CONCURSO ESCOLAR DE TRABAJOS ESTADÍSTICOS

(Curso escolar 2009-2010)

30 de mayo de 2010

Título:

¿Nacemos al azar?

Número de alumnos: 5

Nivel que cursan los alumnos:

2º de Bachillerato. Modalidad: Ciencias de Humanidades y Sociales

¿Nacemos al azar?

Resumen

Este trabajo trata de responder a la pregunta: ¿Nacemos al azar como si lanzásemos un dado de doce caras, o hay causas que influyen en el nacimiento en determinados meses del año?

Índice

1. Introducción **4**

1.1. Metodología del trabajo 4

1.2. Software estadístico 5

2. Planteamiento del problema **6**

2.1. Buscando modelos 6

3. Obtención de datos **7**

4. Buscando patrones **8**

5. Contrastando hipótesis **11**

5.1. Para un mes cualquiera 11

5.2. Para comparar un mes determinado en dos años distintos 12

5.3. Para comparar un mismo mes a lo largo de los años 14

5.4. Para contrastar la equiprobabilidad en cada año 15

5.5. Acumulando nacimientos en años sucesivos 17

6. Estudio en nuestro centro **20**

7. Conclusiones **20**

A. El contraste Chi-Cuadrado **21**

A.1. Contraste para k proporciones 21

A.2. Test de ajuste de distribuciones 22

1. Introducción

Es común haber tenido al menos alguna vez la experiencia de participar en una reunión familiar, una conversación entre amigos o en la celebración de un cumpleaños en la que surge el comentario de que algún mes o época del año presenta más relevancia en los nacimientos. También es común, asentir sin análisis, que así ocurre en efecto, ya que presuponemos que durante las vacaciones, o en las fiestas locales, las parejas engendren más niños y eso cause un sesgo en la distribución de los nacimientos de la población.

El objetivo de este trabajo es abordar estadísticamente esa cuestión para los nacimientos de los niños de la Comunidad de Cantabria, aprovechando que es donde hemos encontrado datos registrados de nacimientos distribuidos por meses.

Una forma vívida de iniciar el estudio de la inferencia estadística es partir de una situación realista en la que sea fácil involucrar a los alumnos como en una controversia, y el problema aquí planteado se presta a ella.

El trabajo consta de una componente informal de análisis exploratorio de datos, mediante tablas y gráficos y la parte formal de inferencia estadística que busca respuestas a preguntas específicas. Se ha procurado evitar el uso de excesivo bagaje matemático pero al mismo tiempo mostrar que los datos numéricos hablan con nosotros por medio de las matemáticas. y que sin ellas no se pueden alcanzar conclusiones.

1.1. Metodología del trabajo

Los alumnos participantes cursan la asignatura de Matemáticas de Ciencias Sociales de 2º de Bachillerato, en cuya programación figuran los contenidos de Inferencia Estadística. Estos contenidos incluyen los conceptos: población y muestra, parámetro y estadístico, estimación de una proporción, la distribución binomial y normal, el contraste de significación para una proporción, hipótesis nula y alternativa, nivel de significación, región de aceptación y de rechazo. Todos estos conceptos en cualquier caso están inmersos en el trabajo.

Se parte del contraste de una proporción que utilizan los alumnos de 2º de Bachillerato de la opción de Ciencias y de Humanidades y a partir de su equivalente con la distribución χ^2, se amplia de una forma gradual por medio del contraste de dos proporcciones hasta

el contraste de más proporciones por medio del contraste Chi-cuadrado.

Se ha cuidado que los contrastes sean graduados, de forma que hasta que los alumnos no han comprendido y trabajado en profundidad un contraste por medio de la realización de las actividades complementarias propuestas, no se accede al siguiente contraste.

Los cálculos se realizan con la hoja de cálculo Calc y se complementan con el uso de comandos estadísticos en R.

En lo posible, se ha evitado el uso excesivo de matemáticas, dado el nivel de los alumnos, pero creemos inadecuado el uso de las fórmulas como si solo fuesen varitas mágicas incomprensibles, y por ello se ha incluido un apéndice explicativo del estadístico de contraste que se utiliza a lo largo de los análisis.

1.2. Software estadístico

Como herramienta de cálculo estadístico se ha utilizado la hoja de cálculo Calc de Openoffice y el paquete estadístico **R** que es un lenguaje y entorno de programación para análisis gráfico y estadístico de datos , dentro de un proyecto de software libre, y uno de los más utilizados en investigación por la comunidad estadística, siendo además muy popular en el campo de la investigación biomédica, bioinformática y de matemáticas financieras.

Como entorno de trabajo y desarrollo de las actividades se ha aprovechado que los alumnos cursan como optativa la asignatura de Tecnología de la Información y la Comunicación que se imparte en el aula de informática. Ello nos ha permitido hacer una introducción en 4 horas del programa R y desarrollar las actividades del trabajo en dicho aula.

En esencia solo se utilizan dos comandos del paquete estadístico R, el de test de proporciones y el test Chi-cuadrado. Esto nos ha permitido trabajar con más profundidad el concepto de nivel de significación, cuyo significado está en la respuesta a la cuestión tantas veces planteada en situaciones de incertidumbre: ¿Cuán raro es este hecho? El nivel de significación permite una cuantificación del mismo, y es, no lo olvidemos, una decisión que debe tomar el diseñador del experimento.

