^2 = \frac{\sum_i n_i (\bar{x}_i - \bar{x})^2}{t-1}$$

amb t-1 graus de llibertat. Per últim la intravariancia

$$S_w^2 = \frac{\sum_{ij} (x_{ij} - \bar{x}_i)^2}{N-t}$$

amb N-t graus de llibertat. Es verifica que la distribució de $\frac{S_b^2}{S_w^2}$ és una F d'Snedecor central amb (t-1) i (N-t) graus de llibertat.

Si el valor del quocient és superior al que donen les taules per una F d'aquests graus de llibertat a un nivell de significació de α, es rebutja la hipòtesi d'igualtat.

Càlculs simplificats

D'una manera semblant a la simplificació dels càlculs de la variancia en una distribució estadística d'un sol factor, en el cas de les variancies entre factors podem calcular més fàcilment aquests valors fent:

$$S_b^2 = \frac{\sum_i n_i (\bar{x}_i - \bar{x})^2}{t-1} = \frac{1}{t-1}\left[\sum_i \frac{\bar{x}_i^2}{n_i} - \frac{\bar{x}^2}{N}\right]$$

i

$$S_w^2 = \frac{\sum_{ij} (x_{ij} - \bar{x}_i)^2}{N-t} = \frac{1}{N-t}\left[\sum_{ij} x_{ij}^2 - \sum_i \frac{\bar{x}_i^2}{n_i}\right]$$

Podem fer servir la notació:

Suma de quadrats del factor	$SC(A) = \left[\sum_i \frac{\bar{x}_i^2}{n_i} - \frac{\bar{x}^2}{N}\right]$
Suma de quadrats corregida del error	$SC(E) = \left[\sum_{ij} x_{ij}^2 - \sum_i \frac{\bar{x}_i^2}{n_i}\right]$
Quadrat mitjà del factor	$CM(A) = S_b^2 = \frac{SC(A)}{t-1}$
Quadrat mitjà de l'error	$CM(E) = S_w^2 = \frac{SC(E)}{N-t}$

I la manera de distribuir els càlculs és:

	Graus de llibertat	Suma de quadrats	Quadrat mitjà	F
Total	N	$\sum_{ij} x_{ij}$		
Mitjana	1	$\frac{\bar{x}^2}{N}$		
Factor	t-1	SC(A)	CM(A)	
Error	N-t	SC(E)	CM(E)	$\frac{CM(A)}{CM(E)}$

Exemples

Imaginem que una determinada component pot fabricar-se de cinc maneres diferents. Un conjunt d'observacions sobre els cinc nivells de fabricació donen els resultats que figuren en la taula:

1	2	3	4	5
0.518	0.713	0.502	0.515	0.713
0.509	0.504	0.511	0.496	0.700
0.481	0.670	0.496	0.506	0.603
0.513	0.697	0.503	0.500	0.693
0.502	0.507		0.487	
	0.684		0.496	
			0.492	

Calculem la suma dels quadrats i la suma de tots els termes

$$\sum_{ij} x_{ij}^{\ 2} = 8{,}292841 \qquad \sum_{ij} x_{ij} = 14{,}511$$

I la suma de cada un dels termes de cada factor

1	2	3	4	5
$\sum x_1 = 2{,}523$	$\sum x_2 = 3{,}755$	$\sum x_3 = 2{,}012$	$\sum x_4 = 3{,}492$	$\sum x_5 = 2{,}709$

I ja podem calcular

$$\frac{\left(\sum_{ij} x_{ij}\right)^2}{N} = \frac{(14{,}511)^2}{26} = 8{,}098812$$

$$\sum_i \frac{\bar{x}_i^2}{n_i} = \frac{2{,}523^2}{5} + \frac{3{,}775^2}{6} + \ldots + \frac{2{,}709^2}{4} = 8{,}236925$$

$$SC(A) = \sum_i \frac{\bar{x}_i^2}{n_i} - \frac{\left(\sum_{ij} x_{ij}\right)^2}{N} = 8{,}236925 - 8{,}098812 = 0{,}138113$$

$$SC(E) = \sum_{ij} x_{ij}^{\ 2} - \sum_i \frac{\bar{x}_i^{\ 2}}{n_i} = 8{,}292841 - 8{,}236925 = 0{,}055916$$

La taula de càlcul simplificada és:

	Graus llibertat	Suma de quadrats	Quadrats mitjans	F
Total	26	8,292841		
Mitjana	1	8,098812		
Factor	4	0,138113	0,034528	
Error	21	0,055916	0,002663	12,966

El valor que dóna una taula F amb 4 i 21 graus de llibertat al nivell del 0,01 és 4,37. El valor calculat és clarament superior, aleshores podem afirmar que hi ha diferències en els cinc procediments.

Podem simplificar els càlculs quan s'equilibren les mostres en els factors, és a dir, quan la mida de les mostres és la mateixa: $n_1=n_2=...$

Considerem com exemple que volem comparar els nivells de renda de tres grups de famílies. Les dades són les que mostra la taula, on hi ha 8 observacions en cada factor

factor 1	1452	1372	1273	1147	1321	1083	1303	1245
factor 2	1266	1080	1091	1302	1514	1088	1218	1326
factor 3	1275	1417	1186	989	1449	1564	1109	1299

Obtenim

$$\sum_{ij} x_{ij} = 30369 \qquad \sum_{ij} x_{ij}^{\ 2} = 38946181$$

i de sumes per factor 10196, 9885 i 10288

Podem calcular

$$\frac{\left(\sum_{ij} x_{ij}\right)^2}{N} = \frac{(30369)^2}{24} = 38428173{,}375$$

$$\sum_i \frac{\bar{x}_i^2}{n_i} = \frac{10196^2}{8} + \frac{9885^2}{8} + \frac{10288^2}{8} = 38439323{,}13$$

$$SC(A) = \sum_i \frac{\bar{x}_i^2}{n_i} - \frac{\left(\sum_{ij} x_{ij}\right)^2}{N} = 38946181 - 38428173{,}375 = 506857{,}875$$

$$SC(E) = \sum_{ij} x_{ij}{}^2 - \sum_i \frac{\bar{x}_i{}^2}{n_i} = 38946181 - 38439323{,}13 = 506857{,}875$$

i formem la taula

	Graus llibertat	Suma de quadrats	Quadrats mitjans	F
Total	24	38946181		
Mitjana	1	38428173,375		
Factor	2	11149,750	5574,875	
Error	21	506857,875	24136,089	0,231

El valor que donen les taules per F amb 2 i 21 graus de llibertat al nivell del 5% és 3,47. Aleshores hem d'acceptar que no hi ha diferències significatives entre els nivells de renda de les famílies.

Model d'efectes aleatoris

Imaginen que els diferents factors estudiats no són tots els que formen la població. Aquesta té un conjunt d'infinits factors dels quals en podem analitzar una mostra aleatòria, però finita, d'aquests factors. En aquest cas és interessant verificar la hipòtesi d'uniformitat entre els diferents factors. Aquesta correspon a un model de $\sigma_A=0$; no existeix variancia entre els diferents factors.

