Teorema del límit central

Carles Rovira Escofet

P08/05057/02306

UOC
Universitat Oberta de Catalunya
www.uoc.edu

Índex

La distribució de la mitjana mostral

En aquesta sessió estudiarem el comportament de la mitjana mostral d'una variable. Per exemple, suposem que volem estudiar la mitjana de l'alçada dels estudiants de la UOC: n'hem seleccionat una mostra a l'atzar, els hem mesurat i hem calculat la mitjana de les alçades dels estudiants de la mostra; ara volem veure com es comporta aquesta mitjana mostral.

Veurem que si sabem que la variable que s'estudia és normal, aleshores la mitjana mostral també és normal però amb desviació típica més petita. I també veurem que si la variable no és normal però la mostra és prou gran, la mitjana també serà aproximadament normal.

1. Distribució de la mitjana mostral per a variables normals

Suposem que tenim una mostra $x_1, ..., x_n$ d'una variable aleatòria normal. Recordem que la mitjana es defineix com:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Aquesta mitjana depèn de la mostra. Usualment tindrem només una mostra, però en podríem prendre moltes de diferents, de manera que a cadascuna li correspondria una mitjana diferent. Això ens dóna peu a parlar de la distribució mostral de la mitjana. Per a indicar que es tracta d'una variable aleatòria, la denotarem per $\overline{X}$.

Observeu que...

... per a una col·lecció de mostres, tindrem la corresponent col·lecció de mitjanes mostrals $\bar{x}_1, ..., \bar{x}_k$.

Per a estudiar-la, haurem de distingir dos casos: quan la desviació típica de la variable que mesurem és coneguda i quan és desconeguda.

1.1. Cas de desviació típica poblacional coneguda

Podem pensar en l'exemple de les alçades dels estudiants de la UOC. Suposem que en un estudi anterior s'havia demostrat que les alçades dels estudiants de la UOC seguien una distribució normal de mitjana 172 cm i desviació típica 11 cm.

Desviació poblacional i desviació mostral

La desviació poblacional és la desviació real de la variable, que en aquest cas suposem coneguda. Quan calculem la desviació a partir de mostres, parlem de *desviació mostral*.

Intuïtivament veiem que la mitjana de les observacions de la mostra que tenim ha de ser un valor proper a 172. També sembla raonable pensar que observacions més grans que la mitjana poblacional, 172, es compensaran amb valors més petits, i que com més gran sigui la mostra, més proper serà el valor de la mitjana mostral a 172.

Pensem ara que tenim una mostra de 100 estudiants de la UOC, fem 10 grups de 10 estudiants i per a cadascun d'aquests grups fem la mitjana aritmètica. Obtenim 10 valors, corresponents a les 10 mitjanes $\bar{x}_1$,..., $\bar{x}_{10}$. Sembla raonable pensar que la mitjana d'aquestes noves dades sigui també 172. D'altra banda, també sembla raonable pensar que aquests nous valors siguin més propers a 172 que les dades originals, ja que en cadascuna de les mitjanes se'ns hauran compensat valors grans amb valors petits.

Si la variable que estudiem segueix una distribució normal amb mitjana μ i desviació típica σ conegudes, aleshores la mitjana mostral és també normal amb la mateixa mitjana μ i desviació típica $\frac{\sigma}{\sqrt{n}}$, on n és la mida de la mostra. Per tant, tipifiquem la variable $\overline{X}$ i obtenim que:

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

segueix una distribució normal estàndard.

Demostració

La demostració d'aquest resultat és conseqüència d'una important propietat de les variables aleatòries normals. La propietat és la següent: si X i Y són variables aleatòries independents amb lleis

$$N(\mu_1, \sigma_1^2) \text{ i } N(\mu_2, \sigma_2^2),$$

respectivament, aleshores $X + Y$ té una llei:

$$N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

En el nostre exemple, la variable que recull totes les possibles mitjanes de cada grup de 10 estudiants segueix una distribució normal de mitjana 172 cm i desviació típica $11 / \sqrt{10} = 3{,}48$ cm. Observem que, efectivament, com més gran és la mostra, més petita és la desviació típica i, per tant, hi ha menys dispersió.

Aquest quocient que ens dóna la desviació típica de la mitjana mostral es coneix com *error estàndard*.

Si σ és la desviació típica de la població i n la mida de la mostra, es defineix l'**error estàndard de la mitjana mostral** com:

$$\frac{\sigma}{\sqrt{n}}$$

Observeu que...

... l'error estàndard és cada cop més petit com més gran és la mida de la mostra.

Exemple d'error estàndard d'una mitjana mostral

Considerem les alçades dels estudiants de la UOC. Suposem que sabem que es tracta d'una variable aleatòria normal de mitjana 172 cm i desviació típica 11 cm i que hem pres una mostra de 300 estudiants a l'atzar. Aleshores podem contestar preguntes del tipus següent:

a) Quina és la probabilitat que la mitjana sigui menor que 170 cm?

La distribució de la mitjana mostral és normal de mitjana 172 cm i desviació típica:

$$\frac{11}{\sqrt{300}} = 0{,}635$$

Tipifiquem la variable per tal d'obtenir una normal (0,1). Hem de calcular:

$$P(\overline{X} < 170) = P\left(\frac{\overline{X} - 172}{0{,}635} < \frac{-2}{0{,}635}\right) = P(Z < -3{,}149) = 0{,}0008$$

ja que Z és una variable aleatòria normal (0,1).

b) Quina és la probabilitat que la distància entre la mitjana mostral (d'aquesta mostra de 300 estudiants) i la mitjana poblacional, 172 cm, sigui més petita que 1 cm?

