UAB

Universitat Autònoma de Barcelona

Consultoria Estadística

LOTERIA DE NADAL

Andrea Acuña Villagaray (1639232)
Maria Marín Méndez (1668394)

4 de Febrer, 2026

Contents

1 Resum

Aquest estudi realitza una anàlisi exhaustiva de la Loteria de Nadal (període 2000-2025) per dets de *web scraping* i modelització avançada, les autores han validat la integritat del sistema, concloent que:

- L'equitat física està garantida: L'ús de boles de fusta de boix amb gravat làser evita biaixos per pes de tinta.
- Convergència a la uniformitat: L'estudi de més de 45.000 boles extretes confirma que totes les terminacions tenen probabilitats similars, seguint la Llei dels Grans Nombres.
- Incapacitat predictiva: S'han testejat models de *Machine Learning* (Random Forest) i estadístics (LMM, GLM Gamma) per intentar predir el premi segons la paritat o la suma de dígits, sense èxit. El sorteig és, a efectes pràctics, atzar pur.

2 Objectiu i motivacions del treball

L'estudi no neix sols d'una curiositat acadèmica, sinó d'una disputa familiar clàssica que es repeteix cada desembre a cada d'una de les autores.

L'interès per aquest tema sorgeix de la dinàmica entre el seu germà, d'esperit profundament científic i recional, i la seva mare. Cada any, quan arriba el moment de comprar els dècims, s'encén el debat on el seu germà sosté amb fermesa que la loteria és una pèrdua de temps i diners, argumentant que estadíticament, "mai toca res" i la seva mare, en canvi, manté viva il · lusió i la tradició, convençuda que, per petita que sigui la probabilitat, "alguna cosa tocarà".

Aquesta tensió entre el rigor científic i la il · lusió popular ha estat el motor per plantejar els següents objectius:

1. **Verificar l'aleatorietat**: Determinar si el sorteig és realment atzarós o si el germà té raó en la seva visió escèptica.
2. **Avaluar la predictibilitat**: Comprovar si existeix alguna estratègia basada en dades històriques que pugui donar la raó a la mare i augmentar les possibilitats de guanyar.
3. **Analitzar mites populars**: Sotmetre a prova creences com la dels números "bonics" o la paritat per veure si tenen fonament real.

3 Loteria de Nadal

En aquest repositori es recull el desenvolupament del Treball Final de Consultoria Estadística 2025, centrat en l'anàlisi exhaustiva de la Loteria de Nadal des de la seva vessant més tècnica. L'objectiu no és sols descriure el sorteig, sinó avaluar amb rigor estadístic si la variabilitat dels resultats històrics (des del 2000 fins a l'anys 2025) respon purament a l'atzar o si existeixen anomalies mesurables en l'homogeneïtat del sistes,

A través de metodologies de web scraping, tests d'independència, etc

La Loteria de Nadal no és sols un sorteig de boles; és l'unic moment de l'any en què un país sencer es posa d'acord per ignorar les lleis de l'estadística. Des d'un punt de vista matemàtic, és un "impost a l'esperança", però des del punt de vista de lesa dades, és un ecosistema fascinant.

L'arquitectura del "GORDO"

El sistema de la Loteria Nacional no treballa amb números a l'atzar, sinó amb una jerarquia rígida que determina les probabilitats reals d'èxit. L'estructura per al sorteig actualment es basa en:

- **Univers numèric:** 100.000 números únics (del 00000 al 99999). Això estableix una probabilitat base de guanyar el primer premi amb un sol dècim del 0.001%.
- **Emissió:** La societat Estatal de Loteries i Apostes de l'Estat (SELAE) ha emès per l'any pasat 197 sèries per cada números. Atès que cada sèrie es divideix en 10 dècims, hi ha un total de 1970 dècims de cada número al mercat.
- **Volum econòmic:** Amb un preu de 20€ per dècim, la recaptació potencial ascendeix a 3.940 milions d'euros. D'acord amb la normativa, el 70% d'aquest import es destina a premis.

Física i Homogeneïtat del Sorteig

Com assenyla el professor Llorenç Badiella, el que percebem com a folklore televisiu és, en realitat, un procés de física aplicada dissenyat per garantir l'equitat absoluta:

- **Les boles:** Hi ha 100.000 boles al bombo gran, totes fabricades en fusta de boix, amb un diàmetre de 3 cm i un pes unificat.
- **Impressió làser:** Per evitar que la pintura alteri el pes (eliminant la teoria que números amb molta tinta, com el 88888, pesen més que l'11111), els números estan gravats amb làser.