Comparem-ho amb els dos exemples anteriors. Els mètodes de fabricació eren només cinc en el primer exemple, i hem considerat les famílies dividides en tres factors en el segon

En el model d'efectes aleatoris es fa servir com eina de contrast la mateixa distribució F d'Snedecor, però quan es rebutja $H_0=\sigma_A=0$ permet calcular variancia entre factors i dins del factors.

Exemples

Considerem que per comprovar la uniformitat en el procés d'embalatge d'una fàbrica es prenen mostres del pes d'un component a l'atzar d'un conjunt de capces a l'atzar. Tinguem present que la quantitat de capces es pot considerar infinita, d'elles es prenen només una mostra de sis. De cada capça, que conté diferents elements, es mesuren alguns. Els resultats són:

1	2	3	4	5	6
48 49	46 49 49	51 50 50 52 49	51 51 52 53	52 50 53	50 50 51 49

Ens permet formar la taula:

	Graus llibertat	Suma de quadrats	Quadrats mitjans	F
Total	21	53059		
Mitjana	1	53001.19		
Factor	5	36.69	7.338	
Error	15	21.12	1.408	5.212

El valor que dóna una taula F al 5% amb 5 i 15 graus de llibertat és 2,90. Hem de rebutjar que el procés sigui uniforme.

El càlcul de la variancia entre les capces és

$$S_b^2 + \frac{N^2 - \sum_i n_i^2}{N(t-1)} S_a^2 = CM(A)$$

d'on

$$1{,}408 + \frac{441-79}{21 \cdot 5} S_a^{\ 2} = 7{,}338$$

i

$$S_a^{\ 2} = 1{,}72$$

La variancia total serà

$$S_T^2 = S_a^2 + S_b^2 = 1{,}72 + 1{,}408 = 3{,}128$$

Podem acabar dient que és major la importància de la variancia entre les capces que la que existeix entre els elements que hi ha en elles.

Taula F-Snedecor

0,01 **F-Snedecor 1%**

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	50	100	1000
1	4052,2	4999,3	5403,5	5624,3	5764,0	5859,0	5928,3	5981,0	6022,4	6055,9	6083,4	6106,7	6125,8	6143,0	6157,0	6170,0	6181,2	6191,4	6200,7	6208,7	6239,9	6302,3	6333,9	6362,8
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,41	99,42	99,42	99,43	99,43	99,44	99,44	99,44	99,45	99,45	99,46	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13	27,05	26,98	26,92	26,87	26,83	26,79	26,75	26,72	26,69	26,58	26,35	26,24	26,14
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,45	14,37	14,31	14,25	14,20	14,15	14,11	14,08	14,05	14,02	13,91	13,69	13,58	13,47
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,96	9,89	9,82	9,77	9,72	9,68	9,64	9,61	9,58	9,55	9,45	9,24	9,13	9,03
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	7,66	7,60	7,56	7,52	7,48	7,45	7,42	7,40	7,30	7,09	6,99	6,89
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47	6,41	6,36	6,31	6,28	6,24	6,21	6,18	6,16	6,06	5,86	5,75	5,66
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67	5,61	5,56	5,52	5,48	5,44	5,41	5,38	5,36	5,26	5,07	4,96	4,87
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11	5,05	5,01	4,96	4,92	4,89	4,86	4,83	4,81	4,71	4,52	4,41	4,32
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71	4,65	4,60	4,56	4,52	4,49	4,46	4,43	4,41	4,31	4,12	4,01	3,92
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40	4,34	4,29	4,25	4,21	4,18	4,15	4,12	4,10	4,01	3,81	3,71	3,61
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16	4,10	4,05	4,01	3,97	3,94	3,91	3,88	3,86	3,76	3,57	3,47	3,37
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96	3,91	3,86	3,82	3,78	3,75	3,72	3,69	3,66	3,57	3,38	3,27	3,18
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80	3,75	3,70	3,66	3,62	3,59	3,56	3,53	3,51	3,41	3,22	3,11	3,02
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	3,61	3,56	3,52	3,49	3,45	3,42	3,40	3,37	3,28	3,08	2,98	2,88
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,62	3,55	3,50	3,45	3,41	3,37	3,34	3,31	3,28	3,26	3,16	2,97	2,86	2,76
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,46	3,40	3,35	3,31	3,27	3,24	3,21	3,19	3,16	3,07	2,87	2,76	2,66
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,43	3,37	3,32	3,27	3,23	3,19	3,16	3,13	3,10	3,08	2,98	2,78	2,68	2,58
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30	3,24	3,19	3,15	3,12	3,08	3,05	3,03	3,00	2,91	2,71	2,60	2,50
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,29	3,23	3,18	3,13	3,09	3,05	3,02	2,99	2,96	2,94	2,84	2,64	2,54	2,43
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	3,06	2,99	2,94	2,89	2,85	2,81	2,78	2,75	2,72	2,70	2,60	2,40	2,29	2,18
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70	2,63	2,56	2,51	2,46	2,42	2,38	2,35	2,32	2,29	2,27	2,17	1,95	1,82	1,70
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50	2,43	2,37	2,31	2,27	2,22	2,19	2,15	2,12	2,09	2,07	1,97	1,74	1,60	1,45
1000	6,66	4,63	3,80	3,34	3,04	2,82	2,66	2,53	2,43	2,34	2,27	2,20	2,15	2,10	2,06	2,02	1,98	1,95	1,92	1,90	1,79	1,54	1,38	1,16

