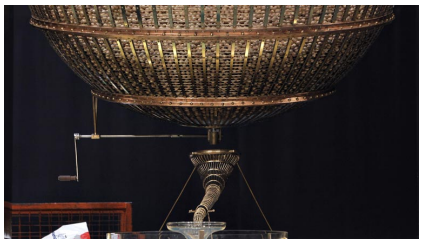
MAT^2

MATerials MATemàtics
Volum 2015, treball no. 5, 13 pp. ISSN: 1887-1097
Publicació electrònica de divulgació del Departament de Matemàtiques
de la Universitat Autònoma de Barcelona
www.mat.uab.cat/matmat

Com fer-se ric amb la loteria (i aprendre Estadística en l'intent)

Llorenç Badiella



L'Estadística Aplicada és una disciplina centrada en l'anàlisi d'informació quantitativa amb la intenció de descriure, predir i, si pot ser, entendre fenòmens d'interès. Aquest objectiu es duu a terme mesurant la variabilitat en els resultats observats i avaluant-ne possibles causes més enllà de la variabilitat natural atribuïble a l'atzar.

Un tema de gran repercussió social que suscita de forma recurrent un gran interès per l'Estadística i la Teoria de Probabilitats és la Loteria de Nadal. Malgrat que tots sabem que la loteria de Nadal és un joc d'atzar profundament injust i que té esperança negativa (per aquell qui juga), no podem resistir la temptació de comprar una o més participacions cada any. Donat que existeix una petita probabilitat de fer-se rics sorgeix el dubte sobre l'existència d'alguna estratègia que permeti aconseguir aquest propòsit amb majors garanties.

De fet, hi ha molts participants que recorren a rituals, tradicions, amulets o conjurs per tal d'atraure la bona sort (fregar el bitllet per la panxa d'una noia embarassada, entrar a l'administració de Loteria que ven els bitllets amb el peu dret o cremar bitllets de sortejos anteriors, òbviament no premiats). També hi ha una tendència força generalitzada a adquirir nombres que representin efemèrides, que tinguin algun valor simbòlic per al participant o a rebutjar nombres que semblen lletjos, com si la component mística dels nombres pogués generar mala o bona sort.

En el present treball intentarem esbrinar si una anàlisi des d'un punt de vista més científic aplicant tècniques Estadístiques permet establir alguna

estratègia que millori les possibilitats que la nostra inversió en la loteria de Nadal sigui més fructuosa. Ja posats, un objectiu complementari consistirà en mostrar l'ús de la prova χ^2 i alguna de les seves variants, justament una de les proves més rellevants en Estadística Aplicada.

1 La Loteria de Nadal

Figura 1: Escena típica del sorteig de la Loteria de Nadal.
(Font: www.lavanguardia.com)

La loteria Extraordinària de Nadal és un joc molt popular que se celebra cada any el 22 de desembre. Essencialment, i sense ànim d'entrar en detalls prou coneguts per tothom, podem resumir la sistemàtica del sorteig indicant que, mitjançant un procediment folklòric (figura 1), s'escullen un nombre de boletes d'un cabàs i se'ls assigna un premi de certa quantia triat d'un altre cabàs.

Figura 2: Bitllets i dècims de loteria.
(Font: www.lavanguardia.com)

Les boletes representen els bitllets que els participants han adquirit amb

anterioritat (figura 2). Més detalls sobre el sorteig es poden consultar arreu, vegeu per exemple la wikipèdia [1].

El primer cabàs conté 100 000 boletes, totes elles fetes amb fusta de boix, de 3 cm de diàmetre i numerades (figura 3).

Figura 3: El bombo amb les 100 000 boletes.
(Font: www.lavanguardia.com)

D'aquest cabàs s'extreuen exactament 1 807 boletes a les que s'assignarà algun dels premis possibles, descrits a la taula 1. Les quantitats en premis corresponen a cada dècim de bitllet que té un cost de 20 euros.

