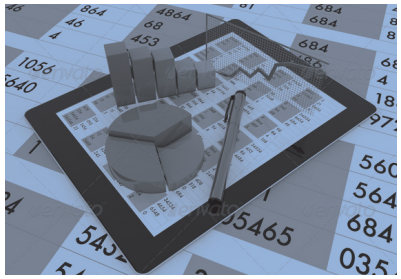
Pre•Impresos 7
Estudiantes

Facultad de Ciencia y Tecnología Departamento de Matemáticas - 2012

Respuesta múltiple en datos categóricos.
Una prueba de hipótesis

Wilson Hely Rodríguez Gámez

UNIVERSIDAD PEDAGOGICA NACIONAL
Educadora de educadores



Pre·Impresos Estudiantes 7

Diseño y Preparación editorial
Universidad Pedagógica Nacional
Fondo Editorial
2012

Pedro José Román
Coordinador Fondo Editorial

Impreso por Javegraf
Bogotá, Colombia

Respuesta múltiple en datos categóricos.
Una prueba de hipótesis

Presentación

De acuerdo con la Misión de la Universidad Pedagógica Nacional, el objetivo de esta publicación es resaltar la importancia de la socialización de las ideas en el campo de las ciencias y su enseñanza; como contribución al fortalecimiento de la docencia y la investigación en educación, tendiendo puentes entre los saberes especializados y la cultura en general. Esto resulta pertinente y significativo en la formación de nuevas generaciones de maestros de ciencias e investigadores en pedagogía, desde campos disciplinares específicos, quienes en su futura práctica profesional afrontarán retos y circunstancias diversos, que el entorno social del país le plantea a la educación.

La serie Pre-Impresos es una iniciativa editorial dirigida a la comunidad académica en general, que divulga la producción intelectual de los estudiantes de la Facultad de Ciencia y Tecnología de la UPN, destacando las experiencias y reflexiones respecto de los temas propios del quehacer disciplinar y pedagógico de los autores. La participación está abierta a todos los integrantes de la comunidad estudiantil que deseen publicar sus trabajos a través de este medio; no hay restricción alguna en cuanto al formato de escritura, número de páginas o tema, con la salvedad de aquellos que estén fuera de los intereses propios de la actividad de la Facultad.

Información:

pre_impresos@pedagogica.edu.co
jcbustos@pedagogica.edu.co
Departamento de Física - UPN
Teléfonos: (57) (1) 3471190 / 5941894 Ext. 242

Respuesta múltiple en datos categóricos. Una prueba de hipótesis

Wilson Hely Rodríguez Gámez[1]
jumbt@hotmail.com;
edma_wrodiguezg992@pedagogica.edu.co

Resumen

Se presenta *una prueba de hipótesis* para casos de *respuesta múltiple en datos categóricos*; en primer lugar, se explican las nociones básicas necesarias en este tipo de pruebas (dato, datos categóricos, respuesta múltiple, prueba de hipótesis, etcétera), con el fin de entender mejor el entorno o contexto. Luego se toma un ejemplo de aplicación, donde se repasan los conocimientos prestablecidos. Desde la perspectiva de la estadística y, en especial, de la inferencia como parte de esta ciencia, se incluye la *estadística de prueba*[2] propuesta por Agresti y Liu. En la parte final de este artículo se consignan varios anexos –citados en el desarrollo del documento–, en los que se abordan los distintos temas tratados, con el ánimo de complementar la información.

Palabras claves: respuesta múltiple, hipótesis, rechazar, distribución, vector, datos categóricos, configuración, prueba.

Abstract

We present a hypothesis test for cases of multiple response categorical data, first, explains the necessary basic knowledge in these tests (data, categorical data, multiple choice, hypothesis testing, so on...), to better understand the environment or context. Then take an example application where knowledge presets are reviewed. From the perspective of statistics and especially of inference as part of this science, it includes the test statistic proposed by Agresti and Liu. At the end of this article are listed several annexes, cited in developing the document, in them are on different topics with depth, with the aim of complementing the information.

Key words: Multiple response, hypotheses, reject, distribution, vector, categorical data, configuration, testing.

1 Estadístico de la Universidad Nacional de Colombia. Estudiante de la especialización en Educación Matemática de la Universidad Pedagógica Nacional. Docente en ejercicio. E_mail: whrodriguezg@unal.edu.co, jumbt@hotmail.com, edma_wrodiguez992@pedagogica.edu.co

2 Concepto a especificarse en el desarrollo del trabajo.

Introducción

Al analizar información es usual que haya dudas de como hacerlo, por razones diversas, generando inseguridad y desubicación, esto conduce a errores por acción u omisión. Cuando esto ocurre se hace necesario contar con una metodología sencilla y clara que permita hacer uso de esta información, para a través de la misma –en este caso– poder tomar decisiones. Se ha advertido que en muchos estudios estadísticos se contemplan situaciones que generan respuestas simples únicamente, esto es, en tales estudios existen preguntas que conducen a únicas respuestas. No siendo común los casos en los cuales las respuestas puedan ser múltiples. *En este escrito se tratará un caso donde se encuentra respuesta múltiple junto con datos categóricos*[3], situación que permitirá mostrar algo de esta combinación (referente al tratamiento o su análisis). En particular interesa realizar una *prueba de hipótesis*, entendiéndose como tal "juzgar estadísticamente si cierta propiedad supuesta para una población[4] es compatible con lo observado en una muestra[5] de ella" (Peña, 1988, p.195). En consecuencia, apoyados en la inferencia estadística, se realizará paso a paso el proceso de prueba, al amparo de la metodología existente y prescindiendo de lenguajes excesivamente técnicos y expresiones matemáticas complejas.