0,1 **F-Snedecor 10%**

n_1

n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	50	100	1000
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	60,47	60,71	60,90	61,07	61,22	61,35	61,46	61,57	61,66	61,74	62,05	62,69	63,01	63,30
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,40	9,41	9,41	9,42	9,42	9,43	9,43	9,44	9,44	9,44	9,45	9,47	9,48	9,49
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,22	5,21	5,20	5,20	5,20	5,19	5,19	5,19	5,18	5,17	5,15	5,14	5,13
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,91	3,90	3,89	3,88	3,87	3,86	3,86	3,85	3,85	3,84	3,83	3,80	3,78	3,76
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,28	3,27	3,26	3,25	3,24	3,23	3,22	3,22	3,21	3,21	3,19	3,15	3,13	3,11
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,92	2,90	2,89	2,88	2,87	2,86	2,85	2,85	2,84	2,84	2,81	2,77	2,75	2,72
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,68	2,67	2,65	2,64	2,63	2,62	2,61	2,61	2,60	2,59	2,57	2,52	2,50	2,47
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,52	2,50	2,49	2,48	2,46	2,45	2,45	2,44	2,43	2,42	2,40	2,35	2,32	2,30
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,40	2,38	2,36	2,35	2,34	2,33	2,32	2,31	2,30	2,30	2,27	2,22	2,19	2,16
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,30	2,28	2,27	2,26	2,24	2,23	2,22	2,22	2,21	2,20	2,17	2,12	2,09	2,06
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,23	2,21	2,19	2,18	2,17	2,16	2,15	2,14	2,13	2,12	2,10	2,04	2,01	1,98
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,17	2,15	2,13	2,12	2,10	2,09	2,08	2,08	2,07	2,06	2,03	1,97	1,94	1,91
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,12	2,10	2,08	2,07	2,05	2,04	2,03	2,02	2,01	2,01	1,98	1,92	1,88	1,85
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,07	2,05	2,04	2,02	2,01	2,00	1,99	1,98	1,97	1,96	1,93	1,87	1,83	1,80
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,04	2,02	2,00	1,99	1,97	1,96	1,95	1,94	1,93	1,92	1,89	1,83	1,79	1,76
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	2,01	1,99	1,97	1,95	1,94	1,93	1,92	1,91	1,90	1,89	1,86	1,79	1,76	1,72
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,98	1,96	1,94	1,93	1,91	1,90	1,89	1,88	1,87	1,86	1,83	1,76	1,73	1,69
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,95	1,93	1,92	1,90	1,89	1,87	1,86	1,85	1,84	1,84	1,80	1,74	1,70	1,66
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,93	1,91	1,89	1,88	1,86	1,85	1,84	1,83	1,82	1,81	1,78	1,71	1,67	1,64
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,91	1,89	1,87	1,86	1,84	1,83	1,82	1,81	1,80	1,79	1,76	1,69	1,65	1,61
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,84	1,82	1,80	1,79	1,77	1,76	1,75	1,74	1,73	1,72	1,68	1,61	1,56	1,52
50	2,81	2,41	2,20	2,06	1,97	1,90	1,84	1,80	1,76	1,73	1,70	1,68	1,66	1,64	1,63	1,61	1,60	1,59	1,58	1,57	1,53	1,44	1,39	1,33
100	2,76	2,36	2,14	2,00	1,91	1,83	1,78	1,73	1,69	1,66	1,64	1,61	1,59	1,57	1,56	1,54	1,53	1,52	1,50	1,49	1,45	1,35	1,29	1,22
1000	2,71	2,31	2,09	1,95	1,85	1,78	1,72	1,68	1,64	1,61	1,58	1,55	1,53	1,51	1,49	1,48	1,46	1,45	1,44	1,43	1,38	1,27	1,20	1,08

0,05 **F-Snedecor 5%**

n_1

n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	50	100	1000
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	242,98	243,90	244,69	245,36	245,95	246,47	246,92	247,32	247,69	248,02	249,26	251,77	253,04	254,19
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40	19,41	19,42	19,42	19,43	19,43	19,44	19,44	19,44	19,45	19,46	19,48	19,49	19,49
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,73	8,71	8,70	8,69	8,68	8,67	8,67	8,66	8,63	8,58	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91	5,89	5,87	5,86	5,84	5,83	5,82	5,81	5,80	5,77	5,70	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,66	4,64	4,62	4,60	4,59	4,58	4,57	4,56	4,52	4,44	4,41	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,98	3,96	3,94	3,92	3,91	3,90	3,88	3,87	3,83	3,75	3,71	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,55	3,53	3,51	3,49	3,48	3,47	3,46	3,44	3,40	3,32	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,26	3,24	3,22	3,20	3,19	3,17	3,16	3,15	3,11	3,02	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,05	3,03	3,01	2,99	2,97	2,96	2,95	2,94	2,89	2,80	2,76	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,89	2,86	2,85	2,83	2,81	2,80	2,79	2,77	2,73	2,64	2,59	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,76	2,74	2,72	2,70	2,69	2,67	2,66	2,65	2,60	2,51	2,46	2,41
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,66	2,64	2,62	2,60	2,58	2,57	2,56	2,54	2,50	2,40	2,35	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,58	2,55	2,53	2,51	2,50	2,48	2,47	2,46	2,41	2,31	2,26	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53	2,51	2,48	2,46	2,44	2,43	2,41	2,40	2,39	2,34	2,24	2,19	2,14
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,45	2,42	2,40	2,38	2,37	2,35	2,34	2,33	2,28	2,18	2,12	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42	2,40	2,37	2,35	2,33	2,32	2,30	2,29	2,28	2,23	2,12	2,07	2,02
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,35	2,33	2,31	2,29	2,27	2,26	2,24	2,23	2,18	2,08	2,02	1,97
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,31	2,29	2,27	2,25	2,23	2,22	2,20	2,19	2,14	2,04	1,98	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,28	2,26	2,23	2,21	2,20	2,18	2,17	2,16	2,11	2,00	1,94	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,25	2,22	2,20	2,18	2,17	2,15	2,14	2,12	2,07	1,97	1,91	1,85
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16	2,14	2,11	2,09	2,07	2,05	2,04	2,02	2,01	1,96	1,84	1,78	1,72
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,99	1,95	1,92	1,89	1,87	1,85	1,83	1,81	1,80	1,78	1,73	1,60	1,52	1,45
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,89	1,85	1,82	1,79	1,77	1,75	1,73	1,71	1,69	1,68	1,62	1,48	1,39	1,30
1000	3,85	3,00	2,61	2,38	2,22	2,11	2,02	1,95	1,89	1,84	1,80	1,76	1,73	1,70	1,68	1,65	1,63	1,61	1,60	1,58	1,52	1,36	1,26	1,11

Regressió

En una variable estadística bidimensional es pot sospitar que existeix alguna mena de relació entre les dues variables que la formen. En aquest cas cal estudiar una mesura de la relació de dependència que pot existir. Seria el cas de considerar el pes i l'alçada d'un grup de persones, o les qualificacions en dues assignatures,..

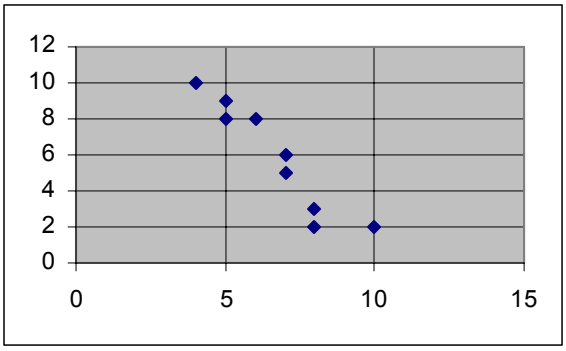
La funció que podem imaginar que lliga les dues variables pot ser de diverses maneres. De totes elles una funció lineal és la més senzilla d'estudiar. Imaginem que si x i y són les dues variables podem ajustar el comportament d'una respecte de l'altra en forma de funció lineal y=ax+b on a i b són constants que cal determinar. La recta es coneix amb el nom de recta de regressió o recta de mínims quadrats.