2. Planteamiento del problema

¿Nacemos al azar como el giro caprichoso de una ruleta, o hay causas que favorecen el nacimiento en determinados meses del año?

2.1. Buscando modelos

Nota del profesor: *Se les invita a los alumnos a plantear problemas en distintos contextos que conlleven un modelo similar al problema planteado.* Estos son algunas de las sugeridas por ellos:

Jenifer Los resultados del lanzamiento de un dado en el que las seis caras son equiprobables.

María El número de accidentes de tráfico registrado por meses.

3. Obtención de datos

¿Donde están los datos necesarios para analizar el problema?. ¿Cómo obtenerlos?

Dado que el muestreo o búsqueda y recogida de datos, conlleva una tarea delicada y a menudo ardua, esa tarea suele ser ya realizada por organismos gubernamentales.

Acudimos en internet a la página del Instituto de ICANE[1]. Entre los enlaces respectivos de su página web, *Banco de datos* -> *Series Temporales*, -> *Demografía* -> *Dinámica Demográfica* -> *Movimiento Natural de Población*, encontramos en -> *Natalidad, Mortalidad y Nupcialidad*, el desglose de nacimientos por meses de los años comprendidos entre 1996 y 2008.

La tabla inferior recoge los datos suministrados para un total de 60.013 nacimientos a lo largo de esos 13 años.

	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
ene	302	318	299	332	371	334	373	349	385	424	434	422	457
feb	296	307	298	337	311	325	338	383	389	365	384	411	433
mar	326	315	282	358	360	345	337	375	402	433	473	440	431
abr	314	341	342	327	351	328	357	390	376	421	424	418	512
may	355	342	340	356	393	396	389	376	414	473	461	441	510
jun	332	337	330	294	359	339	368	374	436	416	384	458	493
jul	328	340	310	324	336	362	395	388	446	444	465	468	526
ago	303	330	347	367	385	390	414	395	433	441	401	453	531
sep	339	287	315	344	372	345	389	415	457	480	440	460	454
oct	283	324	330	332	358	402	407	360	459	437	486	518	478
nov	345	324	318	295	376	351	400	363	433	446	456	451	451
dic	336	334	324	355	369	363	393	355	430	487	421	439	484

Tabla 1: Nacimientos por meses de 1996 a 2008 extraídos del ICANE

Después de realizada la consulta en la página citada, los datos se suministran en formato Excel, que abrimos sin problemas con *Calc* de *OpenOffice* que utilizamos en la asignatura de TIC. La incorporación como fuente de datos al programa R ha sido también directa por la variedad de formatos de importación que tiene este programa estadístico.

[1] `http://www.icane.es`

4. Buscando patrones

Con tantos números como ofrecen las tablas estadísticas el primer mecanismo de búsqueda de patrones serán los gráficos. Representemos los nacimientos registrados para cada año y observemos sus gráficas.

Vamos a hacer en primer lugar un análisis exploratorio de los gráficos con objeto de encontrar alguna característica llamativa que revele algún patrón asociado a los distintos meses. ¿Que meses registran el número máximo, y el número mínimo de nacimientos?. ¿Algún detalle con respecto a las estaciones?. ¿Algún mes que presente cierta peculiaridad?, etc.

En algunos años se aprecia algún mes, como por ejemplo octubre, con valores altos en el número de nacimientos, pero esa conducta desaparece en otros años. Cuando en algunos gráficos creemos encontrar una pauta o regularidad, los demás gráficos parecen eliminarla. Podemos repetir esa observación también para el mes de mayo.

Tal vez se pueda destacar que el mes de enero en al menos 8 de los años presenta un valor bajo en nacimientos en comparación con los otros meses.

En cualquier caso, con la figura 1, no vemos nada concluyente, lo que hace pensar en un comportamiento muy aleatorio y sospechar de momento que haya bastante diferencia de nacimientos en los distintos meses.

A continuación hacemos una gráfico conjunto para los 13 años. El diagrama de cajas de la figura 2 es muy interesante, ya que muestra la variabilidad dentro de cada año, y a su vez es comparable de unos años a otros.

- Las cajas, según pasa el tiempo van desplazándose verticalmente, lo que revela que con los años ha habido un aumento progresivo de nacimientos y por tanto aumentado la natalidad.

- Los años 2004, 2006 y 2008 son los que parecen mostrar mayor dispersión. El mayor coeficiente de variación se registra en el año 2006 y la mayor desviación típica en el año 2008.

- Otro detalle de interés son los valores anómalos de alguno de los años. En concreto, en los diagramas de cajas de los años 2000, 2005 y 2007. Esos valores contribuyen a