En las primeras secciones se abordan los conceptos básicos para contextualizar el entorno de las pruebas de hipótesis, al igual que la respuesta múltiple en datos categóricos[6]. Luego se desarrolla un ejemplo –en una de las últimas secciones- donde se puede aplicar y validar los conceptos. A la par que se despliegan los distintos temas, se implementa una "alfabetización" como parte de una *cultura estadística*[7] que, según Batanero (2002), debe entenderse como contribuir al conocimiento, las destrezas adquiridas, el razonamiento, la intuición, la actitud, etcétera. Razón por la cual, el primer objetivo en este trabajo es explorar y enseñar con sencillez una metodología propia en estos casos, el segundo objetivo es ilustrar como a partir de *pequeñas* muestras se puede generalizar una(s) conclusión(es) a toda una población.

3 Tipo de datos definido en alguna de las secciones siguientes
4 Conjunto de todas las unidades de estudio.
5 Hará referencia de aquí en adelante a una parte de la población seleccionada por muestreo probabilístico.
6 Desde la importancia que tiene un dato o un conjunto de ellos y pasando por las maneras que son tratados, hasta llegar a su análisis y algunos hallazgos.
7 Ver más de cultura estadística en el anexo 1.

La relevancia de un dato y algunos tipos de datos

Naturaleza-hombre constituye un conjunto de múltiples relaciones, el hombre ha encontrado allí objetos y entidades. Estos tienen características o atributos que pueden ser observables o discernibles, es decir, se pueden ver directa o indirectamente a través de uno o varios pasos adicionales. Por ejemplo, se ve el color de los ojos de una persona directamente, pero no se puede ver de la misma manera, la velocidad a que va un vehículo en movimiento, aquí se ha de medir el espacio y el tiempo para obtener la velocidad (por discernimiento con dos medidas). Las características o atributos generalmente son registrados por procesos de medición (altura, peso, volumen, etcétera) en la mayoría de los casos, pero también son registrados por las cualidades (color, contextura, parecer, estado, entre otras). En este sentido, hay objetos, entidades y alguien que da información de ellos; lo que arriba se llamó los atributos o características.

Lo anterior lleva a definir *dato* como *aquello que da información de un objeto o entidad –en cuanto a sus características o atributos–*. En estadística no siempre los datos son cuantitativos, también se tienen cualitativos, la diferencia esencial entre estos dos tipos, es que mientras los primeros dan cuenta numérica (cuantifican) de algún objeto de estudio, los segundos hablan principalmente de cualidades del mismo, de allí los respectivos nombres. Los datos cualitativos para algunos efectos y en especial en este escrito son llamados *datos categóricos*, en el entendido de que una cualidad puede formar un(os) grupo(s) que llamaremos *categoría(s)*. Vale la pena decir también que los datos cuantitativos –aunque no sean categóricos– los podemos categorizar (Díaz, 2009), por ejemplo, podemos categorizar el tiempo de un día en un par de categorías: mañana y tarde.

En sentido amplio, si se tienen varios objetos, se puede extraer el mismo dato de cada uno de ellos, pero como ese dato (en la mayoría de los casos) no es igual para todos los objetos; aparece la noción de variable, *valor(es)* o *categoría(s)* que puede(n) *tomar un dato, al ser registrado para distintos objetos*. Por consiguiente, se llamaran variables numéricas a las procedentes de datos cuantitativos y variables categóricas aquellas que provengan de datos categóricos o cualitativos.

Procedencia, recolección y análisis de los datos

Visto que un solo dato, a pesar de su importancia, es poco significativo estadísticamente, pues, de él no se puede extraer mucho ni sacar inferencia alguna, se debe intentar contar con varios de ellos y, a la vez, comenzar a hacer un esbozo de análisis sobre estos. Antes de proyectar algún procedimiento de análisis, unas de las preguntas obligatorias y pertinentes que surgen son:

- ¿De donde provendrán los datos?
- ¿Cómo se recolectarán los datos?
- ¿Qué hacer con los datos?

Con respecto al primer interrogante, tenemos claro (por lo visto en la sección anterior), que los datos provienen de objetos o entidades, estos a su vez serán de aquí en adelante nuestras unidades de estudio. El conjunto de todas las unidades de estudio es lo que se llama *población,* universo o colectivo, el cual es o será el referente, él se conforma debido a un interés particular que se tenga en sus elementos.

Sin embargo –y pasando al segundo interrogante–, en la mayoría de casos no se puede estudiar toda la población por diversas razones, entre estas se tienen: gran tamaño, viabilidad, factibilidad, costos, etcétera. Razón(es) por la(s) cual(es) es preciso estudiar "pedazos", "porciones", esto es, *muestras* que sean inferiores en tamaño a la población misma, pues, no tendría ningún sentido si no fuera así. La finalidad de una muestra de acuerdo con Canavos (2001), aparte de ser de donde se recolectaran los datos, es que sea representativa de la población y de ella se infieran algunas "cosas"[8]. Las conclusiones que se obtengan en la muestra, deben ser susceptibles de generalizarse o extenderse a la población (figura 1). Por tanto, la extracción y el tamaño de una muestra de una población obedece esencialmente a varios aspectos, entre estos:

- Que sea obtenida con un método aleatorio (totalmente imparcial).
- Que todas las posibles muestras en la población tengan la misma probabilidad de ser seleccionadas (probabilidad de selección).
- Que todas las unidades de estudio tengan la misma probabilidad de ser incluidas en las distintas muestras (probabilidad de inclusión).
- Que para establecer el tamaño de tal muestra, se debe tener en cuenta el tamaño de la población, la variabilidad de la misma y una estimación del error entre otros.

