NOTAS Y PRÁCTICA 2: búsqueda y formulación de Leyes Estadísticas

**Problema 1** *(Leyes de permanencia)*

*i) Busque en https://www.estadisticaciudad.gob.ar/eyc/?p=50852, la cantidad de nacimientos de varones y mujeres ocurridos en los últimos lustros en la ciudad de Buenos Aires. Investigue la existencia o no de algún tipo de regularidad estadística Construya tablas, histogramas, funciones, etc.*

*ii) De haber "encontrado" algún tipo de regularidad, formule un modelo probabilístico que dé cuenta del tipo de comportamiento que tienen los nacimientos en Bs. As., es decir, describa el espacio muestral y su función de probabilidad.*

OBSERVACIÓN 1: *del experimento al modelo*
Como habrá notado, en la tabla de nacimientos de la ciudad de Buenos Aires aparece una regularidad estadística: la proporción de nacimientos entre hombres y mujeres se mantiene alrededor de un índice. Compare con sus compañeros si los respectivos índices coinciden o no lo hacen y si encuentran diferencias analicen a qué se deben. Esta regularidad se llama *permanencia estadística*. ¿Pero es exclusiva de la ciudad de Buenos Aires?, podemos decir que no (busque en otras ciudades, países y confirme o no la respuesta dada). La proporción de nacimiento entre varones y mujeres se encuentra, próxima a 1 pero no llega a ser 1 (imagine cómo sería el proceso de establecer un nuevo índice que dé cuenta no sólo de lo que ocurra en nuestra ciudad). Asumamos que la proporción es igual a

$$\frac{\text{número de varones}}{\text{número de mujeres}} = 1{,}05$$

Este índice es nuestro modelo matemático y la función de probabilidad, segundo modelo, se basa en él.

OBSERVACIÓN 2: *Esquema de la urna*
Podemos figurarnos que el sexo se determina por la extracción de una bolilla de una urna que contiene, por ejemplo, 205 bolillas de las cuales 100 han sido marcadas como "mujeres" y 105 como "varones". Después de cada extracción, la bolilla se repone en la urna, cuyo contenido es cuidadosamente mezclado, de modo que todas las extracciones se realicen en condiciones idénticas. Esta interpretación lleva el nombre de *esquema de la urna* y se extiende a caso más complicados, en los cuales el fenómeno estudiado no presenta sólo dos posibilidades. Es suficiente imaginar una urna que contenga diversas categorías de bolillas marcadas de modo distinto.

OBSERVACIÓN 3: *del modelo a la ley*
A la *frecuencia* (noción experimental), se la ha sustituido por la idea de *probabilidad*. Esta operación está justificada por la *ley de los grandes números* (debida a Jacques Bernoulli $[1654-1705]$) que dice: Si en una prueba la probabilidad de un acontecimiento o suceso que llamamos A, es $p_A$ y si éste se repite una gran cantidad de veces $n_A$, la relación entre las veces que se produce el suceso y la cantidad total de pruebas N, es decir, la frecuencia $f_A = n_A/N$ del suceso, tiende a acercarse cada vez más a la probabilidad $p$, ¿en qué sentido tiende?, lo expresamos de la siguiente manera:

si el número de pruebas es "suficientemente grande", "resulta totalmente improbable" que la diferencia entre $f_A$ y $p$ supere cualquier valor prefijado por pequeño que sea. En símbolos:

$$\lim_{n\to\infty} P(|f_A - p_A| < \varepsilon) = 1$$

**Problema 2** *Del modelo a la simulación de experimentos*
*Asumiendo la existencia del modelo probabilístico correspondiente a la tirada de una moneda no cargada (modelo extraído de la realidad), podremos con ayuda del computador simular experimentos más complejos.*

*i) Con la sentencia "RANDBETWEEN" genere una tira de 10 elementos aleatorios formada por ceros y unos. El cero representa la cara de una moneda y el 1 la ceca. Confiando en que el azar existe y que la computadora lo tiene en su poder, realice un histograma que muestre los resultados de este experimento.*

*ii) Haga lo mismo con 100 tiradas.*

*iii) Idem con 1000 tiradas*

*iv Idem con 10.000 tiradas.*

*v) Alguien afirma que en una tirada de 1000 veces de una moneda el resultado que obtuvo fue de 2 caras y 998 cecas, ¿está mintiendo esta persona? ¿Ud. le creería? Volveremos sobre esta pregunta en el problema 4.*