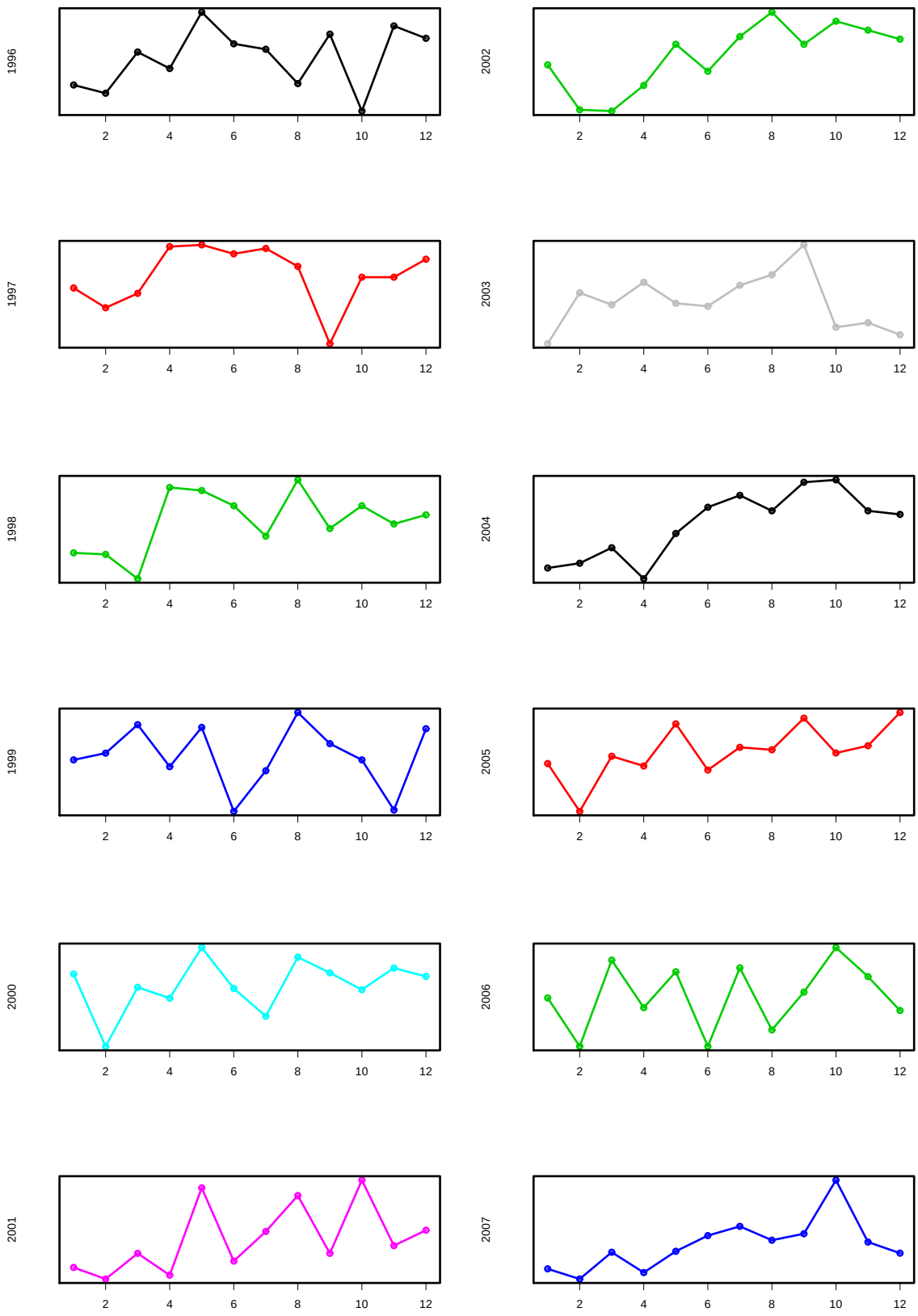

Figura 1: Número de nacimientos mensuales para da año

aumentar la dispersión en los años en que se registran. De hecho la mayor amplitud se presenta en el año 2005.

Actividades complementarias

► Identificar en la figura 2, los años con valores anómalos, y buscar dichos valores en la Tabla 1.

► Realizar el cálculo del rango o amplitud para cada uno de los años. Comparar los resultados obtenidos con los gráficos de cajas de la figura 2.

► Realizar el cálculo de la desviación típica para cada uno de los años. Comparar los resultados obtenidos con los gráficos de cajas de la figura 2.