Prèviament convé veure si una recta dóna o no un bon ajust del comportament de les variables. Resulta fonamental el càlcul de coeficient de correlació lineal entre elles. Valors propers a ±1 donen previsiblement un bon nivell d'ajust, en forma directa o inversament proporcional. Cal observar, tot i així, que valors de correlació no significatius no indiquen que no existeixi relació entre les variables, només que aquesta no és lineal. De la mateixa manera pot resultar que variables sense cap mena de relació tinguin coeficients de correlació alts.

El càlcul ha de ser sempre posterior a una representació gràfica de les variables. Un núvol de punts ja dóna una acceptable idea de si existeix o no relació lineal entre les variables.

Exemples

Un núvol de punts com el del gràfic indica clarament un nivell força acceptable de correlació inversa entre les variables



Si les dades (x,y) són les que s'indiquen, una taula de càlcul ha de ser com ara:

x	y	f	$f.x^2$	$f.y^2$	f.x.y
8	3	1	64	9	24
10	2	1	100	4	20
7	5	1	49	25	35
5	9	1	25	81	45
6	8	1	36	64	48
8	2	1	64	4	16
4	10	1	16	100	40
5	8	1	25	64	40
7	6	1	49	36	42
60	53	9	428	387	310

D'on obtenim un coeficient de correlació r=-0,9463, força significatiu. Observem que és negatiu, indica una relació inversa.

Els càlculs són

$$s_x = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} = \sqrt{\frac{428}{9} - \left(\frac{60}{9}\right)^2} = 1{,}763$$

$$s_y = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2} = \sqrt{\frac{387}{9} - \left(\frac{53}{9}\right)^2} = 2{,}885$$

i la covariança

$$s_{xy} = \frac{\sum xy}{n} - \bar{x} \cdot \bar{y} = \frac{310}{9} - \left(\frac{60}{9} \cdot \frac{53}{9}\right) = -4{,}81$$

El problema més interessant són les prediccions que poden fer-se sobre una determinada variable si en coneixem l'altra. A aquest efecte convé tenir present que les rectes que permeten aquests càlculs són diferents. Només en el cas d'un coeficient de correlació r=1 o r=-1 podem considerar la mateixa recta.
De la taula anterior obtenim

$$\bar{x} = 6{,}67 \quad \bar{y} = 5{,}89 \quad s_x^2 = 3{,}11 \quad s_y^2 = 8{,}32 \quad s_{x,y} = -4{,}81$$

Si ara volem fer prediccions sobre un valor de y que ha de correspondre a un x determinat ens cal fer servir la recta de regressió de y sobre x

$$y - 5{,}89 = \frac{-4{,}81}{3{,}11}(x - 6{,}67)$$

diferent de l'equació que s'hauria de fer servir si, conegut un valor de y, volem estimar un valor de x

$$x - 6{,}67 = \frac{-4{,}81}{8{,}32}(y - 5{,}89)$$

Coeficient de correlació

Habitualment es considera el coeficient de correlació de Pearson com

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

És un indicador del nivell de correlació lineal entre les variables. Si aquestes estan lligades per altre tipus de relació (que no sigui lineal) el coeficient de Pearson pot no resultar significatiu. Els valors extrems 1 i –1 indiquem una relació lineal perfecte (directa o inversa). Un coeficient 0 indica que no hi ha relació lineal entre les variables.

De ρ^2 s'anomena variancia residual i s'interpreta com el percentatge de casos que poden explicar-se a partir d'una relació lineal. Valors de $\rho \geq 0{,}7$ és consideren significatius.

Un coeficient de correlació no paramètric és el coeficient de correlació d'Spearman que es fonamenta amb les diferències d'ordre en les variables, es calcula

$$c = 1 - \frac{6 \cdot \sum D_i^2}{n(n^2 - 1)}$$

n són els parells de dades i D_i les diferències entre el rang que té assignat cada element en una variable i l'altra. Els valors extrems són els mateixos que en cas de la correlació de Pearson, si bé no es deixa influenciar per valors extrems.

Exemples

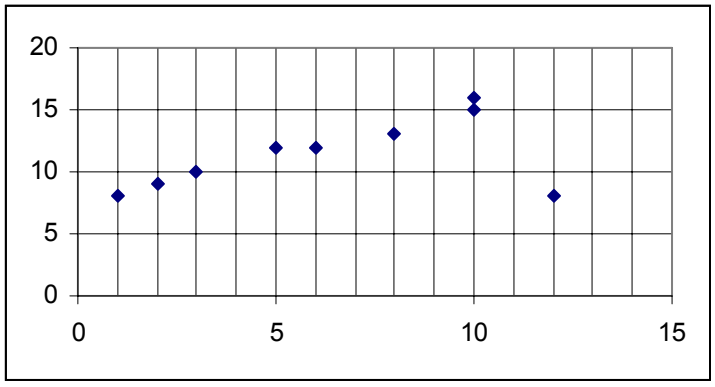
Si les observacions de nou dades són les de la taula, formem les ordenacions d'una i altra variable, calculem el quadrat de la diferència d'ordenació i obtenim

x	y	Rx	Ry	D	D^2
6	12	5	4,5	0,5	0,25
10	15	7,5	7	0,5	0,25
8	13	6	6	0	0
10	16	7,5	8	-0,5	0,25
1	8	1	1	0	0
5	12	4	4,5	-0,5	0,25
3	10	3	3	0	0
2	9	2	2	0	0
12	18	9	9	0	0
					1

$$c = 1 - \frac{6 \cdot \sum D_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 1}{9(81-1)} = 0{,}925$$

Al formar les columnes Rx i Ry cal tenir present que les dades repetides impliquen una modificació en la ordenació. Així, per exemple, en la sèrie x la setena i vuitena posició estan ocupades per x=10. Associem a aquest valor el punt mitjà de les ordenacions 7 i 8.

La representació de les dades és



Fent servir les mateixes dades, el càlcul del coeficient de Pearson és

x	y	x^2	y^2	x.y
6	12	36	144	72
10	15	100	225	150
8	13	64	169	104
10	16	100	256	160
1	8	1	64	8
5	12	25	144	60
3	10	9	100	30
2	9	4	81	18
12	8	144	64	96
57	103	483	1247	698

$$s_x = \sqrt{\frac{483}{9} - \left(\frac{57}{9}\right)^2} = 3{,}68$$

$$s_y = \sqrt{\frac{1247}{9} - \left(\frac{103}{9}\right)^2} = 2{,}75$$

$$s_{xy} = \frac{698}{9} - \left(\frac{57}{9}\right)\left(\frac{103}{9}\right) = 5{,}07$$

i el coeficient de correlació

$$r = \frac{s_{xy}}{s_x s_y} = 0{,}5$$

diferent del d'Spearman calculat anteriorment.