- **Mecànica:** Es fan servir dos bombos simultanis, el gran per als números i el petit per a les 1805 boles de premis. Els sorteif només finalitza quan el bombo de premis queda totalment buit.

El repartiment del "Pastís"

Tot i que el focus està en el "Gordo", la realitat és que el sorteig és una "pluja fina" de premis petits, per augmentar l'esperança per l'any que ve.

Premi	Import per dècim	Boles premiades	Probabilitat
1r premi ("el Gordo")	400.000€	1	0,001%
2n premi	125.000€	1	0,001%
3r premi	50.000€	1	0,001%
4rt premi	20.000€	2	0,002%
5é premi	6.000€	8	0,008%
La Pedrea	100€	1.794	1,794%
Reintegrament	20€	1 de cada 10 xifres	10,00%

Premi per proximitat i derivat

Més enllà de les boles extretes, existeixen premis calculats per la relació numèrica amb els guanyadors:

- **Aproximacions (a):** Premien els números immediatament anterior i posterior del "Gordo", 2n premi i del tercer, on s'obtén 200€, 125€ i 96€ respectivament.
- **Centenes (c):** Es premien els 99 números que comparteixen les tres primeres xifres amb els quatre primers premis (tota la centena), on s'obtén 100€.
- **Terminacions (t):** Es premien els números que coincideixen en les dues últimes xifres amb els tres primers premis, on s'obtén 100€.
- **Reintegrament:** el retorn del valor del dècim (20€) si l'última xifra coincideix amb la del primer premi. Hi ha un 10% de probabilitat d'obtenir-lo.

L'impacte Fiscal

És important destacar que el premi "net" és inferior al nominal per a les quantitats grans. L'agència Tributària aplica un 20% d'impost a la quantitat que superi els 40.000€ (que estan exempts):

- Gordo: Es tributa pel 20% de 360.00€. Premi net = 328.000€
- 2n Premi: Es tributa pel 20% de 85.000€. Premi net = 108.000€
- 3r Premi: Es tributa pel 20% de 10.000€ = 48.000€

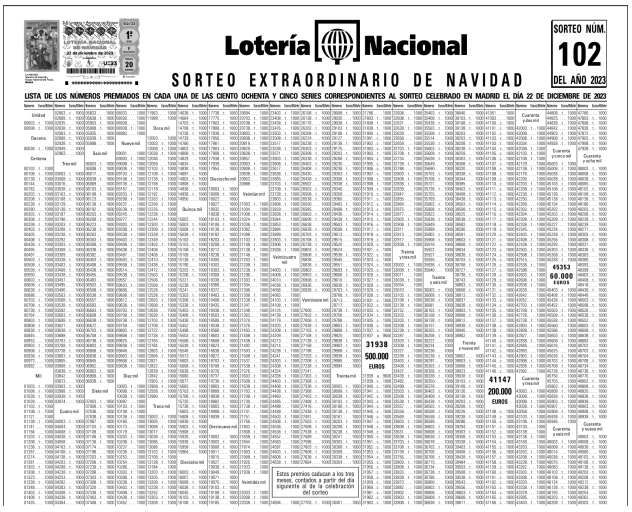
4 Lectura de dades i preprocessament

A continuació, s'introdueix el conjunt de dades utilitzat en l'estudi i es descriu el procés seguit per a la seva obtenció i el preprocessament necessari. Aquest pas és fonamental per garantir la integritat de l'anàlisi descriptiu i la validesa dels models posteriors. L'objectiu inicial és explorar el comportament general dels números premiats mitjançant taules de síntesi i visualitzacions gràfiques.

4.1 Font d'obtenció i naturalesa de les dades

El conjunt de dades utilitzat en aquest estudi s'ha extret principalment dels arxius oficials publicats per la Sociedad Estatal de Loterías y Apuestas del Estado (SELAE), l'entitat responsable de la Loteria de Nadal. Aquests arxius, disponibles de manera sistemàtica des de l'any 2000, contenen la totalitat dels números premiats en cada sorteig, incloent-hi la seva categoria i l'import corresponent.