Categoria	Import del Premi	Nombre de boletes premiades
1era (La Grossa)	400 000	1
2ona	125 000	1
3era	50 000	1
4rta	20 000	2
5ena	6 000	8
Premis menors (pedrea)	100	1 794
Total		1 807

Taula 1: Premis de la Loteria de Nadal

Cal tenir en compte que les boletes premiades són en realitat moltes més que les triades, ja que el sorteig incorpora reintegraments i altres bonificacions a tots aquells nombres amb certa semblança als nombres que han rebut els premis principals.

1.1 Material i mètodes

Les dades emprades per a les diferents anàlisis provenen de la lectura automatitzada dels llençols de resultats dels anys 2011 a 2014 (figura 4) que

apareixen publicats l'endemà del sorteig als diferents mitjans de comunicació per tal que hom comprovi si li ha tocat algun premi.


Lotería Nacional
SORTEO EXTRAORDINARIO DE NAVIDAD
SORTEO NÚM. 102 DEL AÑO 2014
LISTA DE LOS NÚMEROS PREMIADOS EN CADA UNA DE LAS CIENTO SESENTA SERIES CORRESPONDIENTES AL SORTEO CELEBRADO EN MADRID EL DÍA 22 DE DICIEMBRE DE 2014
Unidad
Decena
Centena
Tres mil
Seis mil
Veinte mil
Veintitrés mil
Ocho mil
Once mil
07637
500.000
EUROS
Quince mil
13437
4.000.000
DE EUROS
Dieciocho mil
Treinta y siete mil
Treinta y dos mil
Cuarenta mil
Cuarenta y tres mil
Cuarenta y seis mil

Figura 4: Secció del llençol de resultats de l'any 2014.

Les dades històriques sobre els bitllets premiats amb la Grossa provenen de la wikipèdia [2].

Totes les proves estadístiques realitzades i presentades tot seguit s'han dut a terme emprant el software lliure R v3.2 [3].

Després d'aquest preàmbul ja tenim tots els ingredients per començar l'anàlisi dels resultats històrics, detectar desequilibris i definir l'estratègia que ens permetrà assolir gloriosos èxits.

2 La terminació de la Grossa

Com a primera aproximació al problema, podem realitzar un petit estudi sobre la terminació de la Grossa (el premi principal). Aquesta anàlisi té certa rellevància donat que tots el bitllets amb la mateixa terminació rebran un reintegrament dels diners apostats.

En la figura 5 s'observa la distribució del nombre d'aparicions de cadascuna de les possibles terminacions per als 204 sortejos realitzats al llarg de la història del sorteig.

La terminació amb menor freqüència d'aparicions és l'1, tan sols agraciada 8 vegades, mentre que el dígit final més freqüent és el 5 amb 33 aparicions. Donat que el total de sortejos és 204, el nombre esperat d'aparicions seria de 20.4 .

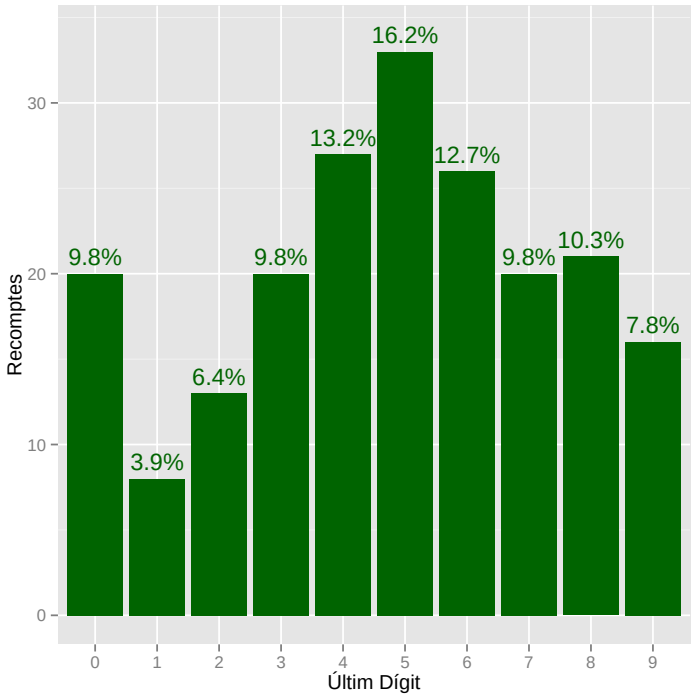


Figura 5: Distribució de l'últim dígit de la Grossa

Per tal de mesurar fins a quin punt aquestes diferències observades són atribuïbles a l'atzar, podem aplicar el test estadístic adient (tests coneguts també com proves de contrast d'hipòtesis).