Error en les estimacions

L'error tipus o error estàndard en les estimacions lineals és defineix com

$$S^2_{y,x} = \frac{\sum (Y - Y_s)^2}{N} = \frac{\sum Y^2 - b\sum Y - m\sum XY}{N}$$

en el cas de ser x la variable independent i y la variable dependent. Y_s és el valor de les imatges estimades en la recta de regressió, m el pendent de la recta i b la imatge de zero. D'una manera semblant en el cas de utilitzar la recta de regressió de x sobre y

Rectes paral·leles a la recta de regressió a distàncies $S_{y,x}$ agrupen el 68% dels valors de la distribució bidimensional, si la distància és $2.S_{y,x}$ el 95%, i així successivament. Quan N és petit cal una correcció

$$\hat{S}_{y,x} = \sqrt{\frac{N}{N-2}} \cdot S_{y,x}$$

La recta de regressió de y sobre x de la primera taula era

$$y - 5{,}89 = \frac{-4{,}81}{3{,}11}(x - 6{,}67)$$

de pendent m=-1,547 , b=16,206, N=9, $\sum xy = 310$, $\sum y = 53$ i $\sum y^2 = 387$. Amb la correcció pertinent l'error estàndard en les estimacions de la recta és

$$\hat{S}_{y,x} = \sqrt{\frac{N}{N-2} \cdot \frac{\sum y^2 - m\sum xy - b\sum y}{N}} = 0{,}975$$

Anàlisi del coeficient de correlació lineal

Un problema de correlació lineal té sentit quan existeixen entre les variables una correlació de tipus lineal prou significatiu. Si les relacions no són lineals (estan basades sobre altre tipus de funcions, com ara logarítmiques o exponencials o polinòmiques,..) o quan el coeficient de correlació no sigui prou significatiu (proper a 1 en valor absolut), no té sentit cap mena de predicció dels tipus lineal. Tampoc en aquell cas on la variable independent excedeixi el rang de la mostra. Com exemple d'aquest últim cas, si un dieta x origina una pèrdua de pes y no podem estimar pèrdues de pes sobre dietes de valors extrems dels mesurats.

Si es vol contrastar la hipòtesi H_0 de r=0 (no hi ha relació lineal significativa) sobre una hipòtesi alternativa $r \neq 0$ en una mostra de mida n es fa servir l'estadístic

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}$$

que segueix una t d'Student de n-2 graus de llibertat.

Si es vol contrastat la hipòtesi de r=k es pot fer servir el fet que la variable $z = \frac{1}{2}\ln\frac{1+r}{1-r}$ segueix una normal de mitjana i variància

$$\left(\frac{1}{2}\ln\frac{1+r}{1-r} + \frac{r}{2(n-1)}; \frac{1}{n-3}\right)$$

fixant d'aquesta manera intervals de valors on cal cercar r a un determinat nivell.

En prediccions basades sobre una recta de regressió que dóna per a $x=x_0$ un valor estimat $f(x_0)=y_0$, un interval centrat en y_0 del veritable valor de y té d'error típic $s_y\sqrt{1-r^2}$. Sobre aquests valors, i fent servir una distribució normal, l'interval

$$y_0 \pm 1{,}96 \cdot s_y\sqrt{1-r^2}$$

té probabilitat 0,95 de contenir el veritable valor de $f(x_0)$

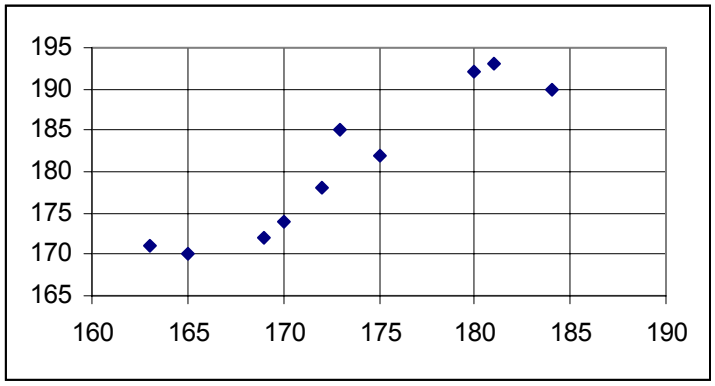
Exemples

En la taula x indica l'altura del pare i y indica l'altura del fill. El càlcul del coeficient de correlació i la recta de regressió de y sobre x és

x	y	x^2	y^2	x.y
165	170	27225	28900	28050
170	174	28900	30276	29580
172	178	29584	31684	30616
184	190	33856	36100	34960
169	172	28561	29584	29068
163	171	26569	29241	27873
175	182	30625	33124	31850
181	193	32761	37249	34933
173	185	29929	34225	32005
180	192	32400	36864	34560
1732	1807	300410	327247	313495

$s_x = 6{,}54 \quad s_y = 8{,}49 \quad s_{xy} = 52{,}26$

el coeficient de correlació lineal és r=0,94. Indica l'existència d'una forta correlació lineal entre les variables, una representació gràfica ho manifesta



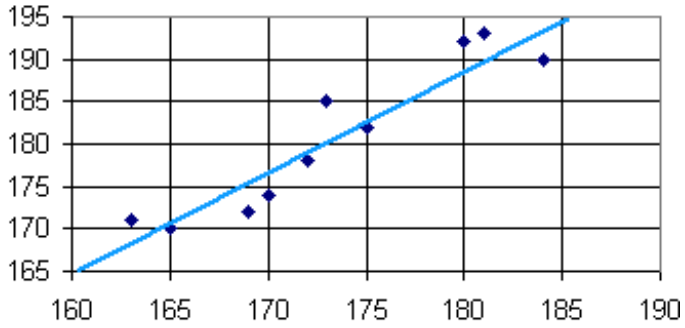
La recta de regressió de y sobre x és

$$y - 180{,}7 = \frac{52{,}26}{42{,}77}(x - 173{,}2)$$

en forma explícita

$$y = 1{,}22x - 31$$

La recta de regressió sobre el gràfic és



Segons la recta de regressió, a un pare d'altura 190 cm correspon una altura de 200,8 en el cas del fill. Un interval amb probabilitat 0,9 d'altura esperada del fill serà

$$\bar{x} \pm \lambda_{0,95} s_y \sqrt{1-r^2} = 200,8 \pm 1,64 \cdot 8,49\sqrt{1-0,94^2}$$

l'altura esperada està en l'interval 200,8±4,8

Si en una mostra on n=16 i el càlcul del coeficient de correlació lineal és r=0,8917 volem contrastar la hipòtesi que no hi ha correlació, calculem $t = \sqrt{16-2}\,\dfrac{0,8917}{\sqrt{1-0,8917^2}} = 7,3745$. Consultada una taula t d'Student, amb 14 graus de llibertat i a un nivell 0,001 tenim t=4,14. Amb aquestes dades cal rebutjar la hipòtesi de correlació nul·la i acceptar que en aquestes dades existeix una correlació lineal prou significativa.