A la Figura següent es pot observar un exemple de la font original d'on s'han obtingut les dades, mostrant l'estructura típica del llistat oficial:


Lotería Nacional

SORTEO EXTRAORDINARIO DE NAVIDAD

SORTEO NÚM. 102 DEL AÑO 2023

LISTA DE LOS NÚMEROS PREMIADOS EN CADA UNA DE LAS CIENTO OCHENTA Y CINCO SERIES CORRESPONDIENTES AL SORTEO CELEBRADO EN MADRID EL DÍA 22 DE DICIEMBRE DE 2023

Estos premios caducan a los tres meses, contados a partir del día siguiente al de la celebración del sorteo

Figure 1: Resultats Sorteig Loteria de Nadal (2023)

A partir d'aquesta font, s'ha realitzat una extracció de la informació rellevant (any, número, premi, lletra), que posteriorment s'ha emmagatzemat en un fitxer amb format `.txt` per al seu processament. La lectura i el tractament de les dades s'ha implementat mitjançant un script d'R extern, dissenyat per automatitzar la unificació dels 26 anys analitzats. També s'ha generat un fitxer en format `.xlsx` per facilitar la reproductibilitat de l'estudi per part de tercers.

4.2 Estructura del Dataset

Per tal de familiaritzar-nos amb la matriu de dades, es presenten a continuació dues taules. La primera mostra un extracte real dels números premiats corresponents al sorteig més recent (2025), mentre que la segona ofereix una visió agregada de tot el període d'estudi. Cada registre inclou el número premiat, la seva categoria (identificada per la "lletra"), l'import del premi en euros i l'any del sorteig.

Table 2: Primers registres del sorteig 2025

numero	lletra	premi	categoria	any
00032	t	1000	Decena	2025
00048	t	1000	Decena	2025
00082	NA	1000	Decena	2025
00093	NA	1000	Decena	2025
00112	NA	1000	Centena	2025

Table 3: Resum global del conjunt de dades (2000-2025)

Variable	Valor
Anys analitzats	26
Total de números premiats	127664
Premi mínim (€)	901.52
Premi màxim (€)	4000000
Categories (Lletra)	a: 156 | c: 11879 | t: 70118 | Bombo: 45511

Podem observar com cada registre té l'import del premi, la seva categoria, la moneda (Euros o pessetes) i l'any del sorteig. En el període de 26 anys que analitzarem, hi han hagut 127.664 números premiats.

També s'utilitza el conjunt de dades de l'històric dels premis de la Grossa des de l'inici del sorteig, l'any 1812, fins a l'actualitat. Aquestes dades han estat obtingudes a partir de fonts disponibles a internet, i podeu veure un extracte de les dades a continuació:

Table 4: Conjunt de dades de la Grossa (1812-2025)

Any	Numero	Terminació
1812	03604	4
1813	08553	3

5 Estadístiques descriptives i anàlisi de l'atzar

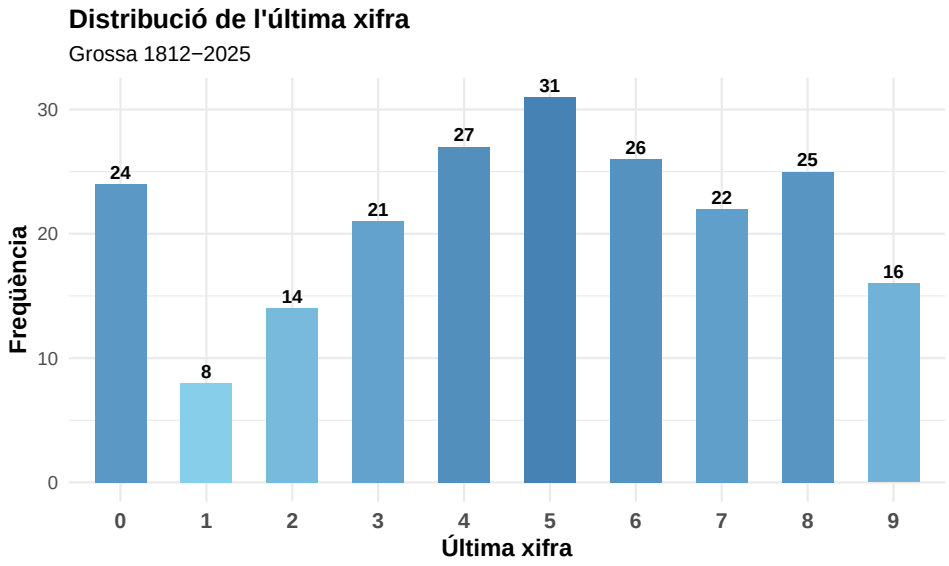
Un cop realitzada la lectura del conjunt de dades, procedim amb l'anàlisi descriptiu per proporcionar una visió global del comportament de les dades per identificar distribucions, així com detectar possibles patrons o irregularitats del sorteig.