**Problema 3** *Extendiendo el campo de aplicaciones*
*Realice la simulación de tirar un dado de seis números con peso repartido homegéneamente*

*i) con todas sus caras iguales*

*ii) ligeramente alargado (elabore una estrategia computacional)*

*iii) ligeramente más alargado aún (modifique la estrategia anterior)*

**Problema 4** *Un segundo modelo: La ley binomial*
*Conociendo la probabilidad de un acontecimiento, podremos conocer la eventualidad de experimentos más complejos.*
*Con Excel construir las distribuciones de probabilidades del las siguientes variables aleatorias (utilizar la función COMBIN para obtener el combinatorio (50,n), con $n = 1, 2, 3...N$):*

*i) se tira una moneda no cargada 5 veces*

*ii) se tira una moneda no cargada 50 veces*

*iii) se tira una moneda no cargada 500 veces*

*iv) Responda con más argumentos a la afirmación hecha en el item v) del problema 2, con ayuda de la ley de los grandes números.*

OBSERVACIÓN 4:
El cálculo de probabilidades presta a la estadística la ayuda de sus métodos deductivos. En el problema anterior el modelo creado nos conduce a la *ley binomial*. En general:
Si se realiza $n$ veces un experimento, cada vez en forma independiente [1] y cada vez con igual binomio de resultados posibles, es decir, éxito y fracaso [2], la probabilidad de que ocurran exactamente $k$ éxitos puede calcularse mediante la siguiente fórmula:

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

[1] En el caso de la moneda, el resultado de una tirada no afecta a la siguiente tirada
[2] cara y ceca en la moneda

Comprobar que es una distribución de probabilidad, calcular su media y desvío típico.
Se deja como tarea realizar el estudio de la ley de Poisson
OBSERVACIÓN 5: Retomemos el problema 4. Los resultados de la experiencia de tirar una moneda $N$ veces plasmado un su correspondiente histograma muestra hasta qué punto serán poco probables o poco frecuentes en una gran cantidad de jugadas los sucesos muy diferentes del suceso más probable (relea el problema 2 y la ley de los grandes números).

**Problema 5** *Teorema del Límite Central*
*Una compañía tiene representantes de venta en todo el país. El número de unidades que el último mes vendió cada representante resultó:* $2, 3, 2, 3, 3, 4, 2, 4, 3, 2, 3, 4, 5, 3, 3, 3, 3, 5, 2, 7$

*i) Graficar en un histograma el comportamiento de esta población. Evaluar la media y desvío típico.*

*ii) Tomar 20 muestras aleatorias de 5 elementos cada una de la población. Evaluar el promedio de cada muestra y graficarlos en un histograma, calcular la media (de las medias) y desvío típico. Comparar con el promedio poblacional.*

*iii) Tomar ahora veinte muestras de 10 datos cada una, repetir los cálculos y la comparación del item anterior*

OBSERVACIÓN 6:
Este teorema afirma que la distribución de medias muéstrales tiende hacia una distribución normal, aunque las muestras procedan de una distribución no normal. Veamos cómo proceder en general

1° De una población podemos extraer una muestra aleatoria de n sujetos. Muestra aleatoria quiere decir que todos los sujetos de la población han tenido en principio la misma oportunidad de ser elegidos

2° De esta muestra podemos calcular la media. Seguimos extrayendo muestras aleatorias y calculando sus medias.

3° Al disponer de un número grande de medias tendríamos una distribución de estas medias; esa distribución es una distribución muestral: no se trata de una distribución de puntuaciones individuales sino de medias de muestras. Un punto importante es que aunque las muestras no tengan una distribución normal, las medias de estas muestras sí tienden a seguir la distribución normal

4° La desviación típica de estas distribuciones muestrales se denomina error típico y se puede estimar a partir de los datos de una muestra. Por lo tanto un error típico es la desviación típica de una distribución muestral, y se interpreta como cualquier desviación típica.