En resumen, aspectos como los anteriores y otros adicionales solo se consiguen con los procedimientos denominados en estadística *muestreo probabilístico*[9]. De manera que las poblaciones pueden ser muchas, pero en un estudio en particular deben ser específicas y únicas en lo posible, pues de ellas serán las muestras. En cuanto a la última de las preguntas surgidas, solo se dice que son una gran cantidad de cosas las que se pueden hacer con los datos: descripción, exploración, inferencia, pronósticos, diagnósticos, entre otras, es decir, diversos tipos de análisis estadísticos. Sin embargo, luego de haber tratado los interrogantes planteados, es necesario tener en cuenta a la hora de obtener y analizar datos el criterio o recomendaciones de los investigadores o profesional en estadística, de esta manera se facilitará un poco la recolección y configuración, consecuentemente se hace más sencillo su posterior análisis.

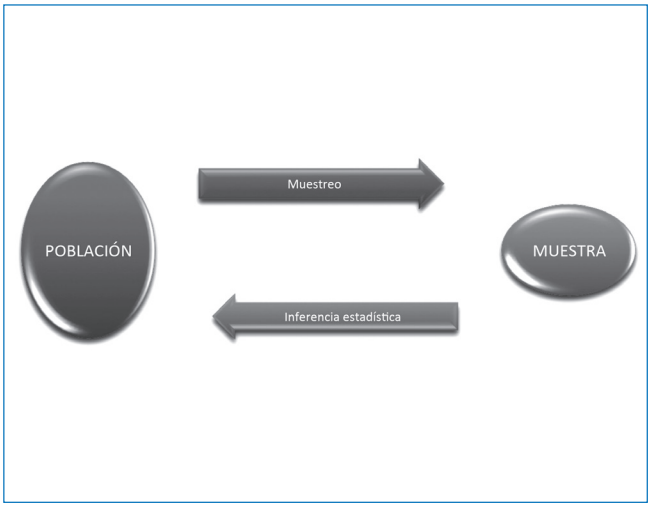


Figura 1. Muestreo probabilístico

8 Por lo general, es acerca de los parámetros o la distribución de la población.

9 Parte del muestreo en estadística.

Configuración de los datos

La configuración de los datos una vez obtenidos, será lo mismo que lo señalado por los estadísticos como *construcción de la base de datos*. La palabra configuración debe entenderse como un *sistema de recolección de información*. Por ejemplo, una base sencilla y muy personal es el documento de identidad, pues en él está configurada información como: nombres, apellidos, fecha de nacimiento, fecha de expedición del documento, una huella digital y otros datos; así tenemos una pequeña base de datos en cada documento de identidad.

La función principal de una base de datos es el almacenamiento de la información y el ordenamiento de la misma, por tanto, se convierte en una herramienta valiosa –para los pasos posteriores en un estudio– (figura 2). Se puede actualizar la información de manera permanente, buscarla con rapidez, copiarla con facilidad: son tres aspectos adicionales muy útiles en las bases, además tienen diferentes presentaciones, se vio en el ejemplo anterior que se presenta como un documento de identidad y, se tienen algunas otras más como:

- Tablas: son configuraciones dispuestas en filas y columnas.
- Formularios, por ejemplo, donde se registran las respuestas a unas preguntas provenientes de una encuesta o registros de consulta.
- Informes, especialmente en sus encabezados son ricos en datos.
- Tarjetas de registros de rendimiento deportivo e inventarios.

En este documento se hace uso solo de las tablas, pues, por su naturaleza, son idóneas en estos casos.

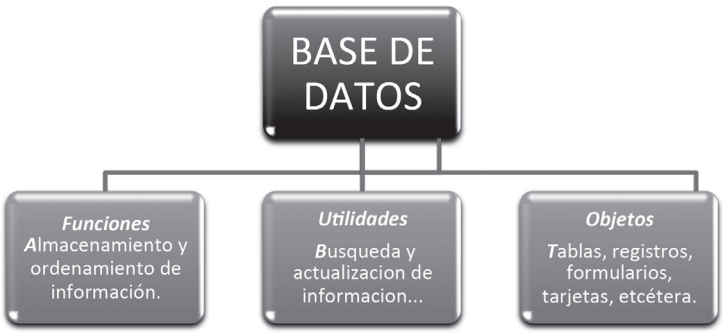


Figura 2. Estructura de una base de datos

Prueba, contraste o juzgamiento de hipótesis

Una forma de concluir *algo*[10] acerca de una población, es a través de los procedimientos llamados en estadística pruebas, contrastes o juzgamiento de hipótesis. Estos procedimientos son en esencia una afirmación (*hipótesis principal o nula, Hp*) sobre la población, luego buscar evidencia estadística obtenida en la muestra, para posteriormente de alguna manera decidir si esa afirmación es rechazada o no[11]. En caso de ser rechazada la hipótesis principal, se cuenta con otra afirmación (*hipótesis alterna, Ha*), que con cierta regularidad es la negación de la hipótesis principal, la cual será aceptada en la eventualidad de que Hp sea rechazada. De este modo, normalmente se tiene una hipótesis principal y otra alterna, las cuales se conocen como sistema de hipótesis. Al respecto, dos conceptos de otros autores: "Una hipótesis estadística es una aseveración o conjetura acerca de la distribución de una población…" (Mayorga, 2004, p. 189).