5.1 Distribució històrica de l'última xifra (1812-2025)

Tot i que l'anàlisi principal d'aquest estudi se centra en el període 2000-2025, s'ha considerat rellevant examinar la distribució històrica de l'última xifra del número guanyador del primer premi des de l'inici del sorteig, l'any 1812.

Aquest estudi és interessant, ja que si un dècim coincideix en l'última xifra amb la del primer premi, el jugador obté el reintegrament, recuperant així els diners invertits. Per aquest motiu, l'estudi de les terminacions pot aportar informació addicional sobre el comportament global del sorteig al llarg del temps.

L'anàlisi permetrà avaluar si hi ha algun patró existent en la freqüència d'aparicions de les xifres.



A partir de les dades analitzades, s'observa que les terminacions 5, 4, 6 i 8 són les que han aparegut amb més freqüència com a última xifra del número guanyador, amb 31, 27, 26 i 25 aparicions, respectivament. En canvi, les terminacions 1 i 9 presenten les freqüències més baixes.

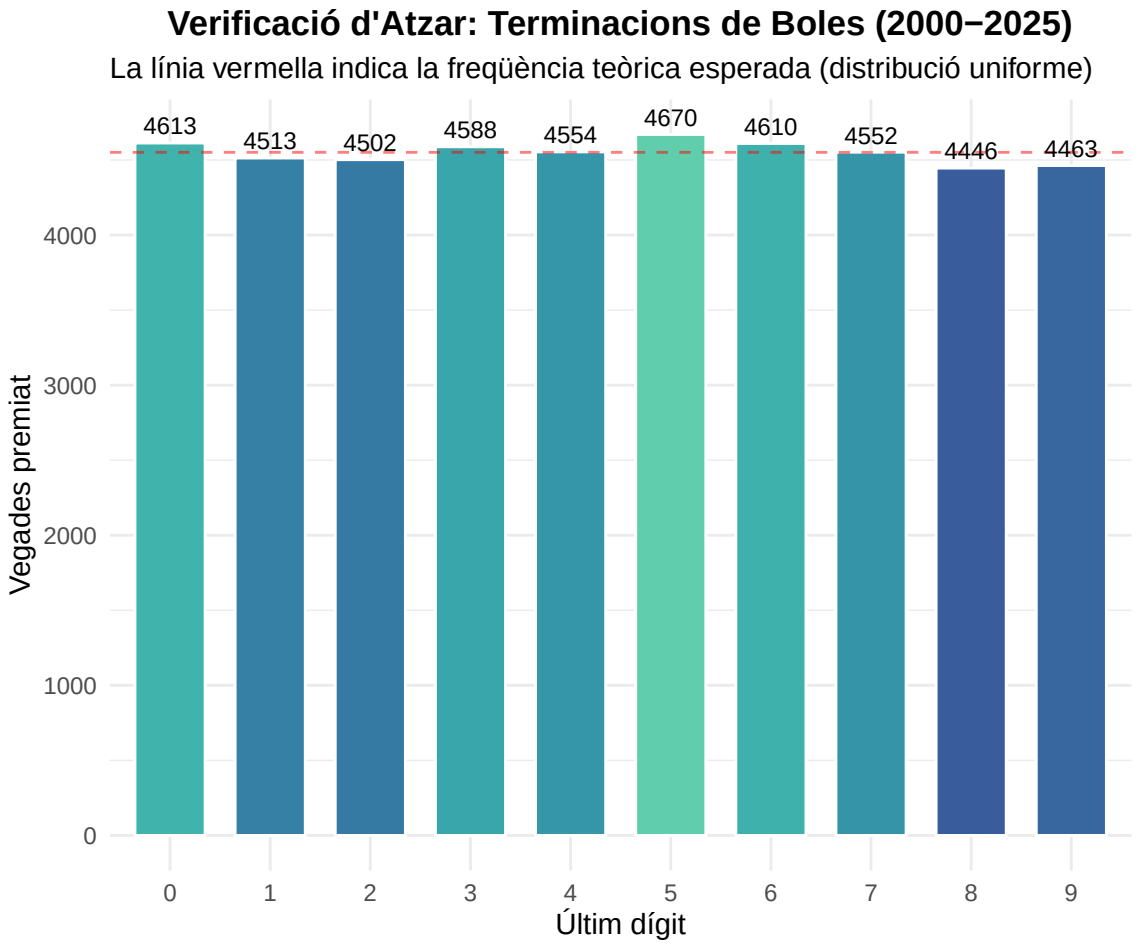
Tot i aquestes diferències, el comportament de les dades és compatible amb la variabilitat d'un procés aleatori. Per tant, en principi, no podrem establir cap predicció fiable sobre futures terminacions.