Per a tots aquells lectors poc familiaritzats amb aquestes tècniques, el procediment vindria a ser el següent:

a) En primer lloc cal establir un escenari neutre i sovint homogeni en relació a allò que es vol comprovar (per exemple que totes les proporcions d'aparició són iguals). Aquest escenari, que ha de ser especialment concret i fàcil de mesurar, rep el nom d'hipòtesi nul·la. De fet, l'interès de la prova rau en constatar que la hipòtesi nul·la no és certa (i per tant les proporcions d'aparició serien diferents), justament l'objectiu de l'anàlisi.

b) Tot seguit, es procedeix a recopilar o recollir dades experimentals sobre el fenòmen que es vol avaluar.

c) El següent pas consisteix en quantificar d'alguna forma la distància entre allò que s'ha observat i el què hauria estat esperable si la hipòtesi nul·la fos certa. La magnitud d'aquesta distància (també anomenada estadístic de la prova) permetrà identificar anomalies. Segons les propietats de les

variables implicades caldrà triar la prova estadística adient, que farà servir la seva pròpia mesura de distància. De fet, darrera de cada mesura de distància hi ha una distribució estadística particular.

d) Les evidències que proporciona el test es concreten amb el p-valor (també anomenat significació estadística), una mesura de la probabilitat d'observar distàncies encara majors sota la hipòtesi nul·la. Per al càlcul del p-valor s'utilitza la distribució estadística corresponent.

e) Concloent, p-valors petits indicaran certes evidències que la hipòtesi nul·la no és compatible amb el resultat observat, permetent rebutjar la suposició inicial. Hi ha cert consens generalitzat a rebutjar la hipòtesi nul·la quan el p-valor és inferior a 0.05 .

En el cas sobre la distribució de l'últim dígit, la hipòtesi nul·la correspon a considerar que la probabilitat d'observar cadascuna de les terminacions en un any qualsevol és la mateixa per a tots els dígits: $p = 0.1$. La prova estadística per a aquest escenari concret rep el nom de prova χ^2 de bondat d'ajust [4, 5], ja que es comparen les freqüències observades de cert esdeveniment amb les freqüències esperades sota una distribució teòrica de referència.

La mesura de distància que utilitza la prova χ^2 de bondat d'ajust és:

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

on O_i i E_i corresponen als valors observats i esperats respectivament per a cada valor i n és el nombre de possibles valors (terminacions). La distribució estadística associada a aquesta distància és la distribució χ^2_{n-1}, on $n-1$ són els graus de llibertat.

En fi, el resultat d'aplicar el test a aquestes dades és:

```
Chi-squared test for given probabilities
Test Statistic = 22.6667, df = 9, p-value = 0.0070
```

En aquesta sortida del programa R, `Test Statistic` dóna el valor de D i `df` dóna els graus de llibertat (*degrees of freedom*), valor que correspon al nombre de categories menys un, és a dir `df` $= 10 - 1 = 9$. Així doncs, mitjançant aquesta tècnica podem rebutjar la hipòtesi que les probabilitats d'ocurrència dels diferents dígits siguin totes iguals amb una significació estadística de 0.0070 . En conclusió, des del punt de vista estadístic, surten més cops les terminacions en 4, 5 i 6, i no seria massa aconsellable jugar a números acabats en 1 o 2. Tot i que aquest resultat sembla engrescador,

aquest possible indici de desequilibri, en cas de ser cert, no seria suficient per tal de capgirar l'esperança negativa del joc.

Cal tenir en compte però, que el sorteig no ha estat homogeni al llarg dels anys i aquest fet podria posar en dubte els resultats anteriors: als inicis del sorteig no es feien servir boletes sinó paperetes, el nombre de boletes que hi participen s'ha incrementat en diverses ocasions i per últim, les boletes del sorteig han estat renovades en diverses ocasions. En resum, caldrà examinar altres detalls més fins per tal de donar alguna credibilitat als pseudo-indicis anteriors donat que el sorteig ha patit serioses variacions estructurals al llarg del temps.