La desviació típica de la variable y és 12,6. Per a un valor de x la imatge que dóna la recta de regressió calculada és y_0=54.3 i r=0,918. Amb una probabilitat del 0.95 la veritable imatge estarà dins de l'interval

$$\left(54,3 \pm 1,96 \cdot 12.6\sqrt{1-0,918^2}\right) = \left(54,3 \pm 9.8\right)$$

Estadística no paramètrica

Al contrastar diferents hipòtesis podem fer servir paràmetres de la població o de la mostra, com ara la mitjana i la desviació tipus. En aquest cas el contrast que fem és paramètric ja que utilitzem aquests valors de diferents paràmetres de la distribució. Quan no fem servir aquests paràmetres diem que utilitzem un contrast no paramètric.
El contrast no paramètric s'ha de fer servir quan la distribució no és normal, o hi ha sospites que no ho sigui, o les informacions sobre les variables venen en forma de rang. En aquest casos hi ha diferents mètodes per contrastar hipòtesis.

Prova dels signes

En una mostra ordenada formada per un conjunt de parells

$$(x_1, y_1)\ (x_2, y_2)(x_3, y_3)...(x_n, y_n)$$

si la distribució de X-Y és simètrica respecte de l'origen es verificarà

$$P(X > Y) = \frac{1}{2}$$

i podem contrastar-ho fent servir una distribució binomial on p=q=1/2. Resulta aplicable quan n>10. En valors superiors de n podem fer servir una aproximació per una normal.
Esperem, aleshores, que les vegades que x sigui superior a y no difereixin significativament de les vegades que y superi x. Associem un signe positiu o negatiu i aquests han de distribuir-se amb

probabilitat 1/2. Com efectes pràctics de càlcul es prescindeix dels parell on x=y, aquells que tindrien un signe 0.

Exemples

Dos crítics han ordenat 10 pintures de l'1 al 10 segons la qualitat que en elles observen. Es vol saber si hi ha diferències significatives en la valoració dels dos crítics

Pintures	1	2	3	4	5	6	7	8	9	10
Crític 1	1	4	2	7	5	10	8	3	9	6
Crític 2	2	5	1	9	6	10	7	3	8	4

La sèrie de signes de la valoració del crític 1 menys la valoració del crític 2 és

-	-	+	-	-	0	+	0	+	+

Prescindim de les pintures 6 i 8 que han estat valorades en la mateixa posició per tots dos cítics, en la sèrie dels 8 signes restants hi ha 4 positius i 4 negatius, aquest és el valor esperat en una distribució binomial n=8 i p=1/2. Aleshores no hi ha motiu per rebutjar la hipòtesi nul·la: Els dos crítics valoren igual.

Un assaig sobre un determinat medicament aplicat a 32 malalts pretén determinar si aquest origina o no un augment de pes. En la sèrie de les diferències de pes en els 32 malalts s'observa que en 18 el pes ha augmentat, en 4 no ha canviat, i en 12 el pes ha disminuït.

Prescindim dels 4 malalts que no han variat de pes, els 28 restants es consideren una distribució binomial on p=q=1/2. Si aproximem per una normal els paràmetres són

$$\mu = 14 \quad \sigma = \sqrt{28 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 2{,}65$$

A un nivell del 0,05 hem d'acceptar valors tipificats que difereixin de la mitjana fins a 1,96. A un nivell del 0,01 els valors poden diferir fins 1,64. Cap d'aquests valors es supera al tipificar els signes positius que ofereixen les diferències de pes

$$z = \frac{18 - 14}{2{,}65} = 1{,}51$$

Cal acceptar que no hi ha diferències de pes que puguin considerar-se significatives.

Una fàbrica d'ordinadors vol provar un nou model de teclat. Per fer-ho demana a 16 mecanògrafes que escriguin un text en un teclat convencional i en el nou model. Els resultat són els que figuren en la taula, on la tercera de les files és el signe de les diferències de pulsacions que s'obtenen en el nou model respecte de l'anterior. Es vol veure si hi ha diferències significatives amb una confiança del 70%

210	205	227	300	286	297	311	331	302	175	177	203	214	258	231	325
219	210	225	303	291	285	312	340	303	176	179	203	215	256	230	326
+	+	-	+	+	-	+	+	+	+	+	0	+	-	-	+

Aproximem una B(15,1/2) a una normal de paràmetres

$$\mu = 7{,}5 \;\; \sigma = \sqrt{15 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 1{,}94$$

En una normal, un interval centrat en la mitjana de probabilitat 0,7 té un valor tipificat de z=1,04. Els extrems d'acceptació són

$$\mu \pm z\sigma = 7{,}5 \pm 1{,}04 \cdot 1{,}94 = \begin{cases} 9{,}51 \\ 5{,}48 \end{cases}$$

En valor 11 (o 4 si pensem en els negatius) queda fora d'aquest interval. Els resultats del segon teclat són superiors als del primer.

Prova de les ratxes

Una ratxa és una successió de fets determinats sense variació. En una successió d'esdeveniments aleatoris la presència de ratxes ha de distribuir-se també aleatòriament. Si no és així podem sospitar que hi ha algun factor que trenca aquesta característica aleatòria i que fa que hi hagi una determinada tendència.

Les proves basades en ratxes són especialment indicades per esbrinar si hi ha alguna component estacional en sèries de valors, una certa persistència cronològica, o si les dades procedeixen d'una mateixa població.

En la successió d'esdeveniments associem signe + o signe - segons els valors siguin superiors o inferiors a un valor de la distribució (habitualment la mitjana o la mediana), anomenem n_1 i n_2 a aquests valors. En la successió de signes determinem la quantitat de ratxes que tenen lloc, aquesta quantitat és R i fem servir que es distribueix normalment amb paràmetres

$$\mu = \frac{2n_1n_2}{n_1+n_2}+1 \quad \sigma = \sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

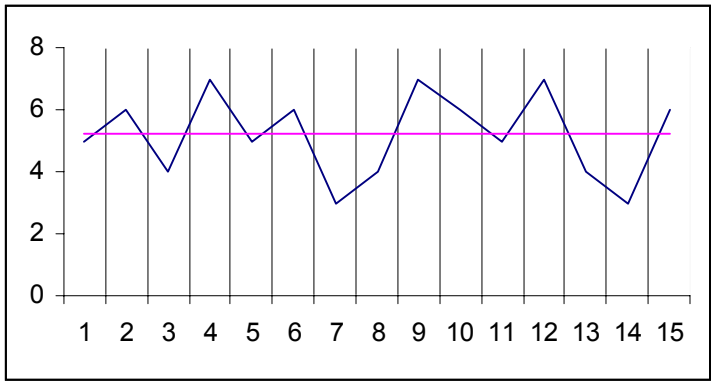
Veiem primer alguns exemples per tal d'entendre el significat de ratxa. Si la successió de signes és com:

+	+	+	-	+	-	-	-	-	+
1			2	3	4				5

S'han produït les 5 ratxes indicades, la primera els tres signes +, al canviar a signe - comença la segona de les ratxes, que dura només una dada, igual que la tercera. La quarta ratxa esta formada per quatre signes - i la cinquena i última de les ratxes el darrer signe +.
En aquest cas $n_1=n_2=5$ ja que hi ha 5 signes + i 5 signes -