Per un raonament semblant (si la distància entre dos números a i b ha de ser més petita que k, s'ha de complir: $|a - b| < k$):

$$P(|\overline{X} - \mu| < 1) = P(-1 < \overline{X} - \mu < 1) = P\left(-\frac{1}{0{,}635} < \frac{\overline{X} - \mu}{0{,}635} < \frac{1}{0{,}635}\right) = P(-1{,}57 < Z < 1{,}57)$$

on Z és una variable aleatòria normal (0,1). Si busquem a les taules de la llei normal (0,1), veiem que aquesta probabilitat és igual a 0,8836.

Tenim així una probabilitat del 0,8836 d'obtenir un valor per a la mitjana mostral que difereixi en menys d'1 cm del valor real de la mitjana quan prenem una mostra de 300 individus.

Observem que enlloc no hem utilitzat que la mitjana fos exactament 172 cm. És a dir, si sabem que la variable "alçada" segueix una normal amb una desviació típica d'11 cm i prenem una mostra de 300 estudiants, sabem que la diferència entre la seva mitjana i la mitjana poblacional μ (que pot ser que no coneguem) serà menor d'1 cm amb una probabilitat del 0,8836.

c) Considerem ara el problema invers. Suposem que desconeixem la mitjana μ de l'alçada dels estudiants de la UOC i volem estudiar una mostra de manera que la diferència entre la mitjana de la mostra i la de la població μ sigui menor que 1 cm amb una probabilitat del 0,95. De quina mida ha de ser la nostra mostra?

Sabem que la variable estadística tipificada:

$$\frac{\overline{X} - \mu}{\frac{11}{\sqrt{n}}}$$

es distribueix com una normal (0,1). D'altra banda, si mirem les taules observem que si Z és una normal (0,1):

$$P(-1{,}96 < Z < 1{,}96) = 0{,}95$$

Per tant:

$$0{,}95 = P\left(-1{,}96 < \frac{\overline{X} - \mu}{\frac{11}{\sqrt{n}}} < 1{,}96\right) = P\left(-1{,}96\ \frac{11}{\sqrt{n}} < \overline{X} - \mu < 1{,}96\ \frac{11}{\sqrt{n}}\right)$$

I si imposem que la diferència $\overline{X} - \mu$ ha de ser menor que 1 cm, obtenim:

$$1{,}96\ \frac{11}{\sqrt{n}} < 1$$

Per tant, $\sqrt{n} > 11 \cdot 1{,}96$, i així: $n > (11 \cdot 1{,}96)^2 = 464{,}8$. Aleshores, si prenem 465 individus per a fer l'estudi, sabem que la diferència entre la mitjana mostral que obtindrem i la mitjana real, serà menor d'1 cm amb una probabilitat del 0,95. Fixeu-vos que com més gran sigui la mida de la mostra, més petita serà la diferència entre la mitjana mostral i la poblacional.

Si es multipliquen el numerador i el denominador per n, podem escriure el resultat que hem vist en aquest apartat d'una altra manera.

Si la variable que estudiem segueix una distribució normal amb mitjana μ i desviació típica σ coneguda, aleshores:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$$

segueix una distribució normal estàndard.

1.2. Cas de desviació típica poblacional desconeguda. La *t* de Student

Fixem-nos que en els exemples estudiats anteriorment necessitàvem dues coses:

- que la variable que s'estudiava fos normal
- que el valor de la desviació típica de la variable fos conegut

Les variables aleatòries normals són habituals

En molts casos és habitual suposar que una variable aleatòria és normal. Alguns exemples són: el pes o l'alçada de les persones, l'error que cometen els aparells de mesura, el pes de la fruita, les vendes setmanals d'una botiga, etc.

Aquests dos fets es coneixen gràcies a estudis previs. Sovint, aquest estudi no s'ha fet, però podem fer la suposició que la variable és normal. En aquest cas, haurem de fer una estimació de la desviació típica amb l'anomenada **desviació típica mostral**:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Observeu que...

... en el cas de la desviació típica mostral, es divideix per $n - 1$, no pas per n.

de manera que en els càlculs de l'apartat anterior reemplaçarem la σ per la s. Llavors, la distribució mostral de la mitjana ja no és una distribució normal com passava quan en lloc de s coneixíem el veritable valor σ de la desviació.

Diversos estudis fets per W.S. Gosset al final del segle XIX van demostrar que en aquest cas s'obté una distribució diferent de la normal, encara que per a mides prou grans s'assemblen força. Aquesta nova distribució es coneix amb el nom de t de Student amb $n - 1$ graus de llibertat. Això vol dir que per cada mida de la mostra, n, tenim, en realitat, una distribució diferent.

W.S. Gosset

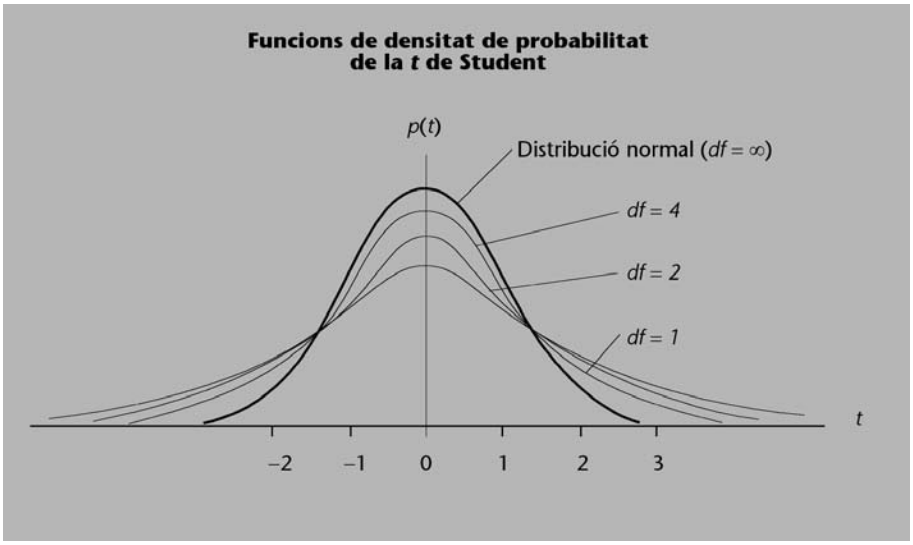
W.S. Gosset treballava a l'empresa cervesera Guinness i utilitzava el pseudònim de *Student* per a signar els seus treballs.