Si volem centrar-nos en la loteria moderna amb el format clàssic de boletes i cabassos conegut per tothom, cal anar amb compte, ja que hi ha hagut canvis importants en el seu format: des de 1983 fins el 2004 hi participaven 66 000 boletes, fins al 2010, 85 000 i a partir de 2011, es van afegir 15 000 boletes més, totalitzant les 100 000 boletes actuals. D'aquesta manera, des de 2011 fins el 2014 el sorteig ha estat en principi homogeni. L'anàlisi centrat en les dades dels darrers quatre anys no requerirà cap ajust especial i serà totalment vàlid.

3 Nombres multi-premiats: Sortejos 2011 a 2014

Com s'ha comentat en la introducció, en cada sorteig en el format actual, es seleccionen 1 807 boletes de les 100 000 boletes disponibles, a les quals se'ls assigna un premi concret. Si bé és cert que hi ha prop del 15% dels bitllets que reben algun tipus de bonificació, l'anàlisi per tal d'avaluar possibles desequilibris en el sistema cal centrar-lo justament en les particularitats de les 1 807 boletes seleccionades. Així doncs, si el sorteig no té cap anomalia, la probabilitat que en un any qualsevol un número concret sigui triat és de 0.01807 i és la mateixa per a tots els números. Una possible estratègia de treball consisteix a comptar per als darrers quatre anys quantes boletes han rebut 1, 2, 3, 4 premis o cap i comparar aquests valors amb la distribució teòrica assumint que el sorteig és neutral.

Si el sorteig és efectivament homogeni, la variable aleatòria (diguem-ne X) que representa el nombre de premis assolits en els darrers quatre sortejos es distribueix seguint una llei binomial amb paràmetres $p = 0.01807$ i $n = 4$. De fet, per a qualsevol valor k entre 0 i 4,

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

Per exemple, la probabilitat que una boleta qualsevol hagi rebut com a mínim una pedrea en els darrers 4 anys és del 7%: $P(X > 0) = 1 - P(X = 0) = 0.0703$.

A la taula 2 es mostra el nombre de repeticions observades per als darrers quatre anys de les 100 000 boletes i es compara amb els resultats esperats sota una distribució binomial teòrica amb els paràmetres indicats. Com a detall, no hi ha cap boleta que hagi estat premiada en tots i cadascun dels darrers quatre sortejos, però n'hi ha dues que han estat escollides en tres ocasions.

Nombre premis	Observats	Esperats
0	92 954	92 965.56
1	6 866	6 843.21
2	178	188.90
3	2	2.32
4	0	0.01
Total	100 000	100 000

Taula 2: Nombres multi-premiats als anys 2011 a 2014 observats vs esperats

Apliquem ara un test χ^2 de bondat d'ajust a la distribució Binomial:

```
Goodness-of-fit test for binomial distribution
Test Statistic = 0.7603, df = 4, p-value = 0.9437
```

El resultat indica manca total d'evidències i per tant no podem rebutjar la hipòtesi d'homogeneïtat. El p-valor és 0.9437. De fet, a la taula anterior es pot comprovar com els valors esperats són efectivament molt propers als valors observats.

4 Període d'inclusió de les boletes al sorteig

Com s'ha comentat en la secció anterior, el sorteig ha anat ampliant el nombre de boletes que hi participen al llarg dels anys. Posats a buscar possibles desequilibris, sembla raonable imaginar que per polítiques d'estalvi, quan el sorteig ha incrementat el nombre de boletes, no s'ha realitzat una renovació exhaustiva, tan sols s'han afegit les noves boletes participants. D'aquesta manera, les boletes més antigues haurien estat exposades a un major desgast, afectant en certa mesura la seva probabilitat de ser escollides. Cal dir que periòdicament s'examinen les boletes manualment i que aquelles que tenen

desperfectes es reemplacen. Per altra part, com s'observa a la figura 6, les boletes (que són fetes amb fusta de boix) tenen diferent tonalitat. De nou podria ser raonable pensar que si les boletes han estat fabricades en diferents lots, poden tenir certes diferències.