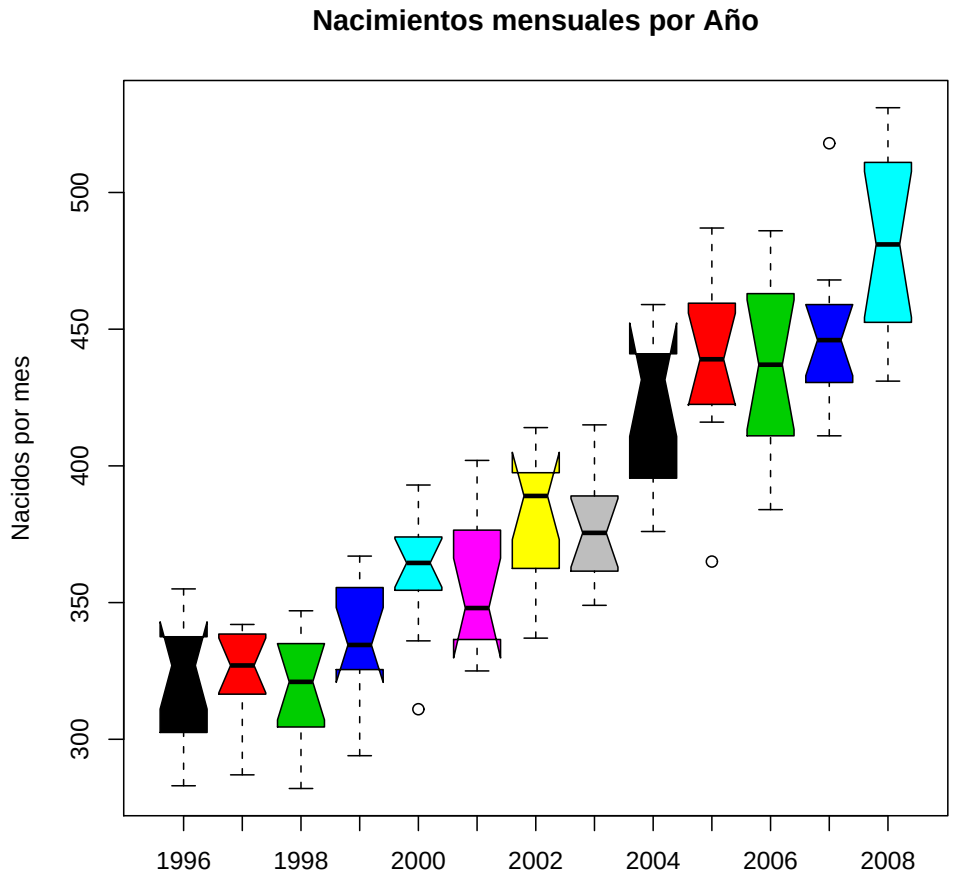

Figura 2: Gráficos de cajas para todos los años

5. Contrastando hipótesis

La hipótesis de partida será que una persona nace al azar de forma similar al resultado de lanzar un dado de 12 caras-meses, con un peso ajustado al número de días que tenga el mes, pues es claro que salvo los años bisiestos, febrero dispone de 28 días y enero de 31 días.

5.1. Para un mes cualquiera

1. Para el primer análisis elegimos por ejemplo el mes de marzo de 1996 (bisiesto[2]), y planteamos como hipótesis inicial que los niños nacen de acuerdo al dado-dodecaedro.

$$\begin{cases} \mathbf{H}_0 \equiv p_{\text{marzo}} = p_0 = \dfrac{31}{366} \\ \mathbf{H}_1 \equiv p_{\text{marzo}} \neq p_0 = \dfrac{31}{366} \end{cases}$$

2. A partir de la tabla, como

Año	1996
Nacidos en Marzo	326
Nacidos en el año	3859

Tabla 2: Contraste para una proporción.

$$p_{\text{marzo}} = \hat{p}_0 = \frac{326}{3859} = 0{,}08447784 \qquad p_0 = \frac{31}{366} = 0{,}08469945$$

Calculamos el estadístico del contraste

$$z = \frac{\hat{p}_0 - p_0}{\sqrt{\dfrac{p_0\, q_0}{n}}} = \frac{0{,}08447784 - 0{,}08469945}{\sqrt{\dfrac{0{,}08469945 \times 0{,}9153005}{3859}}} = -0{,}04944297 \tag{1}$$

Aceptamos la hipótesis nula $\mathbf{H_0}$ pues el valor obtenido $|z| < z_{0{,}975} = 1{,}96$, utilizando un nivel de confianza de 0.95.

[2]Un año es **bisiesto** si es divisible por 4, excepto el último de cada siglo (aquel divisible por 100), salvo que este último sea divisible por 400.

Nota del profesor: *Se introduce en este momento la distribución Chi-Cuadrado para generalizar el contraste anterior.*

Para el contraste anterior podemos utilizar de forma equivalente el estadístico χ^2 de la expresión (7) con $k = 1$, que corresponde al cuadrado del valor hallado en (1)

$$d = \chi^2 = \sum_{i=1}^{1} \frac{(x_i - n_i\, p_0)^2}{n_i\, p_0(1 - p_0)} = \mathbf{0{,}0024}$$

y al ser el valor obtenido $\chi^2 < \chi^2_{1;0,95} = 3{,}8415$, de igual forma, aceptamos la hipótesis de partida, al nivel de confianza de 0.95.

Actividades complementarias

▶ Realizar el contraste anterior para otros meses de otros años, utilizando primero la hoja de cálculo, y a continuación con *R*.

▶ Realizar el contraste al nivel 0.99 para el valor anómalo del año 2005 que se aprecia en la figura 2, utilizando la hoja de cálculo y el comando de *R*.

▶ De los 156 meses de los 13 años, solo hay 3 meses con estadístico significativo a un nivel de confianza del 0.99. Están en los años 2002, 2007 y 2008. Observando la tabla 1 hay que detectarlos, realizar los contrastes y llegar a una conclusión.

5.2. Para comparar un mes determinado en dos años distintos

Planteamos la pregunta, ¿*los nacimientos ocurridos en un mes determinado se comportan de la misma forma en dos años distintos*?.

1. Para este contraste elegimos por ejemplo el mes de marzo de 1996, y el mes de marzo de 2008 y planteamos la hipótesis que la proporción de los niños que nacen en marzo es la misma para esos dos años.

$$\begin{cases} \mathbf{H}_0 \equiv p^{1996}_{\text{marzo}} = p^{2008}_{\text{marzo}} = p_{\text{marzo}} \\ \mathbf{H}_1 \equiv p^{1996}_{\text{marzo}} \neq p^{2008}_{\text{marzo}} \end{cases} \tag{2}$$

2. El contraste para dos proporciones de poblaciones independientes se puede realizar, cuando se dan las condiciones de aproximación a la distribución normal mediante la expresión (3), siendo las proporciones de cada población $\hat{p}_1 = \dfrac{x_1}{n_1}$ $\quad$ $\hat{p}_2 = \dfrac{x_2}{n_2}$

Y el estadístico de contraste está dado por

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0 \hat{q}_0 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{3}$$

donde $\hat{p}_0 = \dfrac{x_1 + x_2}{n_1 + n_2}$, y z es la distribución normal estándar $N(0,1)$,

3. A partir de la tabla,

Año	1996	2008
Nacidos en Marzo	326	431
Nacidos en el año	3859	5760

Tabla 3: Comparación de 1 mes en 2 años distintos.