5.2 Verificació d'homogeneïtat en el període modern (2000-2025)

Més enllà del primer premi, hem analitzat els **127.664 registres** de números premiats en els darrers 26 anys. En explorar aquest volum de dades, hem detectat una distinció crucial per a la validesa de l'estudi:

1. **Premis de bombo (Atzar físic):** Són els números que surten físicament del bombo (Pedrea i premis majors). Al nostre dataset, són aquells que no tenen cap lletra associada, bàsicament són aquells que representen l'extracció directe i aleatòria.
2. **Premis derivats:** Són aquells premis que no surten d'una bola pròpia, sinó que es concedeixen per la relació numèrica amb els guanyadors principals, com ja ho haviem nombrat abans.

A continuació, comparem la distribució de les terminacions de les boles extretes per veure si l'atzar és realment homogeni:



Aquest gràfic ens revela un convergència clara cap a la uniformitat estadística que valida la integritat del sortieig en el període modern. A diferència de la mostra històrica reduïda de la Grossa, l'estudi de les més de quaranta-cinc mil boles extretes del bombo confirma que les terminacions segueixen la Llei de Grans Nombre i es distribueixen de manera equilibrada al voltant de la mitjana teòrica.

Aquest equilibri actua com a una evidència empírica de l'equitat mecànica del sistema, demostrant que el disseny físic de les boles i el seu gravat làser garanteixen que cap número tingui una probabilitat d'extracció superior a la resta.

Finalment, els resultats refirmen el rigor metodològic d'haver separat els premis d'extracció directa dels derivats per càlcul, ja que només així s'ha pogut verificar que el bombo funciona com un generador d'atzar pur sense biaixos detectables.

6 Modelització i validació estadística

En aquesta secció, apliquem tècniques d'estadística avançada per sotmetre a prova l'aleatorietat del sorteig. L'objectiu és determinar si existeix alguna variable (com el rang numèric, la paritat o la suma de dígits) que permeti obtenir un avantatge predictiu, o si, per contra, ens trobem davant d'un sistema d'atzar pur.