La **distribució *t* de Student amb *n* graus de llibertat**, que denotarem per t_n, és molt semblant a la distribució normal (0,1): és simètrica al voltant del zero, però la seva desviació típica és una mica més gran que la de la normal (0,1), és a dir, els valors que pren aquesta variable estan una mica més dispersos. Això no obstant, com més gran és el nombre de graus de llibertat, n, més s'aproxima la distribució t_n de Student a la distribució normal (0,1). Considerarem que podem aproximar la t_n per una normal estàndard per a $n > 100$.

El valor real i la distribució t_n de Student

Observeu que quan coneixem el valor veritable de σ, la variable $\bar{X}$ segueix sempre una distribució normal, però la seva variància depèn de n.

El gràfic següent representa les funcions de densitat de la *t* de Student per a diferents valors de *n* i, en línia més gruixuda, la densitat d'una distribució normal (0,1).



Si σ és desconeguda i *n* és la mida de la mostra, calcularem l'error estàndard mitjançant el quocient:

L'error estàndard és més petit com més gran és la mida de la mostra.

$$\text{Error estàndard} = \frac{s}{\sqrt{n}}$$

Aquest error estàndard ens permet d'obtenir un resultat nou important.

Si la variable que estudiem segueix una distribució normal amb mitjana μ i desviació típica desconeguda, aleshores:

$$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

segueix una distribució t_{n-1}, és a dir, una *t* de Student amb $n - 1$ graus de llibertat.

Òbviament, la manera més fàcil de calcular probabilitats relacionades amb una *t* de Student és amb qualsevol programari estadístic o, fins i tot, un full de càlcul. De totes maneres, com en el cas de la normal, comentarem com podem fer servir unes taules estadístiques.

Les taules que ens donen la distribució de la *t* de Student són semblants a les de la distribució normal estàndard. Això no obstant, i atès que per a cada valor dels graus de llibertat tenim una distribució diferent, les taules habituals només ens serveixen per a vuit probabilitats determinades (per a altres valors cal

utilitzar algun programari apropiat). La manera d'emprar les taules és la següent: cerquem a la primera columna el nombre de graus de llibertat, ens situem en aquella fila i determinem quins punts ens deixen la probabilitat acumulada que ens interessa.

Exemple d'utilització de les taules de la *t* de Student

Una empresa indica en un paquet d'arròs que el pes mitjà del paquet és de 900 grams. En una inspecció, hem analitzat el pes en grams de 10 paquets d'arròs, i hem obtingut les dades següents:

890	901	893	893	896
895	894	895	904	899

a) Quina és la probabilitat que la distància entre la mitjana poblacional i la mitjana mostral sigui més gran de 3 grams?

És raonable pensar que el pes en grams d'un paquet d'arròs és una variable aleatòria normal amb mitjana el pes que indica el paquet, i amb una desviació típica determinada. És a dir, de mitjana, els paquets haurien de tenir 900 grams, però a causa dels errors de mesura dels aparells que omplen els paquets, alguns en contindran una mica més de 900 grams i d'altres, una mica menys. Suposem, doncs, que la variable d'interès (el pes del paquet) és normal, però no sabem res de la seva desviació típica. Amb les nostres dades podem estimar la desviació típica i obtenim:

$$s = 4{,}19$$

Aleshores, podem utilitzar el fet que $(\bar{x} - \mu) / (s/\sqrt{n})$ és una observació d'una *t* de Student amb $n - 1$ graus de llibertat (en el nostre exemple, com que tenim 10 dades, serà una *t* de Student amb 9 graus de llibertat). Podem ara calcular:

$$P(|\overline{X} - \mu| > 3) = 1 - P(-3 < \overline{X} - \mu < 3) = 1 - P\left(-\frac{3}{\frac{4{,}19}{\sqrt{10}}} < \frac{\overline{X} - \mu}{\frac{4{,}19}{\sqrt{10}}} < \frac{3}{\frac{4{,}19}{\sqrt{10}}}\right) =$$

$$= 1 - P(-2{,}26 < t_9 < 2{,}26)$$

on ja sabem que t_9 és una *t* de Student amb 9 graus de llibertat. Podem calcular aquesta probabilitat a les taules:

$$P(-2{,}26 < t_9 < 2{,}26) = 1 - 2P(t_9 \geq 2{,}26) = 1 - 2 \cdot 0{,}025 = 0{,}95$$

Aleshores:

$$1 - P(-2{,}26 < t_9 < 2{,}26) = 1 - 0{,}95 = 0{,}05$$

Per tant, a partir d'aquestes dades, tot sembla indicar que l'empresa enganya els seus clients. En efecte, si es pren una mostra de mida 10, la probabilitat que la diferència entre la mitjana mostral i la real sigui més gran de només 3 grams és d'un 5%. En canvi, la mitjana de la nostra mostra és de 896 grams, 4 grams menys que la quantitat que indica el paquet.

En aquest cas, els valors que ens han sortit ens han permès d'utilitzar les taules. Altres vegades necessitarem utilitzar l'ordinador.