Figura 6: Boletes ja alineades un cop s'ha realitzat el sorteig.
(Font: www.lavanguardia.com)

Per tal de dur a terme aquesta nova anàlisi procedirem a comparar la distribució del nombre de premis rebuts als darrers quatre anys en funció del grup d'antiguitat al qual pertany cada boleta:

- Període 1: Boletes 0 a 65999, conjunt de boletes antigues, configuraven el sorteig des de 1983 a 2004.
- Període 2: Boletes 66000 a 84999, introduïdes el 2005.
- Període 3: Boletes 85000 a 99999, introduïdes el 2011.

Els resultats de la distribució del nombre de premis en els darrers 4 anys segons període són els següents:

Nombre de premis	Període 1	Període 2	Període 3
0	92.93%	92.74%	93.32%
1	6.89%	7.08%	6.52%
2	0.18%	0.17%	0.16%
3		0.01%	
Total	100%	100%	100%

Taula 3: Premis de la Loteria de Nadal

Per tal de mesurar la significació estadística dels resultats, apliquem una nova versió del test χ^2, en aquest cas el test d'independència entre variables

(període i nombre de premis) que permetrà validar els resultats observats amb els resultats que hom esperaria si no hi hagués cap relació entre les variables analitzades:

```
Pearson's Chi-squared test
Test statistic = 13.0793, df = 6, p-value = 0.0418
```

Per a les taules de doble entrada, el nombre de graus de llibertat és sempre (nombre de files -1) $\times$ (nombre de columnes -1). Per tant, en el nostre cas, `df` $= (3-1)\times(4-1) = 6$. S'assoleix certa significació estadística, per sota de 0.05 , permetent concloure que el període i el nombre de premis assolits no són del tot independents. Aquestes diferències detectades, tot i que existeixen des d'un punt de vista estadístic, són més aviat irrellevants: si recalculem el percentatge de boletes escollides dins de cada període (recordem que cada any s'escullen un 1.81% de les boletes), obtenim per a les respectives etapes: 1.81% , 1.86% i 1.71%. Això indicaria que les boletes del tercer període, tenen una probabilitat de ser escollides subtilment menor.

5 Pes de la tinta

La darrera proposta d'anàlisi consisteix a examinar amb detall les característiques de les boletes, concretament els números que duen impresos, donat que podrien alterar el seu pes, la seva dinàmica o qualsevol altra característica que podria ser rellevant. Si els números fossin gravats o bé pintats, podríem imaginar l'existència d'alguna mena de relació entre la probabilitat de ser triades i la quantitat de tinta emprada (o el volum del gravat). A la figura 6 no s'observa clarament si els números tenen algun tipus de relleu, però algunes fonts ([1], per exemple) indiquen que els números són en realitat impresos amb làser (fet que faria irrellevant la qüestió plantejada). Per si de cas i com a última temptativa, procedirem a realitzar l'anàlisi corresponent, ja que gairebé hem esgotat les possibilitats d'assolir el nostre objectiu.

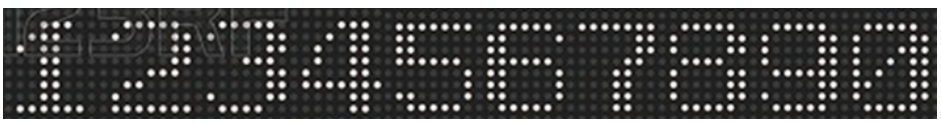


Figura 7: Estimació de la quantitat de tinta per número.

Per determinar la quantitat de tinta de cada número, cal localitzar una tipografia prou semblant a la que s'utilitza en les boletes i tot seguit quantificar els píxels de cadascun dels dígits, obtenint finalment un indicador de la quantitat de tinta de cadascuna de les boletes (figura 7).

Podem representar la quantitat de tinta de les diferents boletes mitjançant un histograma del nombre de puntets (figura 8). El valor mitjà és de 133.5 .