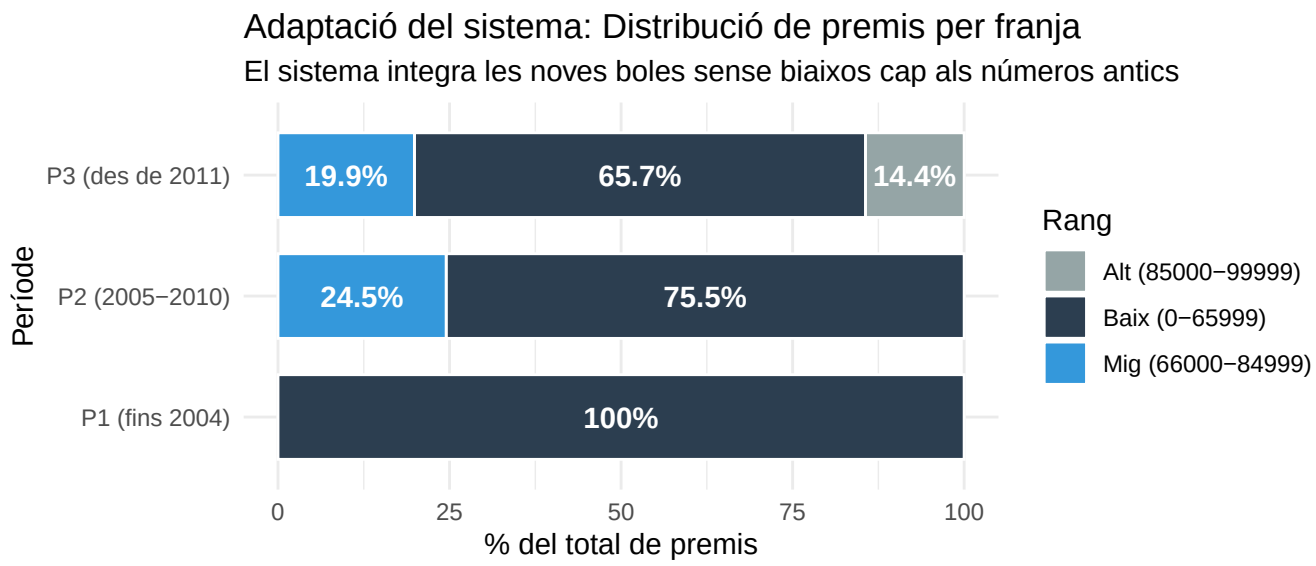
6.1 L'impacte de l'evolució del bombo

Un dels canvis estructurals més rellevants a estudiar al llarg del període estudiat és l'ammpliació progressiva del nombre total de boles al bombo:

- P1 (fins al 2004): 66.000 números.
- P2 (2005-2010): 85.000 números.
- P3 (des de 2011): 100.000 números.

Aquest increment té un efecte directe en l'espai mostral del sorteig. L'anàlisi següent avalua si el bombo s'ha adaptat correctament a aquestes ampliacions, desplaçant proporcionalment els premis cap a les noves franges numèriques.

Per avaluar aquest aspecte, s'han agrupat els números premiats en tres rangs (baix, mitjà i alt), i s'ha avaluat el percentatge de premis assignats a cada franja dins de cada període.



Al gràfic podem observar que, durant el segon període (2005-2010) amb l'ampliació del bombo fins als 85.000 números, els premis es distribueixen gairebé com s'esperava. La distribució teòrica dels premis és del 77.6% per a la franja baixa i del 22.4% per a la franja mitjana. Els percentatges observats s'ajusten bé als valors teòrics, amb una lleugera diferència que es pot explicar amb la variabilitat inherent d'un procés aleatori.

Durant el període P3, la distribució esperada dels premis segons la mida de cada franja és del 66% per a la franja baixa, del 19% per a la franja mitjana i del 15% per a la franja alta. Els percentatges observats es tornen a ajustar molt bé als valors teòrics, sense tenir evidència de biaix, la qual cosa reforça la hipòtesi que els premis es distribueixen de manera aleatòria.

6.2 Anàlisi de variables ocultes: Mite vs. Realitat

Existeix la creença popular que certs números ("bonics", parells, o amb certes sumes) tenen més probabilitat de ser premiats. Per contrastar aquesta hipòtesi amb rigor matemàtic, hem construït un Model Lineal Mixt (LMM).

- Variable Resposta: Logaritme base 10 del premi (*log10(premi)*), per normalitzar la distribució.
- Predictors Fixos: Paritat del número (Parell/Senar) i Suma dels cinc dígits del número (ex: 12345 = 15).
- Efecte Aleatori: Any del sorteig $(1|any)$, per controlar la variabilitat econòmica i estructural de cada edició.

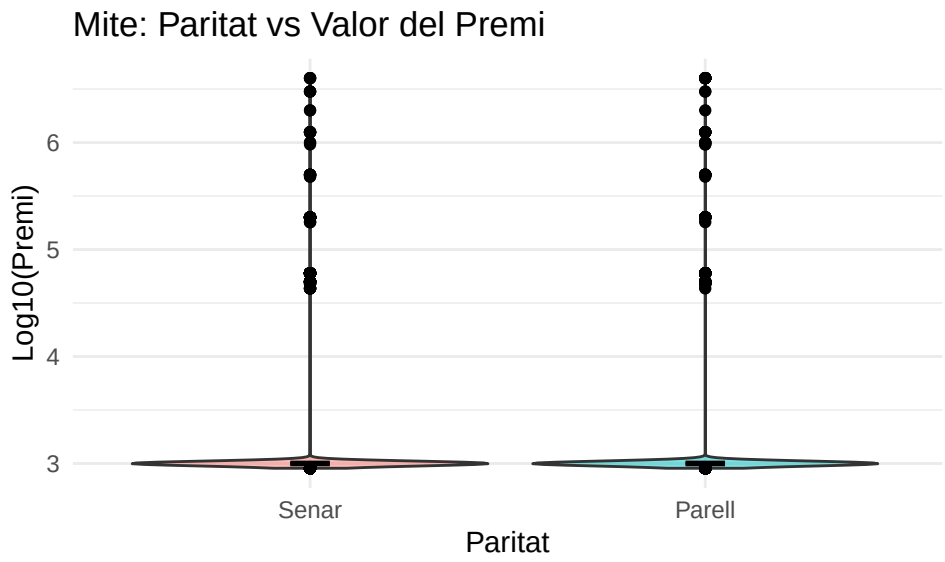
El model s'ha ajustat amb lmer() i mostra l'efecte de la paritat i de la suma dels dígits sobre el logaritme del premi.