> Una hipótesis estadística es una afirmación con respecto a alguna característica desconocida de una población de interés. La esencia de probar una hipótesis estadística es decidir si la afirmación se encuentra apoyada por la evidencia experimental que se obtiene a través de una muestra aleatoria. En forma general, la afirmación involucra ya sea a algún parámetro o a alguna forma funcional no conocida de la distribución de interés a partir de la cual se obtiene una muestra aleatoria. La decisión acerca de si los datos muéstrales[12] apoyan estadísticamente la afirmación se toma con base en la probabilidad, y si esta es mínima, entonces será rechazada. [Canavos, 2001, p. 303].

En el proceso de decidir si la hipótesis principal es rechazada o no, se precisa de la evidencia estadística relacionada anteriormente, mejor conocida como *estadística de prueba*[13]. Dicha estadística tiene por naturaleza una distribución[14] teórica o exacta, que

10 Se refiere en especial a planteamientos estadísticamente coherentes.
11 Recordemos que según lo visto con anterioridad las conclusiones en la muestra son susceptibles de generalizarse a la población de la cual procede
12 Se refiere a datos obtenidos en una muestra.
13 Expresión matemática, usada en la prueba, que no involucra los parámetros de una población, entiéndase por parámetro un valor fijo pero desconocido, por ejemplo, la varianza poblacional suponiendo que no se conoce.
14 Hace referencia a la distribución de las distintas probabilidades de los valores que toma una variable aleatoria.

se debe establecer, además, se debe comparar el valor de la estadística de prueba, con algún valor pertinente de su misma distribución teórica. Luego se verifica que la estadística de prueba –en cuanto su valor–, no supere el valor teórico (para algunos casos), si el valor de la estadística de prueba supera el valor teórico, se considera que hay evidencia estadística suficiente para rechazar la hipótesis principal a favor de la hipótesis alterna.

Aun cuando todo lo anterior se realice con rigor, persisten los siguientes riesgos:

- Rechazar una hipótesis principal, siendo cierta (*error tipo I*).
- No rechazar una hipótesis principal, siendo falsa (*error tipo II*).

De modo que, antes de realizar la prueba, se debe fijar o determinar un *nivel de significancia (α),* que se desee o determine, por ejemplo, α = 0,05; esto controlará la posibilidad de rechazar una hipótesis principal cuando sea cierta (error tipo I) y consecuentemente se obtiene un *nivel de confianza* del (1-α) 100%. De manera sucinta en la figura 3 se consignan los pasos para realizar una prueba de hipótesis.

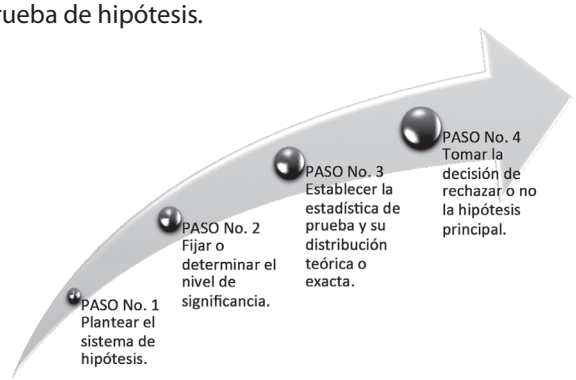


Figura 3. Pasos básicos para realizar una prueba de hipótesis

Del error tipo I y II

La tabla 1 contempla dos tipos de errores (I y II) cuando nos enfrentamos a tomar alguna decisión con respecto a una hipótesis principal, el tipo I es el error que se comete al rechazar la hipótesis cuando es cierta, en tanto que el tipo II es el cometido al no rechazarla cuando la hipótesis es falsa. Debemos notar que solo es posible cometer uno de estos dos errores, en efecto, si la hipótesis es realmente cierta solo estaremos en riesgo de cometer el error tipo I y si la hipótesis es falsa el riesgo es con respecto al error del segundo tipo.

	Hipótesis principal (Hp)	
Decisión	**Cierta**	**Falsa**
Se rechaza	Error I	Correcto
No se rechaza	Correcto	Error II

Tabla 1: Errores tipo I y II

Es necesario también tener una medida que dé cuenta del riesgo de cometer alguno de estos errores; esta medida debe ser una probabilidad, más exactamente una probabilidad condicional. Canavos (2001) escribe:

> *Definición 1*: la probabilidad de rechazar Hp dado que Hp es cierta, se define como la probabilidad (o tamaño) del error de tipo I y se denota por α, y se encuentra entre 0 y 1.
>
> *Definición 2:* la probabilidad de no rechazar Hp dado que Hp es falsa, se define como la probabilidad (o tamaño) del error de tipo II y se denota por β, y se encuentra entre 0 y 1 (p. 305).