Exemples

Durant quinze mesos consecutius una empresa ha estudiat els milers d'unitats venudes i els ha comparat amb la mitjana de vendes d'aquest període. Vol estudiar, amb un nivell del 90%, si les diferències de vendes respecte de la mitjana segueixen responen a l'atzar o si es deuen a components estacionaris. El gràfic de les dades és



Formem una taula amb els signes i les ratxes

-	+	-	+	-	+	-	-	+	+	-	+	-	-	+
1	2	3	4	5	6	7		8		9	10	11		12

D'on R=12, $n_1=7$ i $n_2=8$. Es produeixen 8 ratxes en 7 signes + i 8 signes -
Ajustem una distribució normal de paràmetres

$$\mu = \frac{2 \cdot 7 \cdot 8}{7+8} + 1 = 8{,}47$$

$$\sigma = \sqrt{\frac{2 \cdot 7 \cdot 8(2 \cdot 7 \cdot 8 - 7 - 8)}{(7+8)^2 (7+8-1)}} = 1{,}85$$

Si tipifiquem R obtenim

$$z = \frac{12 - 8{,}47}{1.85} = 1{,}91$$

Aquest valor és superior al valor que donen les taules d'una N(0,1) de 1,64. Hem d'acceptar que les diferències no són aleatòries

Un professor vol veure si l'ordre de lliurament d'unes proves ve determinat per la qualificació en aquestes proves. Per això corregeix les proves segons l'ordre i escriu en elles + o - si la qualificació és positiva o negativa. La sèrie de valors que anota és

+	+	-	-	-	-	+	+	-	+	+	+

Observa 7 signes +, 5 signes - i 5 ratxes. A un nivell del 0,05 ho ha de contrastar amb una normal de paràmetres

$$\mu = \frac{2 \cdot 7 \cdot 5}{12} + 1 = 6{,}83$$

$$\sigma = \sqrt{\frac{2 \cdot 7 \cdot 5(2 \cdot 7 \cdot 5 - 7 - 5)}{(7+5)^2 (7+5-1)}} = 1{,}6$$

El valor tipificat por diferir fins ± 1,96. En aquest cas és

$$z = \frac{5 - 6{,}83}{1{,}6} = 1{,}143$$

d'on s'ha d'acceptar que les ratxes es deuen a l'atzar i no hi ha relació en l'ordre i la qualificació.

Prova dels rangs de signes de Wilcoxon

Fins ara s'han considerat els signes de les diferències entre parells de valors i s'ha prescindit de la seva magnitud. D'aquesta manera hi ha una pèrdua d'informació ja que una diferència positiva pot originar-se de dos parells de valors diferents, però semblants, o bé de dos valors que siguin molt diferents. En un exemple anterior s'han observat diferències de rapidesa en dues mecanògrafes sobre dos teclats d'ordinador diferents, però no s'ha considerat el valor d'aquestes diferències. En el primer parell de dades de valors 210 i 219 pulsacions el signe seria també positiu en el cas que les diferències fossin majors, 210 i 270 o més properes com 210 i 211

La prova de Wilcoxon assigna a cada parell de n dades un rang des de 1 fins n. El rang 1 correspon a la diferència absoluta més petita, el rang n el de la diferència absoluta major. Quan s'han establert aquests rangs es considera el signe d'aquesta diferència fent que aquest sigui positiu o negatiu, tenim aleshores uns rangs amb signe. La suma dels rangs positius es denota com T_+ i s'espera que aquest no difereixi significativament de la suma dels rangs negatius quan formulem una hipòtesi d'igualtat.

Si considerem uns parells de resultats (x,y) les diferències, els rangs i els rangs-signes són:

x	y	diferències	rangs	rang-signe
4	7	-3	3,5	-3,5
3	6	-3	3,5	-3,5
7	2	5	5	5
6	4	2	2	2
6	7	-1	1	-1

Observem que la diferència més petita entre els parells de valors correspon al parell (6,7), ocupa el rang 1 (el primer) en la columna de rangs, però la diferència és negativa. El parell (7,2) és aquell un s'observa una màxima diferència, de 5 unitats i positiva. Veiem que els dos primers parells de la taula tenen la mateixa diferència i han d'ocupar els llocs tercer i quart dels rangs. Considerem la mitjana entre aquests llocs i en cada un d'ells el rang serà 3,5. En aquesta taula serà $T_+=7$.

Els valors estan tabulats quan n<10. En valors superiors s'utilitza una aproximació normal de mitjana

$$\mu = \frac{n(n+1)}{4}$$

i variància

$$s^2 = \frac{n(n+1)(2n+1)}{24}$$

Exemple

Un grup de 20 estudiants ha fet dues proves A i B amb els resultats que figuren en la taula. Es calculen les diferències A-B, el rang i el rang-signe

x	y	diferència	rang	rang-signe
45	35	10	11,5	11,5
36	40	-4	4	-4
61	58	3	2	2
45	58	-13	15	-15
59	65	-6	7	-7
58	48	10	11,5	11,5
65	70	-5	6	-6
65	81	-16	18	-18
54	62	-8	9,5	-9,5
59	48	11	13	13
28	35	-7	8	-8
87	66	21	20	20
36	48	-12	14	-14
67	69	-2	1	-1
58	72	-14	16	-16
68	76	-8	9,5	-9,5
59	74	-15	17	-17
75	57	18	19	19
68	72	-4	4	-4
66	62	4	4	4

Observem les tres diferències 4 en valors absolut. Han d'ocupar els llocs tercer, quart i cinquè en el rang. Les associem al rang 4, que és la mitjana dels rangs 3,4 i 5. Si observem les dues diferències 8 en valor absolut, han d'ocupar els llocs novè i desè, les fem correspondre a un rang 9,5

La suma de les diferències positives és $T_1=81$ i n=20. El contrast s'ha de fer sobre una normal de paràmetres

$$\mu = \frac{20 \cdot 21}{4} = 105$$

$$s = \sqrt{\frac{20 \cdot 21 \cdot 41}{24}} = 26{,}79$$

El valor T_+ tipificat és

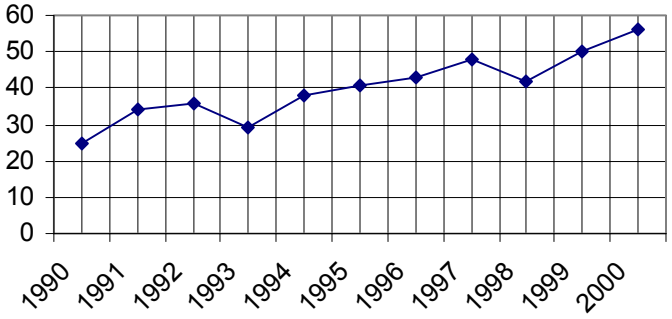
$$z = \frac{81 - 105}{26{,}79} = -0{,}896$$

que està dins de l'interval d'acceptació a un nivell 0,05. S'accepta H_0 i no hi ha diferències en les dues proves.