2. Resum

En aquesta sessió hem estudiat la distribució de la mitjana de dades que provenen d'una distribució normal, i n'hem diferenciat dos casos: quan la variància poblacional és coneguda i quan la variància és desconeguda. Per a estudiar aquest darrer cas, hem hagut d'introduir la distribució *t* de Student.

Exercicis

1. La despesa mensual de la família mexicana Robles segueix una distribució normal de mitjana 3.000 pesos i variància 500. Suposem que la despesa de cada mes és independent de la dels altres mesos. Si l'ingrés anual és de 37.000 pesos, quina és la probabilitat que no gastin més del que guanyen? Quant haurien de guanyar per a tenir una seguretat del 99% que no gastaran més del que han guanyat?

2. Hem fet una enquesta entre els homes d'una població determinada i, a partir dels resultats, deduïm que el pes dels homes d'aquesta població segueix una distribució normal de mitjana 72 kg. Per tal de saber si les dades que hem obtingut són fiables, pesem 4 dels enquestats i obtenim una mitjana de 77,57 kg amb una desviació típica de 3,5 kg. Tenim prou motius per a pensar que els enquestats han mentit quan ens han dit el seu pes?

Solucionari

1. Anomenem X_A la despesa anual. Com que la despesa mensual X_M segueix una llei normal de mitjana 3.000 i desviació típica $\sqrt{500}$ i:

$$12 \cdot 3.000 = 36.000 \text{ i } \sqrt{12 \cdot 500} = 77,4597$$

sabem que $\frac{X_A - 36.000}{77,4597}$ segueix una distribució normal estàndard.

Per tant, la probabilitat que la família Robles gasti menys de 37.000 pesos és:

$$P(X_A < 37.000) = P\left(\frac{X_A - 36.000}{77,4597} < \frac{37.000 - 36.000}{77,4597}\right) = P(Z < 12,9099)$$

on Z és una distribució normal estàndard. Si mirem les taules de la distribució normal estàndard observem que la probabilitat que sigui més petita que 3 ja és 1. Per tant, la probabilitat és 1, és a dir, podem assegurar amb gairebé un 100% de certesa que no gastaran més del que guanyen.

Per a respondre la segona pregunta, hem de trobar una quantitat G tal que:

$$P(X_A < G) = P\left(\frac{X_A - 36.000}{77,4597} < \frac{G - 36.000}{77,4597}\right) = 0,99$$

Si mirem les taules de la normal, observem que la quantitat:

$$\frac{G - 36.000}{77,4597}$$

hauria de ser igual a 2,33, i per tant, si resolem l'equació següent:

$$\frac{G - 36.000}{77,4597} = 2,33$$

obtenim que cal que G = 36.180,4811 per a tenir una seguretat del 99% que aquesta família no gastarà més del que guanya.

2. Observem que la diferència entre la mitjana de les nostres dades i el valor poblacional és de 5,57. Calcularem la probabilitat que, si escollim 4 dels enquestats a l'atzar, la mitjana del pes d'aquests individus difereixi en 5,57 kg o més de la mitjana que coneixem de la població. Per tant, hem de calcular:

$$P(|\overline{X} - \mu| \geq 5,57)$$

Si aquesta probabilitat fos petita, ens indicaria que els enquestats segurament han mentit sobre el seu pes. Amb l'ajut de les taules, calculem la probabilitat del complementari:

$$P(|\overline{X} - \mu| < 5,57) = P(-5,57 < \overline{X} - \mu < 5,57) = P\left(-\frac{5,57}{\frac{3,5}{\sqrt{4}}} < \frac{\overline{X} - \mu}{\frac{3,5}{\sqrt{4}}} < \frac{5,57}{\frac{3,5}{\sqrt{4}}}\right) =$$

$$= P(-3,18 < t_3 < 3,18) = 1 - 2P(t_3 \geq 3,18) = 1 - 0,05 = 0,95$$

on t_3 és una t de Student amb 3 graus de llibertat. Hem d'utilitzar la t de Student perquè sabem que la variable d'interès segueix una distribució normal, però en desconeixem la desviació típica (només tenim la desviació típica de la mostra). Per tant:

$$P(|\overline{X} - \mu| \geq 5,57) = 1 - P(|\overline{X} - \mu| < 5,57) = 0,05$$

Així, doncs, sembla que ens han mentit, ja que la probabilitat que la diferència entre les mitjanes dels pesos que ens han dit i 72 és molt petita, de l'ordre de 0,05.

Observeu que podem fer tots aquests càlculs amb les taules de la t de Student.

El teorema del límit central

La distribució de la mitjana mostral d'una població normal és una distribució normal amb la mateixa mitjana poblacional i amb desviació típica, l'error estàndard. Aquest fet ens permet de calcular probabilitats quan tenim una mostra d'una variable amb distribució normal i desviació típica coneguda. Quan no coneixem la desviació típica de la variable, també podem fer càlculs amb la distribució *t* de Student.

En aquesta sessió veurem com hem de procedir quan no sabem si la variable d'interès segueix una distribució normal o no, o quan sabem del cert que la seva distribució no és normal.

Quan la mostra és prou gran, la solució ens ve donada per un dels resultats fonamentals de l'estadística: el teorema del límit central. Ho introduirem amb un cas particular: l'estudi de la binomial.

1. Aproximació de la binomial a la normal

Suposem que juguem diàriament a un número d'una loteria que, entre altres premis, retorna l'import jugat a tots els números que acaben en la mateixa xifra que el número guanyador.