Dos distribuciones muestrales, con sus errores típicos, nos van a interesar de manera especial:

1. la distribución muestral de las medias
2. la distribución muestral de las diferencias entre medias de la misma población

Estas distribuciones muestrales son modelos teóricos que a partir de los datos de una muestra nos van a permitir inferir conclusiones acerca de la población a la que pertenece la muestra. Conociendo el error típico de estas distribuciones podemos estimar entre qué limites se encuentra la media de la población o si dos muestras proceden de poblaciones distintas con media distinta.

La media de una muestra podemos interpretarla como una estimación (solamente una estimación sujeta a error) de la media de la población. Esta estimación será más precisa:

1. Si la muestra es aleatoria porque en ese caso representa mejor las características de la población
2. Si la muestra es grande (si la muestra comprendiera a toda la población tendríamos el dato exacto, no una estimación).

El error típico, como es la desviación típica de todas las posibles muestras de esa población, nos va a permitir localizar entre qué límites se encuentra la media de la población. Este planteamiento es semejante al que nos encontramos en los sondeos de opinión, como son las encuestas pre-electorales. Si el 48 % de los sujetos entrevistados dice que va a votar a un determinado candidato, esto no quiere decir que el 48 % exacto de la población le vaya a votar. Sin embargo los datos obtenidos de una muestra nos van a permitir estimar un tanto por ciento mínimo probable y un tanto por ciento máximo probable de votantes a ese candidato: entre esos dos tantos por ciento se va a encontrar el tanto por ciento definitivo cuando todos hayan votado. De los datos de una muestra extrapolamos a la población, por eso se trata de estadística inferencial.

Por último, el T.L.C. vale si la población tiene media y varianza finitas.

**Problema 6**
*Asumiendo que*

$$f(x) = \frac{1}{\pi\gamma\left[1 + \frac{x-x_0}{\gamma}\right]^2}$$

*es una función de distribución de probabilidad (llamada distribución continua de Cauchy)*

*i) Grafíquela con $x_0 = 0$ y $\gamma = 1$*

*ii) En el mismo sistema incorpore la gráfica de $N(0, 1)$*

*iii) Conjeture sobre la media y la varianza de la distribución de Cauchy*

Continuando con el teorema central del límite, una de las formulaciones es:
Si $X_1, ..., X_n$ son variables aleatorias independientes con media $\mu_i$ y varianza $\sigma_i^2$ , independientemente del tipo de distribución que sigan las variables $X_i$, la suma de todas ellas, $Y = X_1 + ... + Xn$ tiende a distribuirse aproximadamente a una normal, con media $\mu = \mu_1 + ... + \mu_n$ y y varianza $\sigma^2 = (\sigma_1^2 + ... + \sigma_n^2)/n$, siendo las aproximaciones mejores a medida que aumenta $n$.

Dado que la variable Binomial, no es más que la suma de $n$ variables independientes, su distribución tienda a aproximarse a la normal a medida que aumenta n, con media igual a la esperanza $E(X) = np$ y varianza $\sigma^2 = np(1p)$, resultado demostrado por De Moivre en 1733. Este autor encontró que si $X$ es una variable $B(n, p)$, la distribución:

$$\frac{X - np}{sqrtnp(1-p)}$$

converge hacia una distribución normal con media 0 y desvío estándar 1.

**Problema 7** *: T.C.L.*
*Dada $B(n, 1/6)$*

*i) Busque ayudándose con la computadora a partir de qué valor de n la distribución se parece a la normal.*

*ii) Algunos autores establecen una cota mínima a partir de la cual aproximan $B(n, p)$ con una normal cuando se se cumple la condición $np(1p) > 5$, otros más exigentes piden $np(1p) > 9$. Compare su resultado con estos puntos de vista.*

**Problema 8** *Usando el T.C.L. estime la probabilidad de sumar más de 200 puntos al lanzar 70 veces un dado no cargados.*
*Sugerencias:*

1. *Asumamos la condición $np(1p) > 9$ para decidir cuándo será confiable aproximar la binomial por una normal. Verifique entonces que dicha condición se cumple.*

2. *Calcule* $E(X) = X_1 + ... + X_{70}$ *y* $\sigma^2 = (\sigma_1^2 + ... + \sigma_{70}^2)/n$.

3. *Por el T.C.L. la variable aleatoria "puntos obtenidos en 70 lanzamientos", es decir,* $B(70, 1/6)$ *se aproxima a una normal* $N(E(X), \sigma)$. *haciendo un cambio de variables Calcule* $P(x > 200)$.