Table 5: Coeficients fixos del LMM: Paritat i Suma de dígits

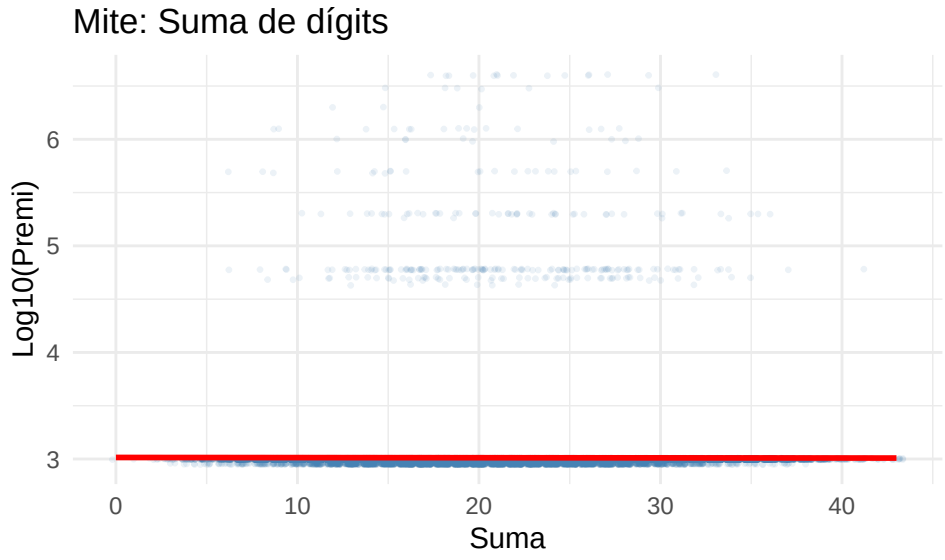
Predictor	Estimate	Std_Error	df	t_value	Pr(>|t|)
(Intercept)	3.0164910	0.0042603	127.9804	708.0411598	0.0000000
es_parellParell	0.0005651	0.0017298	45487.4368	0.3266915	0.7439027
suma_digits	-0.0002532	0.0001376	45456.1750	-1.8394395	0.0658571

Aquesta taula mostra que els coeficients de ser parell i la suma dels dígits són molt propers a zero. Podem observar que p-valors són majors que el nivell de significació del 0.05, i per tant, no són significatius. Indicant que cap de les variables tenen un impacte rellevant sobre el premi.

Per a complementar aquesta anàlisi, hem representat gràficament les variables:



Aquest gràfic ens permet observar si la distribució dels premis varien segons si el número és parell o senar. Tant i com hem vist, les distribucions són molt similars i les medianes gairebé coincideixen, per tant, no hi ha evidència que la paritat afecti el premi.



En aquest segon gràfic observem si existeix alguna relació lineal entre la suma dels dígits del número i el valor del premi. La línia vermella representa la tendència estimada pel model.

Com es pot apreciar, la recta de regressió és pràcticament horizontal i la majoria dels punts es distribueixen de manera uniforme al voltant d'aquesta línia. Això indica qeu la suma dels dígits no exerceix cap influència significativa sobre el logaritme del premi. En ambdós casos, els gràfics serveixen per visualitzar que les creences populars són només mites, i reforcen les conclusions del LMM.

6.3 Comparativa de capacitat predictiva

Per explorar fins a quin punt és possible "guanyar a la banca", hem comparat tres enfocaments per predir la quantia del premi:

1. Model Mixt (LMM): Estadístic clàssic, que incorpora efectes aleatoris per any del sorteig
2. GLM Gamma: Model lineal generalitzat, ideal per a valors monetaris positius.
3. Random Forest: Algorisme de *Machine Learning* capaç de detectar patrons complexos i no lineals.

Nota: Per raons computacionals, hem pres una submostra de 5000 observacions per al Random Forest.

Com a mesura de rendiment dels models utilitzarm l'RMSE (Root Mean Squared Error), que reflecteix l'error mitjà entre els valors reals i els predits:

Model	**RMSE (Error promig en €)**	**Interpretació**
Model Mixt	22844.01	Incapaç de reduir l'error usant l'any.
GLM Gamma	22848.32	Captura la mitjana però no la variància extrema.
Random Forest	16033.3	Malgrat la seva complexitat, no troba patrons.