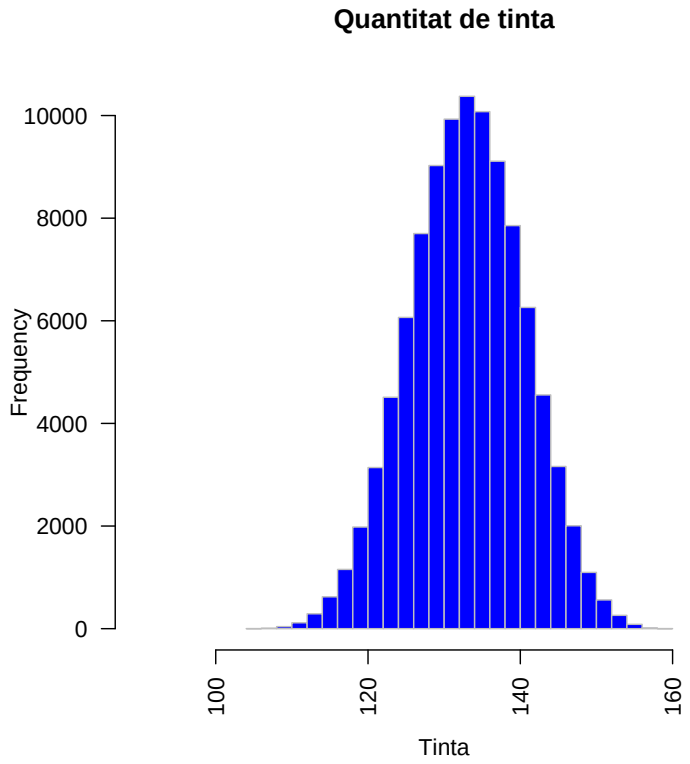


Figura 8: Distribució de la tinta

A partir d'aquesta quantitat estimada de tinta procedim a classificar els diferents números en quatre grups: menys de 128 puntets, entre 128 i 133, entre 134 i 139 i finalment 140 o més. Si bé és cert que la discretització de variables no és recomanable degut a la pèrdua d'informació, aquesta categorització permetrà emprar de nou la prova χ^2 d'independència, només aplicable quan es disposa de dues variables categòriques.

```
Pearson's Chi-squared test
Test Statistic = 3.5763, df = 6, p-value = 0.7338
```

En l'anàlisi de les taules de contingència és habitual col·lapsar categories molt poc freqüents per tal d'evitar-ne la seva sobre-representació i millorar la validesa del test. En el nostre cas, les categories corresponents a 2, 3 i 4 premis s'han agrupat. Així doncs `df` $= (3-1) \times (4-1) = 6$. El resultat

d'aplicar la prova estadística pertinent ens indica que no podem establir cap relació entre la quantitat de tinta i la freqüència d'obtenció de premis, el p-valor és de 0.7338 .

6 Conclusió

La principal conclusió del següent treball és que no hem aconseguit establir cap estratègia per fer-nos rics jugant a la loteria de Nadal.

Tampoc hem pogut detectar cap anomalia rellevant, tan sols hem descobert un petit indici de desequilibri associat al període d'introducció de les boletes al sistema, que caldrà validar emprant les dades dels propers sortejos. Encara que es confirmés, continuaríem lluny de capgirar l'esperança negativa del joc.

En fi, malgrat els resultats negatius de l'estudi, com que ben segur continuarem comprant loteria (a veure si toca), només ens queda desitjar a tothom bona sort!

Nota final per a tots aquells que encara tinguin dubtes sobre l'existència de patrons amagats o confiïn en poders místics dels nombres: les dues boletes que han estat seleccionades en 3 dels darrers quatre sortejos corresponen als nombres 71362 i 74515 .

Referències

[1] https://es.wikipedia.org/wiki/Sorteo_Extraordinario_de_Navidad

[2] http://es.wikipedia.org/wiki/Anexo:Sorteo_Extraordinario_de_Navidad

[3] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria (2014). URL: http://www.R-project.org/.

[4] Agresti, A., & Kateri, M. *Categorical data analysis*. Springer Berlin Heidelberg (2011).

[5] Peña, D. *Fundamentos de Estadística*. Alianza Editorial (2014).

Servei d'Estadística Aplicada &
Dept. de Matemàtiques
Universitat Autònoma de Barcelona
badiella@mat.uab.cat

Publicat el 11 de desembre de 2015