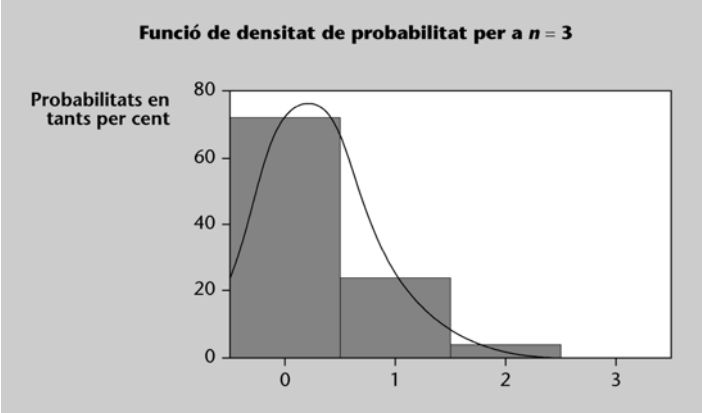
Considerem la variable $X(n)$ que ens dóna el nombre de cops que ens han tornat l'import jugat quan s'han realitzat n sortejos. En aquest cas, sabem que la variable aleatòria $X(n)$ segueix una distribució binomial de paràmetres n i $p = 0{,}1$. En efecte, s'han fet n sortejos (és a dir, s'ha repetit un mateix experiment n cops de manera independent) i en cada sorteig la probabilitat que ens tornin els diners és $p = 1/10 = 0{,}1$ (probabilitat d'èxit). Observem, però, què passa en augmentar el valor de n amb la funció de densitat de probabilitat de la variable $X(n)$. Si dibuixem aquesta funció de densitat de probabilitat per a $n = 3$, obtenim la gràfica següent:

Binomial

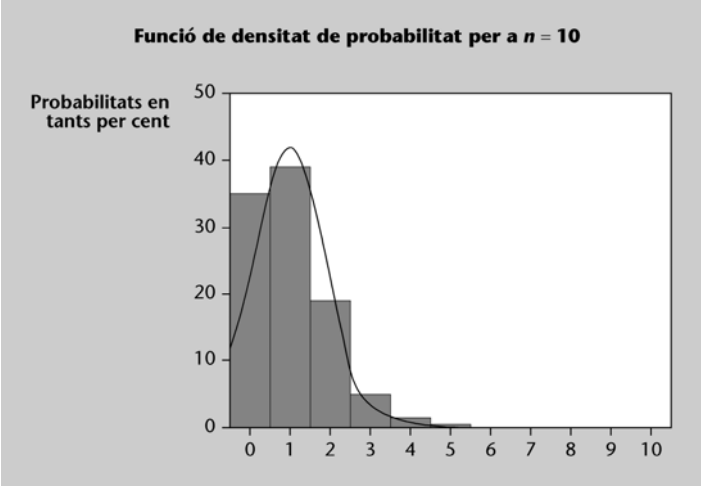
Si X segueix una distribució binomial de paràmetres n i p, aleshores:

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

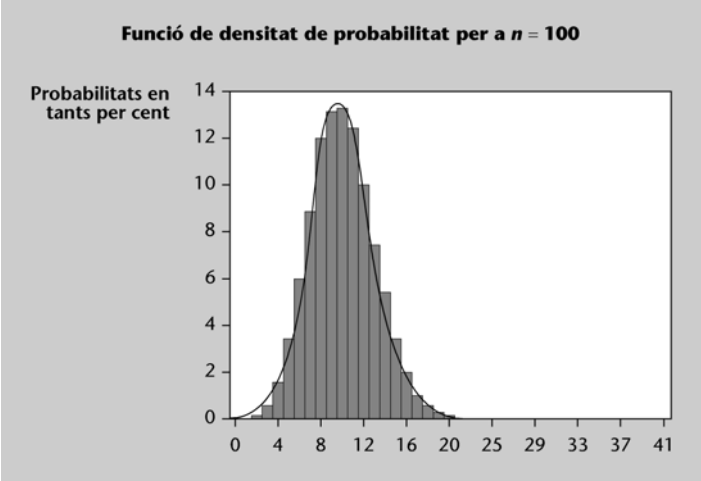
per als $k \in \{0, \ldots, n\}$.



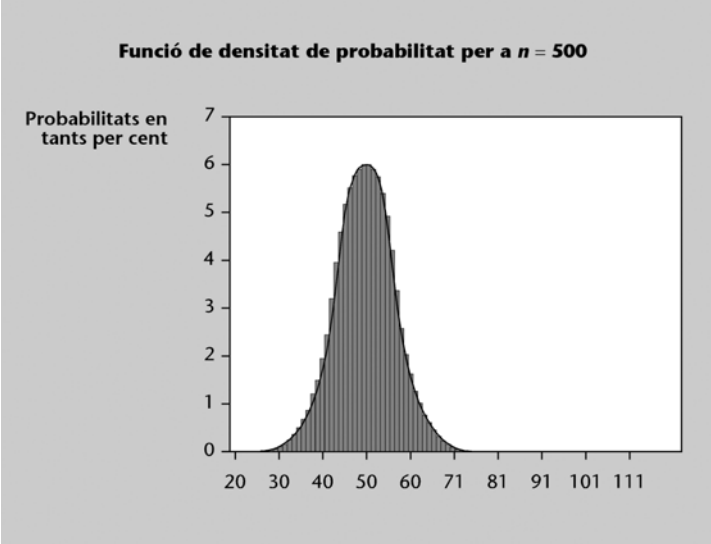
Si ara considerem $n = 10$, els possibles valors van del 0 al 10, i el gràfic de la funció de densitat de probabilitat és:



Si prenem $n = 100$, el gràfic és:



I si per exemple prenem $n = 500$, el gràfic de la funció de probabilitat és:



Veiem, doncs, que el perfil d'aquest gràfic cada cop s'assembla més al de la funció de densitat de probabilitat d'una variable aleatòria normal. La conclusió que traiem d'aquest experiment és que si n és prou gran, la variable aleatòria $X(n)$ és aproximadament normal. Determinarem ara la mitjana i la desviació d'aquesta variable aleatòria, que seran els corresponents a la mateixa $X(n)$:

- L'esperança d'aquesta variable és:

$$n \cdot p = 0{,}1 \cdot n$$

- i la variància:

$$np(1 - p) = n(0{,}1) \cdot (0{,}9) = 0{,}09n$$

Aquests seran els paràmetres de la variable aleatòria normal que aproxima la distribució de $X(n)$. Així, doncs, si n és prou gran, $X(n)$ es comporta com una $N(0{,}1n; 0{,}09n)$.