Calculamos el estadístico del contraste (3)

$$z_{1-\alpha/2} = \frac{0,0845 - 0,0748}{\sqrt{0,0787 \times 0,9213 \left(\frac{1}{3859} + \frac{1}{5760}\right)}} = 1,72303$$

Aceptamos la hipótesis de partida pues el valor obtenido $|z| < z_{0,975} = 1{,}96$, utilizando un nivel de confianza de 0.95.

Para el contraste anterior podemos utilizar de forma equivalente el estadístico χ^2 de la expresión (7) con $k = 2$, que corresponde al cuadrado del valor hallado

$$d = \chi^2 = \sum_{i=1}^{2} \frac{(x_i - n_i\, p_0)^2}{n_i\, p_0 (1 - p_0)} = \mathbf{2{,}968829}$$

y al ser el valor obtenido $\chi^2 < \chi^2_{1;0,95} = 3{,}8415$, de igual forma, aceptamos la hipótesis de partida, al nivel de confianza de 0.95.

,

Actividades complementarias

▶ Realizar el contraste anterior para por ejemplo el mes de Junio de 1996, y el mes de Junio de 2008. Se aceptará la hipótesis inicial al comprobar que el estadístico $\chi^2 = 0{,}0058 < \chi^2_{1;0,95} = 3{,}8415$.

5.3. Para comparar un mismo mes a lo largo de los años

Vamos a extender el análisis anterior a los datos disponibles de los trece años. Para que sea una extensión de los casos anteriores elegiremos el mes de marzo y nos preguntamos, *¿la probabilidad de nacer en el mes de Marzo se mantiene en los trece años, de 1996 a 2008?*

$$\begin{cases} \mathbf{H}_0 \equiv p^{1996}_{\text{marzo}} = p^{1997}_{\text{marzo}} = \ldots = p^{2008}_{\text{marzo}} = p_{\,0} \\ \mathbf{H}_1 \equiv p^{i}_{\text{marzo}} \neq p_{\,0} \text{ para algún año } i \end{cases} \tag{4}$$

	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
mar	326	315	282	358	360	345	337	375	402	433	473	440	431
Totales	3859	3899	3835	4021	4341	4280	4560	4523	5060	5267	5229	5379	5760

Tabla 4: Nacimientos registrados en el mes de Marzo de 1996 a 2008

En el contraste planteado en (4), p_0 se estima con $\hat{p}_0 = \dfrac{x_1 + \ldots + x_{13}}{n_1 + \ldots + n_{13}} = 0,0812657$, y utilizando el estadístico χ^2 de la expresión (7) con $k = 13$, como en casos anteriores

$$d = \chi^2 = \sum_{i=1}^{13} \frac{(x_i - n_i\, p_0)^2}{n_i\, p_0(1 - p_0)} = 19{,}9485$$

- Obtenemos $\chi^2 = 19{,}9485$ que no es significativo al nivel 0.99, pues es menor que $\chi^2(12, 0{,}99) = 26{,}21697$

Como conclusión **aceptamos que la probabilidad de nacer en el mes de Marzo en cualquiera de los años estudiados es la misma** .

Actividades complementarias

► Realizar el contraste anterior para los 11 meses restantes y obtener la tabla siguiente

	ene	feb	mar	abr	may	jun	jul	ago	sep	oct	nov	dic
χ^2	5.46	14.22	19.95	17.08	8.86	16.09	10.22	17.38	18.36	29.05	12.90	9.33

Es interesante la conclusión, pues salvo el mes de octubre, los valores obtenidos de χ^2 son inferiores al valor crítico $\chi^2(12, 0{,}99) = \mathbf{26.22}$ y por tanto aceptamos que **para cada mes, la probabilidad de nacer en él es la misma en cualquiera de los años estudiados**.

En el caso de **octubre** es fácil observar en la Tabla 1, que es el mes con más amplitud, con valor máximo de 518 en 2007 y un mínimo en 1996 con 283 nacimientos.

5.4. Para contrastar la equiprobabilidad en cada año

En los casos anteriores se ha contrastado una proporción o más de dos proporciones en años diferentes o independientes. Ahora plateamos la pregunta original que era, *¿nacemos al azar como si lanzásemos un dado de doce caras?*.

Vamos a realizar el contraste para el año 1996

$$\begin{cases} \mathbf{H}_0 \equiv p^{1996}_{\text{enero}} = p^{1996}_{\text{feb}} = \ldots = p^{1996}_{\text{dic}} \\ \mathbf{H}_1 \equiv p^{1996}_i \neq p^{1996}_j \text{ para algún } i \neq j \end{cases} \tag{5}$$

En la tabla siguiente elegimos los datos para el año 1996.

dias	mes	O_i	p_i	E_i	$(O_i - E_i)^2/E_i$
31	ene	302	0.0847	326.86	1.890
29	feb	296	0.0792	305.77	0.312
31	mar	326	0.0847	326.86	0.002
30	abr	314	0.0820	316.31	0.017
31	may	355	0.0847	326.86	2.423
30	jun	332	0.0820	316.31	0.778
31	jul	328	0.0847	326.86	0.004
31	ago	303	0.0847	326.86	1.741
30	sep	339	0.0820	316.31	1.627
31	oct	283	0.0847	326.86	5.884
30	nov	345	0.0820	316.31	2.602
31	dic	336	0.0847	326.86	0.256
366		N =3859			**17.537**

- La columna O_i registra los datos o nacimientos observados en cada mes.