Respuesta múltiple en datos categóricos

En algunos casos, la *respuesta* a una pregunta para un estudio puede ser *múltiple*, en otras palabras, quien responde puede elegir una o más opciones. Para ejemplificar esto, pensemos en qué se respondería si se pregunta: ¿qué mascota prefiere entre perros, gatos y canarios?, algunos responderían que perros, otros que gatos, y el resto que canarios. Sin embargo, se podría optar por responder que se prefiere dos de las tres categorías o las tres o ninguna, en fin, en total serian ocho posibles respuestas que se podrían dar. Matemáticamente se pueden expresar esas respuestas en vectores de tres componentes (triplas), si se asigna 1 a la preferencia de la mascota y 0 si no, para cada caso; de este modo, aquel que elige los perros podemos expresarlo como sigue (1,0,0), el que elige perros y gatos (1,1,0) y de manera análoga para las demás respuestas posibles. Pero si se condiciona a que la respuesta deba ser una y solo una, se tendrían únicamente tres posibles respuestas (1,0,0), (0,1,0), (0,0,1), esto ultimo se conoce como *respuesta*

simple. Además se puede estar interesado, en la comparación de grupos prestablecidos cuando se presenten este tipo de respuesta, es decir, establecer algún tipo de asociación entre los diferentes grupos.

Acerca del vector respuesta

Podemos notar el *vector respuesta*[15] (tabla 2, renglón resaltado) de una persona para varias categorías como sigue: $(X_{ih1},X_{ih2},X_{ih3},\ldots,X_{ihc})$, donde cada una de las componentes del vector tomará solo los valores 1 o 0, de manera muy universal cada componente será de la forma X_{ihj} y sus subíndices significan o denotan:

- *i* denota la persona que responde, así *i* puede tomar los valores 1, 2,…, hasta el número total de personas que haya, supongamos n_h personas en el grupo *h*.
- *h* denota el grupo a que pertenece la persona que responde, luego *h* puede tomar los valores 1, 2,…, hasta el número de grupos que haya, supongamos *r* grupos.
- *j* denota la categoría que elija o no la persona, entonces j puede tomar valores 1, 2,…, hasta la cantidad *c* de categorías que exista.

Entonces $X_{ihj}=1$ significa que: "La i-ésima persona, en el h-ésimo grupo *eligió* la categoría j-ésima", en tanto que sí $X_{ihj}=0$ significa que: "La i-ésima persona, en el h-ésimo grupo *no eligió* la categoría j-ésima" De esta manera se obtendrán vectores con solo unos o ceros por componentes y tantos como personas hayan respondido (tabla 2).

Grupo	Categoría 1	Categoría 2	…	…	…	Categoría C
1	.	.	.	.	.	.
	(Xi11	*Xi12*	……..	……	……..	*Xi1c)*
	.	.	.	.	.	.
.	.	.	.	.	.	.
	(Xih1	*Xih2*	…….	……..	……..	*Xihc)*
	.	.	.	.	.	.
r	.	.	.	.	.	.
	(Xir1	Xir2	……..	…….	…….	Xirc)
	.	.	.	.	.	.

Tabla 2. Base de datos general

15 Vector que contiene la respuesta de determinada persona, con solo 1 o 0 en cada componente, expresando la elección o no respectivamente.

Un ejemplo de aplicación

Para precisar mejor lo referente a la respuesta múltiple en datos categóricos y a la vez desarrollar una prueba de hipótesis, a continuación se explicará mediante un ejemplo práctico trabajado por Higgins (2005).

Un inspector en jardinería, luego de establecer una *muestra* (se supone obtenida por muestreo probabilístico) de 117 hogares de todos los hogares (*población*) en tres ciudades distintas (*grupos*) les pregunta: "¿Dónde adquieren los implementos para trabajar en su jardín?". Las tres posibles respuestas simples se citan a continuación.

a. En tiendas especializadas (G).
b. En tiendas de rebaja (D).
c. En supermercados (S).

Una vez configurada la información obtenida por parte del inspector, se tiene la siguiente base de datos que se muestra en la tabla 3.

CIUDAD	G (1,0,0)	D (0,1,0)	S (0,0,1)	G,D (1,1,0)	G,S (1,0,1)	D,S (0,1,1)	G,D,S (1,1,1)	TOTAL
1	5	10	6	3	2	6	1	33
2	13	7	3	8	4	3	2	40
3	5	13	10	2	3	9	2	44
TOTAL	23	30	19	13	9	18	5	117

Tabla 3: Base de datos

En la base (tabla 3), configurada a partir de los datos proveniente de los hogares en las tres ciudades luego de ser preguntados, se observa que las filas o renglones son las ciudades *(grupos)* y las columnas son las posibles respuestas que se dieron (para las tres *categorías* G, D, S y *algunas de sus combinaciones*). Hay que tener en cuenta, por ejemplo, que el número 5 en la primer celda (primera fila y primera columna) de la base significa que: "Cinco hogares en la ciudad No. 1 manifestaron adquirir los implementos para trabajar en su jardín en tiendas especializadas (G)", mientras que el número 8 (segunda fila y cuarta columna) en la misma base significa que: "Ocho hogares en la ciudad No. 2 manifestaron adquirir los implementos para trabajar en su jardín en tiendas especializadas (G) o en tiendas de rebaja (D)". La columna y fila resaltadas corresponden a los subtotales de cada fila y columna correspondiente, como se ha dicho, se tiene la *base de datos*, configuración en forma de tabla.

Ahora, si se supone que el inspector está interesado en establecer algún grado de asociación entre las tres ciudades, para diversos fines, pensemos que quiere establecer si tienen un grado de asociación fuerte o débil, en el primer caso –si la asociación resultara fuerte– será lo mismo el comportamiento de las tres ciudades entre sí; por el contrario, si el grado de asociación es débil, las tres ciudades serí an en esencia distintas, los dos casos anteriores con referencia a la adquisición de los implementos para el jardín. Se puede imaginar, también, que el inspector quiere proveer implementos para el jardín y desea saber si puede elegir cualquier ciudad para su domicilio. Recogiendo lo más importante, el sistema de hipótesis para este caso queda planteado así:

Hp: la *distribución de probabilidades*[16] de las tres ciudades es la misma.
Ha: la *distribución de probabilidades* de las tres ciudades es distinta.