Sigui X una variable aleatòria amb distribució binomial de paràmetres n i p. Si n és gran, aleshores la distribució de X és aproximadament normal amb esperança $\mu = np$ i variància $\sigma^2 = np(1 - p)$. A la pràctica s'acostuma a utilitzar aquesta aproximació quan np i $n(1 - p)$ són tots dos més grans que 5, o bé quan $n > 30$.

Aquest resultat ens permet de simplificar força els càlculs en algunes situacions.

Exemple de la loteria

Quina és la probabilitat aproximada que en un any ens hagin tornat els diners almenys 50 vegades? De fet, hem de calcular la probabilitat $P(X(365) \geq 50)$. Si volguéssim obtenir el valor exacte d'aquesta probabilitat, pel fet que $X(365)$ és una binomial de paràmetres 365 i $p = 0{,}1$, hauríem de fer el càlcul següent:

$$P(X(365) \geq 50) = 1 - P(X(365) < 50) =$$
$$= 1 - P(X(365) = 0) - P(X(365) = 1) - P(X(365) = 2) - \ldots - P(X(365) = 49)$$

on cadascuna d'aquestes probabilitats es trobaria mitjançant la fórmula de la binomial que ja coneixem, en el nostre cas:

$$P(X(365) = k) = \binom{365}{k}(0{,}1)^k(0{,}9)^{365-k}$$

En canvi, si renunciem a demanar que la probabilitat sigui exacta, i ens conformem amb una molt bona aproximació, podem fer servir el fet que la distribució de $X(365)$ es pot aproximar per una normal de paràmetres $\mu = 365 \cdot 0{,}1 = 36{,}5$ i $\sigma^2 = 365 \cdot 0{,}09 = 32{,}85$. Així:

$$P(X(365) \geq 50) = P\left(\frac{X(365) - 36{,}5}{\sqrt{32{,}85}} \geq \frac{50 - 36{,}5}{\sqrt{32{,}85}}\right)$$

i si anomenem Z una variable aleatòria normal (0,1), aquesta probabilitat serà aproximadament:

$$P\left(Z \geq \frac{50 - 36{,}5}{\sqrt{32{,}85}}\right) = P(Z \geq 2{,}36) = 0{,}0091$$

Per tant, la probabilitat aproximada que ens tornin els diners 50 cops o més al llarg de l'any és únicament del 0,0091.

Observeu que hem calculat $P(X(365) \geq 50)$, però que aquesta quantitat és la mateixa que $P(X(365) \geq 49{,}5)$, ja que la variable només pren valors naturals. Fixeu-vos que si l'aproximem per la normal, obtindrem:

$$P(X(365) \geq 49{,}5) = P\left(\frac{X(365) - 36{,}5}{\sqrt{32{,}85}} \geq \frac{49{,}5 - 36{,}5}{\sqrt{32{,}85}}\right)$$

$$= P\left(Z \geq \frac{49{,}5 - 36{,}5}{\sqrt{32{,}85}}\right) = P(Z \geq 2{,}26) = 0{,}0119$$

que és una quantitat lleugerament diferent de l'obtinguda abans. Es diu que aquest valor s'ha obtingut fent una correcció de continuïtat, ja que aproximem una variable discreta per una de contínua. Podem considerar bons tots dos resultats.

1.1. Estudi de la proporció

Hem vist que quan n és gran, podem aproximar una binomial (n,p) per una normal de paràmetres $\mu = np$ i $\sigma^2 = np(1-p)$. D'altra banda, sabem que ens podem mirar la variable aleatòria binomial com la suma de n variables aleatòries amb distribució de Bernoulli de paràmetre p. Si dividim aquesta suma per n, obtenim clarament la proporció d'èxits.

Una **proporció** correspon a fer la mitjana de n variables aleatòries de Bernoulli de paràmetre p, on n és la mida de la mostra i p, la probabilitat d'èxit de cada esdeveniment individual.

Exemple de la loteria

En l'exemple de la loteria, podem pensar que $X(n)$, el nombre de vegades que ens han tornat els diners en n sortejos, és una suma de n variables, cadascuna de les quals val 1 si aquell dia concret ens han tornat els diners, i 0 en cas contrari. La suma de les n variables ens dóna el nombre de vegades que ens han tornat els diners en els n sortejos, i si dividim per n obtenim la proporció de sortejos en què ens tornen els diners.

Exemple de càlcul d'una proporció

Si volem calcular la proporció de catalans que s'han connectat avui a Internet, podem considerar que a cada català li correspon una variable Bernoulli que val 1 si es connecta o 0 si no es connecta. Per a calcular la proporció, hem de dividir el nombre de catalans que s'han connectat pel nombre total de catalans.

Com que hem vist que la suma de n distribucions de Bernoulli de paràmetre p, que és una binomial (n,p), és aproximadament una distribució normal amb mitjana np i variància $np(1-p)$, és clar que la proporció (que és la suma de les n distribucions de Bernoulli dividida per n), tindrà esperança p i desviació típica $\sqrt{p(1-p)/n}$.