- La columna p_i no son exactamente igual a $1/12$ pues como ya quedó claro no todos los meses tienen el mismo número de días, siendo por tanto la probabilidad $p_2 = 0{,}0792$ de nacer en febrero con 29 días en este año bisiesto y algo más alta la probabilidad de los meses con 31 días con $p_i = 0{,}0847$.

- La columna E_i registra los valores esperados para cada mes si repartimos el total de los 3859 nacidos de este año de acuerdo a las p_i, que son $E_i = N \times p_i$

- En la última columna se hallan los sumandos del estadístico χ^2 de la fórmula

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - n_i\, p_i)^2}{n_i\, p_i} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{6}$$

- El valor obtenido en la expresión (6) se considera aceptable si no cae en la región crítica, sombreada en negro en la figura, es decir cuando $\chi^2 < \chi^2_{0{,}99,11} = 24{,}72497$.

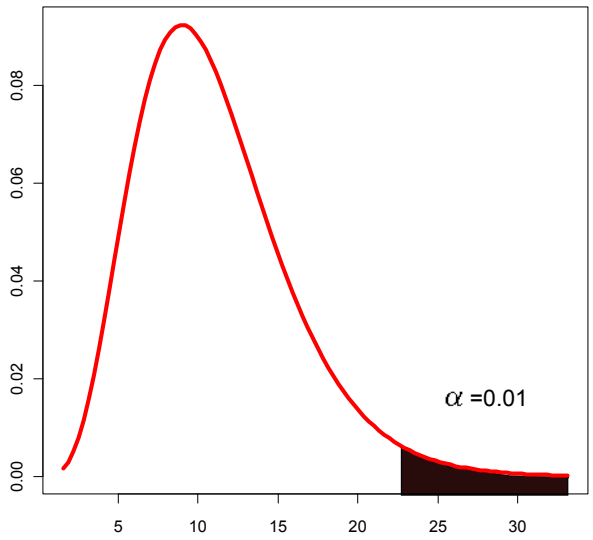

Figura 3: Función de densidad χ^2 con 11 g.l.

- Una vez realizados los cálculos que figuran arriba en la hoja de cálculo, se comprueba el mismo resultado con el comando de **R**:

- Se concluye tras el contraste, dado que $\chi^2 = 17{,}537 < \chi^2_{0{,}99,11} = 24{,}72497$, que aceptamos que **para el año 1996 los niños nacen aleatoriamente de acuerdo al dado-dodecaedro**.

Actividades complementarias

► Realizar el contraste anterior para los 12 años restantes de 1997 a 2008.

Se ha realizado el contraste siguiendo los pasos del ejemplo anterior en la hoja de cálculo y con el program **R** para todos los años y los valores del estadístico χ^2 obtenidos en cada año se muestra en la tabla siguiente

	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
χ^2	17.54	7.99	13.94	16.34	10.87	14.49	13.75	17.57	19.66	17.30	20.60	15.81	23.74

Tabla 5: Valores de χ^2 de 1996 a 2008, todos inferiores al valor crítico $\chi^2_{0,99,11} = 24{,}72497$

Lo que nos lleva a **aceptar que para los años registrados, los niños nacen aleatoriamente de acuerdo al dado-dodecaedro**.

5.5. Acumulando nacimientos en años sucesivos

Nota del profesor: *Se propone a los alumnos que investiguen lo que ocurre cuando se acumulan los nacimientos de años anteriores.*

A continuación vamos a ir agregando los nacimientos de años anteriores, como se puede apreciar en la Tabla 6 y que se reproduce de forma gráfica en la Figura 4.

Año 2008 La columna del año 2008 en la Tabla 6 corresponde a la distribución por mes de nacimiento de todos los menores de 13 años. La suma hace un total de 60.013 niños.

Lo más llamativo son los valores de los meses de **enero con 4800**, febrero con 4577, **marzo con 4877** y **mayo con 5246**.

No resaltamos el valor del mes de febrero con tan solo 4577 nacimientos, ya que se puede deber al menor número de días de este mes, aunque más adelante haremos el contraste adecuado que determine si se debe realmente a eso solo.

	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
ene	302	620	919	1251	1622	1956	2329	2678	3063	3487	3921	4343	4800
feb	296	603	901	1238	1549	1874	2212	2595	2984	3349	3733	4144	4577
mar	326	641	923	1281	1641	1986	2323	2698	3100	3533	4006	4446	4877
abr	314	655	997	1324	1675	2003	2360	2750	3126	3547	3971	4389	4901
may	355	697	1037	1393	1786	2182	2571	2947	3361	3834	4295	4736	5246
jun	332	669	999	1293	1652	1991	2359	2733	3169	3585	3969	4427	4920
jul	328	668	978	1302	1638	2000	2395	2783	3229	3673	4138	4606	5132
ago	303	633	980	1347	1732	2122	2536	2931	3364	3805	4206	4659	5190
sep	339	626	941	1285	1657	2002	2391	2806	3263	3743	4183	4643	5097
oct	283	607	937	1269	1627	2029	2436	2796	3255	3692	4178	4696	5174
nov	345	669	987	1282	1658	2009	2409	2772	3205	3651	4107	4558	5009
dic	336	670	994	1349	1718	2081	2474	2829	3259	3746	4167	4606	5090

Tabla 6: Nacimientos por meses acumulados de 1996 a 2008

Años desde 2003 a 2008 Lo comentado en el párrafo anterior se puede reproducir para las columnas de la Tabla 6 de los años que van desde 2003 a 2008.