Para llevar a cabo la prueba, se fija[17] un nivel de significancia de α = 0,05, por consiguiente, tendremos un nivel de confianza del (1-α) 100% = (1-0,05) 100% = 95%. A continuación, se toma como estadística de prueba (ϕ^2)[18] la sugerida por Agresti y Liu en 1999, que no es otra cosa que una suma de sumas. Ahora bien, para obtener ϕ^2 debemos tener antes los valores observados y esperados, de la *elección* y *no elección* de las categorías G, D y S, en las distintas ciudades.

CIUDAD	Categoría G		Categoría D		Categoría S	
	ELIG. G	NO ELIG. G	ELIG. D	NO ELIG. D	ELIG. S	NO ELIG. S
1	O = 11 E = 14,1	O = 22 E = 18,9	O = 20 E = 18,6	O = 13 E = 14,4	O = 15 E = 14,4	O = 18 E = 18,6
2	O = 27 E = 17,1	O = 13 E = 22,9	O = 20 E = 22,6	O = 20 E = 17,4	O = 12 E = 17,4	O = 28 E = 22,6
3	O = 12 E = 18,8	O = 32 E = 25,2	O = 26 E = 24,8	O = 18 E = 19,2	O = 24 E = 19,2	O = 20 E = 24,8
EST_ PRUEBA	15,5		1		5,2	

Tabla 4. Valores observados y esperados

La tabla 4, obtenida en su totalidad de la tabla 3, contiene o relaciona los valores observados (O) y los valores esperados (E) para las categorías G, D y S, respectivamente, en cuanto a la elección o no de las categorías, hechas por los distintos hogares. Así, por ejemplo, O = 11 = 5 + 3 + 2 + 1 en la primera celda de la categoría G, corresponde a el total de hogares en la muestra que eligieron G (ELIG. G) en la ciudad 1, sin importar que hayan elegido más categorías, en tanto que E = 14,1 = [33 × (23 + 13 + 9 + 5)] / 117 = (33 × 50) / 117 en la misma celda corresponde por definición de valor esperado, al cociente de multiplicar el subtotal que aparece para la ciudad 1, por la suma de los subtotales de los que eligieron G, sin importar la ciudad, entre el total de hogares en la muestra 117.

En la fila resaltada se relacionan los sumandos de la estadística de prueba; en consecuencia, como ya se dijo, la tabla 4 es extraída de la tabla 3, al realizar pasos análogos al anteriormente expuesto, para cada una de las tres categorías, junto con las tres ciudades. Luego el valor hallado para la estadística de prueba establecida corresponde a ϕ^2 = 15,5 + 1 + 5,2 = 21,7, donde, por ejemplo, 15,5 corresponde la siguiente suma en la categoría G.

$$15{,}5 = \frac{(11-14{,}1)^2}{14{,}1} + \frac{(22-18{,}9)^2}{18{,}9} + \frac{(27-17{,}1)^2}{17{,}1} + \frac{(13-22{,}9)^2}{22{,}9}$$

$$+ \frac{(12-18{,}8)^2}{18{,}8} + \frac{(32-25{,}2)^2}{25{,}2}$$

El valor de ϕ^2 obtenido se debe comparar con el valor[19] de su distribución teórica *ji*-cuadrado[20] (X^2_6) con 3 × (3-1) = 6 grados de libertad (por ser *c*(*r*-1) los grados de libertad, siendo para este caso c = 3 y r = 3, donde *c* son las tres categorías y *r* son los tres grupos o ciudades) al 95% de confianza, que equivale a X^2_6=12,6. Ahora, *como* ϕ^2 *es mayor que* X^2_6($\phi^2 > X^2_6$), *lo que nos indica que hay evidencia estadística para rechazar Hp en favor de Ha*. En efecto, la distribución de probabilidad de las ciudades no es igual, a lo menos dos son distintas. En conjunto, diremos que las ciudades en cuanto a las preferencias, para la adquisición de implementos para el jardín son *significativamente distintas.*

Asimismo y con respecto al inspector, este podrá concluir que no es lo mismo en todas las ciudades colocar una tienda de productos para el cuidado de los jardines, en lo concerniente a las preferencias de los hogares, pues estadísticamente son diferentes (recordemos que nos hemos quedado con la hipótesis alterna).

16 Hace referencia a la función conjunta de probabilidad de cada ciudad.
17 Recordemos que el nivel de significancia no solo se puede fijar, en ocasiones también se puede determinar.
18 Matemáticamente nuestra estadística de prueba es:

$$\phi^2 = \sum_{j=1}^{c}\sum_{h=1}^{r}\left\{\frac{(O_{hj1}-E_{hj1})^2}{E_{hj1}} + \frac{(O_{hj0}-E_{hj0})^2}{E_{hj0}}\right\}$$

donde O_{hj1}, E_{hj1} es el valor observado y esperado respectivamente, en el grupo *h* que eligieron la categoría *j*, lo mismo que O_{hj0}, E_{hj0} son los valores observados y esperados respectivamente, en el grupo *h* que *no* eligió la categoría j. ϕ^2 tiene distribución teórica *ji*-cuadrado ($X^2_{c(r-1)}$) con *c*(*r*-1) grados de libertad, siendo *c* el numero de categorías y *r* el numero de grupos.