Utilitat de les proporcions

L'estadística cada cop s'utilitza més i les enquestes apareixen cada dia als diaris. Ens interessa saber quina proporció d'electors votaran un determinat partit, quina proporció de ciutadans rebutja un determinat pla o una determinada llei que està preparant el govern, quina proporció de consumidors estaran interessats en un nou producte que volem treure al mercat, etc.

Per tant, quan la mida de la mostra, n, és gran, la distribució de la proporció és aproximadament una distribució normal d'esperança p i desviació típica $\sqrt{p(1-p)/n}$. En aquest cas, $\sqrt{p(1-p)/n}$ correspon a l'error estàndard.

Exemple de distribució de la proporció

Preguntem a una mostra d'habitants d'una població la seva opinió sobre la possible construcció d'un pantà. La probabilitat que un individu concret de la població estigui d'acord

amb la construcció del pantà és p, i n és el nombre d'habitants entrevistats. El 30% dels enquestats estan a favor de la construcció del pantà, és a dir, podem establir que $p = 0{,}3$. Si hem preguntat a 400 habitants, aleshores trobem que la distribució de la proporció d'habitants que estan a favor de la construcció del pantà, que denotarem per p, és:

$$N\left(0{,}3;\ \frac{0{,}3(1-0{,}3)}{400}\right) = N(0{,}3;\ 0{,}0005)$$

Per a calcular la probabilitat que la proporció d'habitants a favor sigui més gran del 40%, hauríem de fer:

$$P(\hat{p} > 0{,}4) = P\left(\frac{\hat{p} - 0{,}3}{\sqrt{0{,}0005}} > \frac{0{,}4 - 0{,}3}{\sqrt{0{,}0005}}\right) = P(Z > 4{,}47) = 0$$

on Z indica una distribució normal estàndard.

2. El teorema del límit central

Sabem que la distribució de la mitjana mostral d'una variable normal o bé té distribució normal o bé es correspon amb una t de Student. També hem vist que si les variables originals segueixen una distribució de Bernoulli, aleshores la seva mitjana és una proporció i, en aquest cas, quan n és prou gran, la seva distribució mostral també és una normal.

El darrer resultat és cert sigui quina sigui la distribució de les dades originals. És a dir, no cal que partim ni de distribucions normals ni de distribucions de Bernoulli, ja que per a mostres de mides prou grans, la distribució de la mitjana mostral és normal sigui quina sigui la distribució original. Aquest resultat fonamental de l'estadística té un nom propi: el *teorema del límit central*.

El **teorema del límit central** diu que si una mostra és prou gran ($n > 30$), sigui quina sigui la distribució de la variable d'interès, la distribució de la mitjana mostral serà aproximadament una normal. A més, la mitjana serà la mateixa que la de la variable d'interès, i la desviació típica de la mitjana mostral serà aproximadament l'error estàndard.

Què vol dir *n* prou gran?

Considerarem que n és prou gran quan, com a mínim, $n > 30$.

Una conseqüència d'aquest teorema és la següent:

Donada qualsevol variable aleatòria amb esperança μ i per a n prou gran, la distribució de la variable $(\overline{X} - \mu)/(\text{Error estàndard})$ és una normal estàndard.

Càlcul de l'error estàndard

Recordem que si la variable té una desviació típica coneguda σ, l'error estàndard es pot calcular com $\sigma/\sqrt{n}$.

Quan σ és desconeguda, calculem l'error estàndard com $s/\sqrt{n}$.

Exemple d'aplicació del teorema del límit central

Una empresa de missatgeria que opera dins la ciutat triga una mitjana de 35 minuts a dur un paquet, amb una desviació típica de 8 minuts. Suposem que durant el dia d'avui han repartit 200 paquets.

a) Quina és la probabilitat que la mitjana dels temps de lliurament d'avui estigui entre 30 i 35 minuts?

b) Quina és la probabilitat que, en total, pels 200 paquets, hagin estat més de 115 hores?

Considerem la variable X = "Temps de lliurament del paquet". Sabem que la seva mitjana és 35 minuts i la seva desviació típica és 8. Però fixeu-vos que no sabem si aquesta variable se-

gueix una distribució normal. Durant el dia d'avui s'han lliurat $n = 200$ paquets. És a dir, tenim una mostra $x_1, x_2, ..., x_n$ de la nostra variable.

Pel teorema del límit central, sabem que la mitjana mostral es comporta com una normal d'esperança 35 i desviació típica:

$$\frac{8}{\sqrt{200}} = 0,566$$

Si fem servir aquesta aproximació, ja podem contestar la pregunta **a**. Hem de calcular:

$$P(30 \leq \overline{X} \leq 35) = P\left(\frac{30-35}{0,566} \leq \frac{\overline{X}-35}{0,566} \leq \frac{35-35}{0,566}\right)$$

que és aproximadament igual a la probabilitat següent:

$$P\left(\frac{30-35}{0,566} \leq Z \leq \frac{35-35}{0,566}\right) = P(-8,83 \leq Z \leq 0) \approx 0,5$$

on Z és una normal (0,1). És a dir, tenim una probabilitat aproximada de 0,5 que la mitjana del temps de lliurament d'avui hagi estat entre 30 i 35 minuts.

Pel que fa a la segona pregunta, d'entrada hem de passar les hores a minuts, ja que aquesta és la unitat amb què ens ve donada la variable. Observeu que 115 hores per 60 minuts ens donen 6.900 minuts. Se'ns demana que calculem la probabilitat següent:

$$P\left(\overline{X} > \frac{6.900}{200}\right) = P(\overline{X} > 34,5)$$

i com que sabem que la mitjana es distribueix aproximadament com una normal de mitjana 35 i desviació típica 0,566 (suposarem sempre que la distribució de la mitjana és normal, ja sigui perquè la variable d'interès és normal o perquè la mostra és prou gran), aquesta probabilitat es pot aproximar per la probabilitat d'una distribució normal estàndard Z:

$$P\left(Z > \frac{34,5-35}{0,566}\right) = P(Z > -0,88) = 1 - P(Z < -0,88) = 1 - 0,1894 = 0,8106$$

2.1. Control de qualitat

Un dels casos més habituals en què podem aplicar el teorema del límit central és a l'hora de fer un procés de control de qualitat.