Sèries temporals

Quan les dades estadístiques s'obtenen cada cert temps i es té present l'instant on s'ha fet aquesta observació, formen sèries temporals. Un conjunt de valors que indiquen les despeses familiars en els dotze mesos de l'any quan s'associen a cada mes constitueixen una sèrie. Les vendes d'una botiga al llarg dels dies del mes, o de les hores del dia,.. també formen una sèrie temporal.

Quan cal una representació gràfica de sèries temporals es reserva l'eix OX per les unitats de temps, així tendències com augment o disminució i cicles estacionals són fàcilment observables. Generalment s'utilitza un polígon de freqüències. Per exemple en aquest gràfic de vendes d'un determinat article al llarg d'uns anys

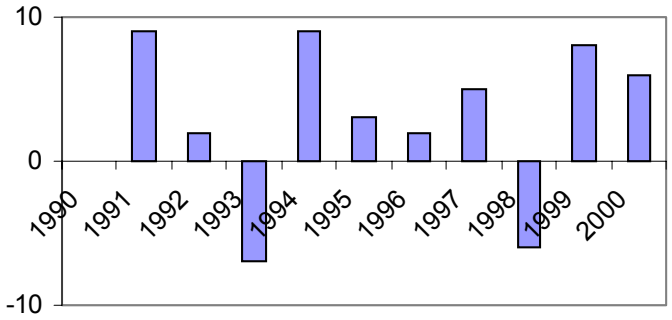


Nombres índex

En les sèries temporals podem indicar i representar les magnituds però moltes vegades és més convenient representar la variació d'aquesta magnitud al llarg dels períodes. Aquesta variació pot ser absoluta, indicant amb signe si la variació és positiva (es produeix un augment) o negativa (una disminució). Amb les dades anteriors les variacions absolutes són

any	vendes	variació
1990	25	0
1991	34	9
1992	36	2
1993	29	-7
1994	38	9
1995	41	3
1996	43	2
1997	48	5
1998	42	-6
1999	50	8
2000	56	6

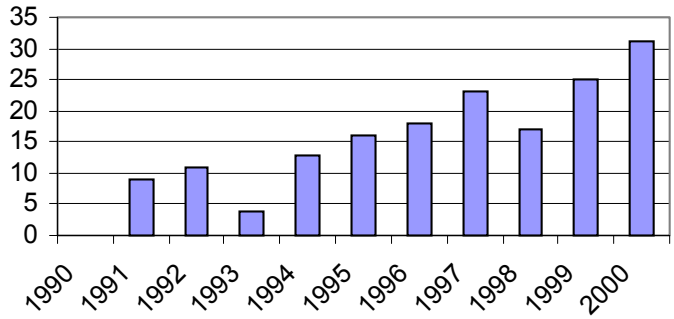
que poden representar-se sobre un gràfic de barres amb signes com ara



Aquestes variacions s'han calculat totes sobre la dada anterior; és la variació de les vendes de l'any n comparades amb les vendes de l'any n-1. Definir clarament la base del càlcul de les variacions és molt important de cara a estudiar les sèries temporals. No és el mateix fer que la base sigui l'any anterior que fer que la base sigui, per exemple, el primer any de la sèrie. Si es calculessin d'aquesta manera, les noves variacions formen una taula com ara

any	vendes	variació	variació 1990
1990	25	0	0
1991	34	9	9
1992	36	2	11
1993	29	-7	4
1994	38	9	13
1995	41	3	16
1996	43	2	18
1997	48	5	23
1998	42	-6	17
1999	50	8	25
2000	56	6	31

i donarà lloc a una gràfica molt més "optimista".



Quan es calculen les variacions en forma de percentatge sobre un valor base, s'obtenen els anomenats nombres índex. Aquests constitueixen la manera habitual de presentar estudis com ara els índex de preus al consum (IPC) o els índexs de borsa. Com s'ha indicar abans la base i la manera de calcular aquests nombres índex és fonamental de cara a interpretar els resultats de les sèries temporals. Habitualment es pren com base el període temporal anterior i es fixa aquest com valor unitari. En les dades que ens serveixen d'exemple els índexs seran

any	vendes	variació %	variació 1990
1990	25	0,00	0,00
1991	34	36,00	36,00
1992	36	5,88	44,00
1993	29	-19,44	16,00
1994	38	31,03	52,00

1995	41	7,89	64,00
1996	43	4,88	72,00
1997	48	11,63	92,00
1998	42	-12,50	68,00
1999	50	19,05	100,00
2000	56	12,00	124,00

La tercera columna calculada sobre els valors anteriors, la quarta columna sobre el primer any. Observem la diferència de valors segons el canvi de la base.

Previsió

L'anàlisi de sèries temporals té sovint una finalitat de predicció: Si les vendes d'un determinat any són conegudes, es poden saber les de l'any vinent? Si l'índex de preus s'ha comportat d'una determinada manera, es por preveure el seu comportament futur? Si unes determinades accions de borsa han augmentat de valor fins ara, ¿seguiran augmentant?...

Aquestes preguntes tenen totes una difícil resposta ja que impliquen un coneixement d'una situació futura que no es controlable. De totes maneres hi ha mètodes de càlcul que volen inferir uns resultats futurs en sèries de temps coneguts els valors anteriors. Tots ells s'han de considerar amb moltes reserves.

Els mètodes de regressió, en particular el mètode més simple de regressió lineal no és aplicable, ja que aquest pressuposa que els errors són independents. En les sèries de temps aquests errors independents és molt qüestionable.

El model de predicció més simple és aquell que dóna al període n la previsió del valor en el període n-1. Imagina que el proper mes, el proper any,.. les dades seran com les del mes anterior, l'any anterior,.. Si aquest model fos vàlid no hi hauria variació en les sèries temporals.

Un segon mètode és coneix amb el nom de mitjanes mòbils. Pronostica el valor en el període n com una mitjana de k valors anteriors. Respon a la suposició que els darrers valors són els que més poden influir en un valor futur. Per a a_{n+1} la previsió de mitjanes mòbils k serà

$$a_n = \frac{a_{n-k+1} + \ldots + a_{n-1} + a_n}{k}$$

Aquest mètode accepta una modificació que consisteix en atribuir un pes determinat a cada un dels valors anteriors formant així una mitjana ponderada. Normalment l'últim valor és aquell a qui s'associa un pes major ja que es considera que un valor futur ve més influenciat per l'anterior. Al ser una mitjana ponderada s'ha de verificar que la suma de tots els pesos sigui 1. La previsió de mitjanes mòbils k ponderades p_i serà ara

$$a_n = \frac{p_k a_{n-k+1} + \ldots + p_2 a_{n-1} + p_1 a_n}{k}$$

on

$$\sum p_k = 1$$