19 Valor que se encuentra en tablas estadísticas de la distribución ji-cuadrado o se puede calcular en diferentes software como R.
20 Ver más de la distribución ji-cuadrado en el Anexo 2.

Anexo 1

La estadística como cultura

El Ministerio de Educación Nacional[21] (2006), como encargado de las políticas educativas en todo el territorio colombiano, contempla cinco tipos de pensamiento para la enseñanza de las matemáticas, estos son:

1. Pensamiento numérico y sistemas numéricos.
2. Pensamiento espacial y sistemas geométricos.
3. Pensamiento métrico y sistemas de medidas
4. Pensamiento aleatorio y sistemas de datos.
5. Pensamiento variacional y sistemas algebraicos y analíticos.

En el pensamiento 4 reposa la estadística y toda su metodología; en cuanto a los lineamientos curriculares es pertinente resaltar algunas apreciaciones del mismo MEN (1998) con respecto a dicho pensamiento.

> El pensamiento aleatorio y los sistemas de datos: Una tendencia actual en los currículos de matemáticas es la de favorecer el desarrollo del pensamiento aleatorio, el cual ha estado presente a lo largo de este siglo, en la ciencia, en la cultura y aun en la forma de pensar cotidiana. La teoría de la probabilidad y su aplicación a los fenómenos aleatorios, han construido un andamiaje matemático que de alguna manera logra dominar y manejar acertadamente la incertidumbre. Fenómenos que en un comienzo parecen caóticos, regidos por el azar, son ordenados por la estadística mediante leyes aleatorias de una manera semejante a como actúan las leyes determinísticas sobre otros fenómenos de las ciencias. Los dominios de la estadística han favorecido el tratamiento de la incertidumbre en ciencias como la biología, la medicina, la economía, la psicología, la antropología, la lingüística..., y aún más, han permitido desarrollos al interior de la misma matemática (p.47).

En sentido amplio, con referencia a la estadística vista como cultura, señala Batanero (2011):

> a nivel internacional la Unesco implementa políticas de desarrollo económico y cultural para todas las naciones, que incluyen no solo la alfabetización básica, sino la numérica. Por ello, los estadísticos sienten la necesidad de difusión de la estadística, no solo como una técnica para tratar los datos cuantitativos, sino como una cultura, en términos de capacidad de comprender la abstracción lógica que hace posible el estudio cuantitativo de los fenómenos colectivos (p.12).

En consecuencia, visto que a nivel local y mundial la estadística cobra importancia a través de una alfabetización y cultura, corresponde a los docentes comenzar la implementación. Si bien es cierto que no se está a la espera del *método super-mejor de enseñanza*, sí se esta a la espera de optimizar recursos y aunar esfuerzos, así poder explorar y enseñar con sencillez todo lo concerniente a la estadística.

Anexo 2

La distribución ji-cuadrado

Si se tienen $Z_1, Z_2, ...Z_n$ n variables aleatorias con distribución normal estándar (Peña, 1988), esto es, estas variables son de valor esperado cero y varianza uno y además independientes e igualmente distribuidas (iid); se puede definir una variable ji-cuadrado (X_n^2) con n grados de libertad, como sigue:

$$X_n^2 = Z_1^2 + Z_2^2 + ... + Z_n^2$$

De modo que, X_n^2 es el resultado de sumar los cuadrados de cada una de las variables inicialmente dadas. Así, el valor esperado (E) y varianza (Var) de nuestra nueva variable serán:

$$E(X_n^2) = n \ , \ Var(X_n^2) = 2n$$

Si se observa la figura 4 donde se encuentran varias distribuciones de variables ji-cuadrado, es sobresaliente que no son simétricas, pues siempre tienen un sesgo bien marcado; la propiedad fundamental de la variable ji-cuadrado es que si se suman dos de ellas, con n_1 y n_2 grados de libertad cada una de ellas, se obtiene una variable también ji-cuadrado, pero con $n_1 + n_2$ grados de libertad.

21 Ente gubernamental colombiano, encargado de la educación en toda la nación.

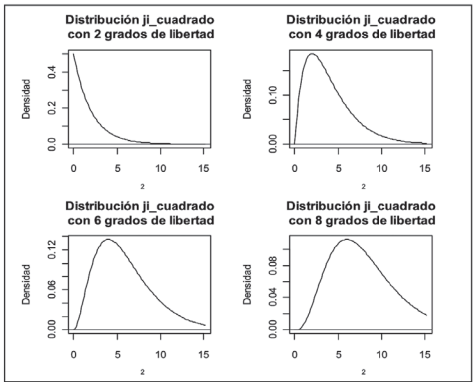


Figura 4. Distribuciones ji-cuadrado para distintos grados de libertad. Hechas en R[22]

Un software útil en el procesamiento de datos

De forma complementaria a toda la temática en este documento tratada, podemos decir algo de las tecnologías de la información y la comunicación (TIC) y dejar al lector la inquietud a manera de prospectiva. Existen muchas tecnologías en informática que están disponibles para la estadística, *software* a modo de paquetes que facilitan la implementación de metodologías para el análisis de datos. SAS, SPSS, MINITAB, STATA, R: entre los más usados, R (por recomendación del autor de este escrito) es en realidad el más idóneo, pues, su manejo se considera relativamente fácil. Algunos aspectos ventajosos de R:

- *Software* siempre disponible, es de uso libre.
- Un análisis soportado en R requiere varios pasos, así el interesado debe saber de ello, es decir, se requiere un usuario bien estructurado.
- Se pueden construir objetos (como funciones), por tanto propicia usuarios siempre analíticos y en procura de resolver problemas.