L'anàlisi de l'Error Quadràtic Mitjà (RMSE) dels models generats mostra la incapacitat de predir el resultat de la loteria. Tots tres models tenen un RMSE elevat, indicant que la predicció exacta del premi és molt difícil.

Això ens torna a confirmar que els premis de la Loteria de Nadal són aleatòris i que cap patró numèric observable permet predir amb precisió la quantia guanyadora.

7 Validació de l'aleatorietat (Test d'Uniformitat)

Per assegurar que el bombo no té "memòria" ni biaixos, apliquem un test uniforme per comprovar que cada xifra té la mateixa probabilitat d'aparèixer en els números guanyadors, apliquem un test de Chi-quadrat.

Table 7: Resultat del test Chi-quadrat per uniformitat

	Estadístic	Grau.de.llibertat	p.valor
X-squared	6.2585	9	0.7138

Estem testejant si els números segueixen una distribució uniforme, és a dir, cada dígit de 0 a 9 apareix amb la mateixa probabilitat. Podem veure que el p-valor és molt gran (>0.05), per tant, no tenim suficient evidències estadístiques per rebutjar la hipòtesi nul · la i acceptem que el sorteig és homogeni.

8 Conclusions

Les conclusions d'aquest treball confirmen de manera rotunda que el Sorteig Extraordinari de Nadal és un esdeveniment basat en l'atzar pur, validant així la postura més escèptica i científica expressada pel germà en el debat familiar. L'anàlisi de la integritat física del procés demostra que l'ús de boles de fusta de boix amb un pes unificat i gravat làser garanteix una equitat absoluta, eliminant qualsevol biaix que pogués derivar de la pintura o el pes dels números. Aquesta aleatorietat es veu reforçada pels tests estadístics d'uniformitat realitzats sobre el període modern, on s'ha obtingut un p-valor de 0,7138. Aquesta xifra obliga a acceptar l'homogeneïtat del sorteig i confirma que, en la pràctica, cada dígit té exactament la mateixa probabilitat de ser extret del bombo.

Pel que fa als mites populars sobre números "bonics" o determinades propietats numèriques, els models lineals mixtos han demostrat que variables com la paritat del número o la suma dels seus cinc dígits no tenen cap impacte significatiu en la quantia del premi obtingut. Les evidències gràfiques han permès visualitzar que les distribucions de premis i les seves medianes són pràcticament idèntiques independentment de si el número és parell o senar , el que desmunta científicament qualsevol estratègia de compra basada en aquests criteris tradicionals. Així mateix, la comparativa de models de capacitat predictiva, incloent-hi algorismes de *Machine Learning* com el *Random Forest*, ha posat de manifest la impossibilitat de predir resultats futurs mitjançant l'estudi de dades històriques, ja que els elevats errors de predicció (RMSE) confirmen que el bombo no té "memòria".

D'aquesta manera, la ciència dóna la raó al germà: la probabilitat de guanyar la Grossa amb un sol dècim es manté immutable en un 0,001% , confirmant que la loteria actua matemàticament com un "impost a l'esperança". No obstant això, per mantenir viva la il · lusió de la mare i maximitzar les opcions de recuperar la inversió, l'estratègia més racional és la diversificació. Comprar números amb terminacions diferents augmenta les possibilitats d'obtenir el reintegrament, que té una probabilitat del 10%. Si es vol seguir la tradició històrica, xifres com el 5, el 4 o el 6 han estat les més freqüents des de 1812 , però cal recordar que el sorteig és, en realitat, una "pluja fina" de premis petits on la Pedrea n'és la gran protagonista amb un 1,794% de probabilitat. En definitiva, el millor premi garantit per a la família és gaudir del folklore i la tradició de compartir el dècim, acceptant que l'atzar n'és l'únic i absolut senyor.

9 Annexos

A continuació, introduirem el directori de Git-Hub on podràs trobar tot el codi d'R utilitzat:

https://github.com/andreaacuvilla/Loteria

10 Referències

Badiella, L. (2005). La física aplicada en la Loteria de Nadal: equitat i mecanismes de sort. Barcelona: Edicions UPC.

La Lotera. (2025). Histórico de números premiados con el Gordo de Navidad. https://lalotera.es/historico-numeros-premiados-gordo-navidad/