La Figura 4 la hemos realizado agregando o acumulando los nacimientos de los años anteriores. Observamos por ejemplo como los nacimientos ocurridos en el mes de Febrero se quedan rezagados respecto a otros meses haciendo por ello el gráfico en cierto modo engañoso y que nos lleve a una conclusión errónea. Nota del profesor: *Realizar un gráfico que evite ese detalle, representando en vez del número de nacimientos acumulados, el número medio de nacimientos por mes acumulando.*

Actividades complementarias

► Realizar el contraste planteado en la expresión (5) con los datos acumulados en años sucesivos que figuran en la Tabla 6 de 1997 a 2008.

Hemos realizado el contraste de la actividad propuesta observando que los valores del estadístico χ^2 obtenidos, **son significativos desde el año 2003 al año 2008**. que corresponden a las columnas sombreadas de la Tabla 6.

El análisis de la figura 5 nos revela que la causa debe estar en los meses de Enero por presentar pocos nacimientos y Mayo por presentar muchos nacimientos

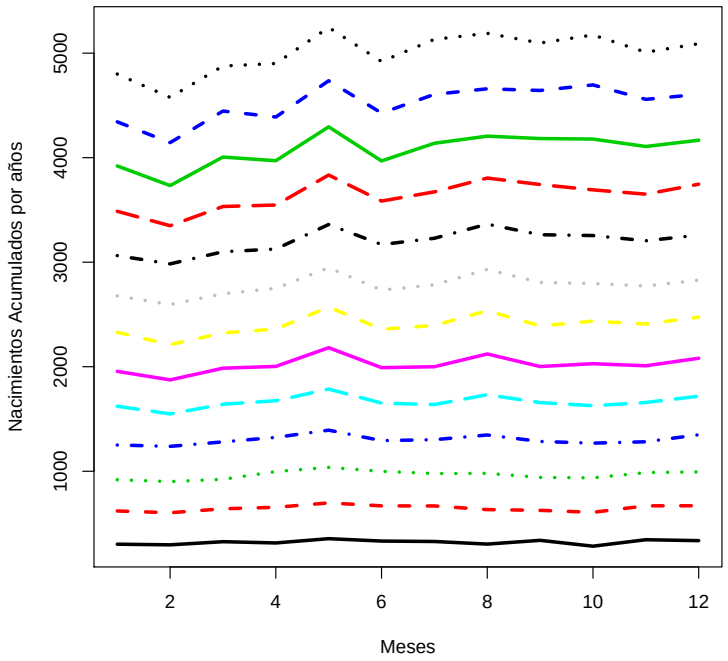

Figura 4: Nacimientos acumulados por años desde 1996 a 2008

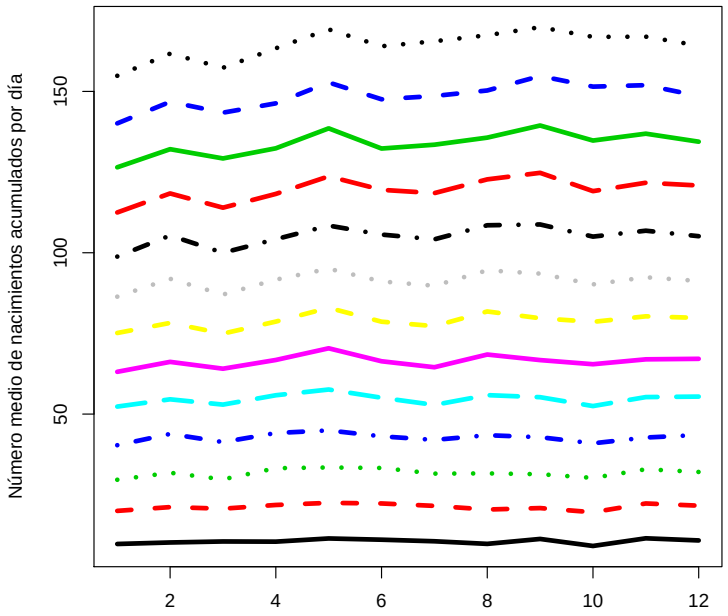

Figura 5: Promedios acumulados por años desde 1996 a 2008

Nota del profesor: *Agrupar el mes más bajo que es Enero junto con el mayor que es Mayo* y de nuevo acudimos a la expresión (5) para realizar de nuevo el contraste con los datos acumulados en años sucesivos que figuran en la Tabla 6 de 1997 a 2008, y concluimos que **agrupando los meses de Enero-Mayo no hay diferencias significativas en los nacimientos acumulados**.

6. Estudio en nuestro centro

Hemos trasladado el problema a nuestro centro con objeto de indagar si la aleatoriedad en el nacimiento mensual se mantiene. Para ello hemos realizado la recogida de datos de las fechas de nacimiento de los alumnos de nuestro centro nacidos en la propia localidad con objeto de detectar influencias locales. Los datos recogidos se muestran en la tabla inferior.