Veamos una síntesis de lo que es R y su relación con la estadística (W. N. Venables, D. M. Smith, 1999).

Entorno R

R es un colectivo integrado de programas para manipulación de datos, cálculo y gráficos. Entre otras características dispone de:

- Almacenamiento y manipulación efectiva de datos.
- Operadores para cálculo sobre variables indexadas (Arrays), en particular matrices.
- Una amplia, coherente e integrada colección de herramientas para análisis de datos.
- Posibilidades gráficas para análisis de datos, que funcionan directamente sobre pantalla o impresora.
- Un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas. (Debe destacarse que muchas de las funciones suministradas con el sistema están escritas en el lenguaje R).

El término "entorno" lo caracteriza como un sistema completamente diseñado y coherente, antes que como una agregación incremental de herramientas muy específicas e inflexibles, como ocurre frecuentemente con otros programas de análisis de datos. R es en gran parte un vehículo para el desarrollo de nuevos métodos de análisis interactivo de datos. Como tal, es muy dinámico y las diferentes versiones no siempre son totalmente compatibles con las anteriores. Algunos usuarios prefieren los cambios debido a los nuevos métodos y tecnología que los acompañan; a otros, sin embargo, les molesta ya que algún código anterior deja de funcionar. Aunque R puede entenderse como un lenguaje de programación, los programas escritos en R deben considerarse esencialmente efímeros.

Estadística y R

Muchas personas utilizan R como una técnica estadística. Nosotros preferimos describirlo como un entorno en el que se han implementado muchas técnicas estadísticas, tanto clásicas como modernas. Algunas están incluidas en el entorno base de R y otras se acompañan en forma de bibliotecas (*packages*). Junto con R se incluyen ocho bibliotecas (denominadas estándar), pero otras muchas están disponibles a través de internet en CRAN (http://www.r-project.org).

Como hemos indicado, muchas técnicas estadísticas, desde las clásicas hasta la última metodología, están disponibles en R, pero los usuarios necesitarán estar dispuestos a trabajar un poco para poder encontrarlas.

Existe una diferencia fundamental en la filosofía que subyace en R y la de otros sistemas estadísticos. En R, un análisis estadístico se realiza en una serie de pasos, con unos resultados intermedios que se van almacenando en objetos, para ser observados o analizados posteriormente, produciendo unas salidas mínimas. Sin embargo, en paquetes como SAS o SPSS, se obtendría de modo inmediato una salida copiosa para cualquier análisis, por ejemplo, una regresión, una prueba de hipótesis o un análisis discriminante.

22 Software de uso frecuente en estadística.

Referencias bibliográficas

Venables, W.N. y Smith, D.M. (23 de octubre de 1999). *CRAN.R-project*. Recuperado el 17 de Noviembre de 2012 de: http://www.r-project.org

Díaz, L.G. (2009). *Análisis estadístico de datos categóricos.* Bogotá, Colombia: Universidad Nacional de Colombia.

Canavos, G.C. (2001). *Probabilidad y estadística aplicaciones y métodos.* Fernández Ciudad, S.L.: McGraw-Hill.

Mayorga, H. (2004). *Inferencia estadística.* Bogotá, Colombia: Universidad Nacional de Colombia.

Ben_Zvi, D. (2005). *The Challenge of Developing Statistical Literacy, Reasoning and thinking.* Haifa, Israel: Kluwer Academic Publishers.

Batanero, C. (2011). *Estadística con proyectos.* Granada, España: ReproDigital.

Peña, D. (1988). *Estadística modelos y métodos.* Madrid, España: Alianza.

Higgins, J. (2005). *An Introduction to Modern Nonparametric Statistics.* Nueva York: E. B. C.

Batanero, C. (2002). *Los retos de la cultura estadística*. Buenos Aires: ReproDigital.

Ministerio de Educación Nacional (MEN) (12 de Marzo de 2006). *Estándares de competencias en matemáticas - Eduteka.* Recuperado el 17 de Noviembre de 2012 de: www.eduteka.org/pdfdir/MENEstandaresMatematicas2003.php

Ministerio de Educación Nacional Colombia (MEN) (4 de febrero de 1998). *Matemáticas - Ministerio de Educación Nacional.* Recuperado el 17 de Noviembre de 2012 de: www.mineducacion.gov.co/cvn/.../articles-89869_archivo_pdf9.pdf

Sobre el Autor

Wilson Hely Rodríguez Gámez, profesional en Estadística de la Universidad Nacional de Colombia, graduado en 2012, actualmente cursa la especialización en Educación Matemática en la Universidad Pedagógica Nacional. Pertenece al cuerpo docente que en educación Media se desempeña en colegios Feyseristas, en el área de matemáticas y que aplica constantemente métodos de enseñanza-aprendizaje. Ha trabajado en el desarrollo de implementaciones de métodos estadísticos con el software R, con profesores de la misma universidad de la que es egresado. Asistió al Seminario de Inducción para la Creación de Instituciones de Educación Formal, auspiciado por el Cadel de la localidad 18 de Bogotá; recientemente, participó en el III Seminario Internacional de Investigación sobre Calidad de la Educación (ICFES 2012). Realizó, en el marco del último seminario relacionado, el taller de Introducción a Modelos Lineales Jerárquicos.