Entendrem per **control de qualitat** el seguiment d'una certa variable aleatòria en un procés de producció a partir de la mitjana de mostres successives.

Establirem un interval, de manera que les mitjanes que caiguin fora d'aquest interval ens indicaran que hi ha alguna anomalia en el procés de producció en aquell instant. Els límits d'aquest interval s'anomenen **límits de control**.

Si μ és l'esperança de la variable d'interès, σ la desviació típica i considerem una mostra d'aquesta variable de mida n, els límits de control vindran donats per $\mu + 3\frac{\sigma}{\sqrt{n}}$ i $\mu - 3\frac{\sigma}{\sqrt{n}}$. És a dir, calculem tres cops l'error estàndard a

banda i banda de la mitjana. Per tant, la longitud de l'interval és dues vegades el triple de l'error estàndard.

Per què agafem aquest interval? Si apliquem el teorema del límit central sobre la variable d'interès, sabem que la mitjana de n dades es distribueix com una normal amb mitjana μ i desviació típica $\frac{\sigma}{\sqrt{n}}$. Es demostra fàcilment que la probabilitat que una mitjana estigui fora de l'interval $\mu + 3\frac{\sigma}{\sqrt{n}}$ i $\mu - 3\frac{\sigma}{\sqrt{n}}$ és de 0,003 (això vol dir que un valor fora d'aquest interval, si el procés funcionés correctament, es pot donar només amb una probabilitat de 0,003). Per tant, quan es doni un valor fora de l'interval pensarem que no és casualitat i que el problema és que la variable no es comporta com suposàvem.

Exemple de realització d'un control de qualitat

Considerem una màquina que omple pots de iogurt. Suposem que, de mitjana, cada pot conté 125 grams de iogurt amb una desviació típica d'1,5 grams. Cada setmana fem un control de la màquina: analitzem una mostra de 30 pots de iogurt i calculem la mitjana d'aquests 30 iogurts. En aquest exemple, l'error estàndard és:

$$\frac{1,5}{\sqrt{30}} = 0,274$$

Per tant, els límits de control seran:

$$125 + 3 \cdot 0,274 = 125,82$$
$$125 - 3 \cdot 0,274 = 124,18$$

Així, doncs, si la mitjana de les mostres setmanals de mida 30 està entre aquests dos valors, considerarem que tot és correcte. Mentre que si és inferior a 124,18 o superior a 125,82 suposarem que hi ha alguna anomalia en el procés de producció, i caldrà revisar-lo.

Per cert, fixeu-vos que per a fer aquest control de qualitat només cal fer malbé 30 iogurts cada setmana.

3. Resum

En aquesta sessió hem presentat un resultat fonamental de l'estadística, el teorema del límit central. L'hem desenvolupat a partir de l'estudi d'una proporció. Hem acabat veient una de les seves aplicacions més usuals, la realització d'un control de qualitat.

Exercicis

1. En un experiment de laboratori es mesura el temps d'una reacció química. S'ha repetit l'experiment 98 vegades i s'obté que la mitjana dels 98 experiments és de 5 segons amb una desviació de 0,05 segons. Quina és la probabilitat que la mitjana poblacional μ difereixi de la mitjana mostral en menys de 0,01 segons?

2. S'estableix un control de qualitat per a un procés de producció de bales. S'ha establert que quan el procés està sota control, el diàmetre de les bales és d'1 cm amb una desviació típica de 0,003 cm. Cada hora es prenen mostres de 9 bales i se'n mesuren els diàmetres. Els diàmetres mitjana de 10 mostres successives, en centímetres, són:

1,0006	0,9997	0,9992	1,0012	1,0008
1,0012	1,0018	1,0016	1,0020	1,0022

Estabiu quins són els límits de control i expliqueu què podeu concloure sobre el procés de producció en aquests instants.

Solucionari

1. Com que la mostra és gran, pel teorema del límit central podem suposar que la distribució de la mitjana és una normal de mitjana μ i desviació típica, l'error estàndard. Per tant, la probabilitat que ens demanen, que és:

$$P(|\overline{X}-\mu|<0,01) = P(-0,01<\overline{X}-\mu<0,01) = P\left(-\frac{0,01}{\frac{0,05}{\sqrt{98}}}<\frac{\overline{X}-\mu}{\frac{0,05}{\sqrt{98}}}<\frac{0,01}{\frac{0,05}{\sqrt{98}}}\right) =$$

$$= P\left(-1,98<\frac{\overline{X}-\mu}{\frac{0,05}{\sqrt{98}}}<1,98\right)$$

es pot aproximar per la probabilitat d'una distribució normal estàndard Z:

$$P(-1,98<Z<1,98) = 1-2\cdot 0,0239 = 0,9522$$

Per tant, la probabilitat que ens demanen és de 0,9522.

2. Observem que la mitjana $\mu = 1$ i que l'error estàndard és:

$$\frac{\sigma}{\sqrt{n}} = \frac{0,003}{\sqrt{9}} = 0,001$$

Per tant, els límits de control seran 1,003 i 0,997. Observem que absolutament totes les mitjanes que hem obtingut de les successives mostres estan dins l'interval format pels dos límits de control. És a dir, no hi ha cap dada superior a 1,003 ni cap dada inferior a 0,997. Per tant, podem concloure que el procés de control ha estat correcte durant el temps que l'hem analitzat, i que no hem detectat cap anomalia.