Mes	ene	feb	mar	abr	may	jun	jul	ago	sep	oct	nov	dic
Alumnos	23	13	24	17	26	17	23	21	21	13	24	27

Tabla 7: Distribución de nacimientos por meses en nuestro centro.

Realizamos el contraste de acuerdo a la expresión (5) y obtenemos el valor $\chi^2 = 10{,}3289 < \chi^2_{0,99,11} = 24{,}72497$, y concluimos que **no hay diferencia significativa en la distribución de los nacimientos por meses de los alumnos de nuestro centro**.

7. Conclusiones

Hemos ido sacando conclusiones y exponiéndolas a medida que hemos realizado los sucesivos contrastes de las actividades.

Como conclusión final podemos decir que los niños nacidos en cada año en Cantabria lo hacen aleatoriamente de acuerdo al modelo del dado-dodecaedro y por tanto el análisis realizado para cada año en particular, no detecta la presencia de causas que hagan determinados meses más relevantes en la distribución de los nacimientos por meses.

Sin embargo, después de 7 años, a medida que con los años se acumulan los nacimientos, se pone de manifiesto una relevancia significativa de los meses de Enero y Mayo, de forma que a lo largo de 13 años con un total de 60.013 nacimientos, se presenta que:

- Hay una ligera tendencia a la baja en el mes de Enero.
- Hay una ligera tendencia al alza en el mes de Mayo.
- El resto de los meses podemos considerarlos como equiprobables.

A. El contraste Chi-Cuadrado

A.1. Contraste para k proporciones

Supongamos que $x_1, x_2, \ldots, x_k$ son v. a. con distribución binomial de parámetros respectivos n_1 y p_1, n_2 y p_2, y n_k y p_k respectivamente. Si las n son suficientemente grandes, cada

$$z_i = \frac{(x_i - n_i p_i)^2}{n_i\, p_i(1-p_i)} \quad i = 1, 2, \ldots, k$$

será normal estándar y si son independientes

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - n_i\, p_i)^2}{n_i\, p_i(1-p_i)}$$

corresponde a una dist. χ^2 con k grados de libertad.

Muestras	Exitos	Fracasos
muestra 1	x_1	$n_1 - x_1$
muestra 2	x_2	$n_2 - x_2$
...	...	...
muestra i	x_i	$n_i - x_i$
...	...	...
muestra k	x_k	$n_k - x_k$

Así, para efectuar el contraste

1.

$$\begin{cases} \mathbf{H}_0 \equiv p_1 = p_2 = \ldots = p_k = p_0 \\ \mathbf{H}_1 \equiv p_i \neq p_0 \text{ algún } i \end{cases}$$

2. Si p_0 es conocido, la discrepancia viene dada por el estadístico

$$d = \chi^2 = \sum_{i=1}^{k} \frac{(x_i - n_i\, p_0)^2}{n_i\, p_0(1-p_0)} \tag{7}$$

donde d se ajusta a una distribución χ^2 con k grados de libertad. Y si p_0 no es conocido, se utiliza el estimador $\hat{p}_0 = \dfrac{x_1 + \ldots + x_k}{n_1 + \ldots + n_k}$, siendo

$$d = \chi^2 = \sum_{i=1}^{k} \frac{(x_i - n_i\, \hat{p}_0)^2}{n_i\, \hat{p}_0(1-\hat{p}_0)} \tag{8}$$

y d ahora se ajusta a una distribución χ^2 con $k-1$ grados de libertad.

3. Fijado α, si $|d| > \chi^2_{1-\alpha,g.l.}$ se **rechaza** la hipótesis H_0.

A.2. Test de ajuste de distribuciones

Este test es un contraste de significación para saber si los datos de una muestra son conformes a una ley de distribución teórica que sospechamos que es la correcta.

Dada una muestra aleatoria simple (x_1, x_2, …, x_n) de X podemos determinar la frecuencia absoluta de cada uno de los k resultados posibles en dicha muestra (es decir, el número de veces que ha ocurrido cada resultado). Designemos por n_1, n_2, …, n_k a estas frecuencias absolutas. La distancia χ^2 entre la distribución de frecuencias observada en la muestra y la distribución de probabilidad especificada por la hipótesis nula se define como

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \tag{9}$$

donde n_1, n_2, …, n_k son las frecuencias absolutas de los k posibles resultados y p_1, p_2, …, p_k son las probabilidades de dichos resultados si es cierta la hipótesis nula.

La distancia χ^2 definida por la ecuación (9) es una distribución χ^2 de Pearson con $k-1$ grados de libertad, es decir $\chi^2(k-1)$.

Referencias

[1] Devore, Jay L., *Probabilidad y Estadística*. Ed. Thomson. 6º Ed. 2005.

[2] Doran, Jody L., *Las Matemáticas en la Vida Cotidiana*. Ed. Addison-Wesley. 3º Ed. 1999.

[3] Engel, Arthur., *Probabilidad y Estadística*, Mestral Universidad, 1998.

[4] Peña, Daniel, *Estadística: Modelos y Métodos (Vols. I y II)*, Alianza Editiorial, 1987.

[5] Tanur, Judith M., *La Estadística. Una guía de lo desconocido.*, Alianza Editorial, 1992.