Análisis de variables de entrada

Introducción

Un sistema es estocástico cuando, al menos, una de sus variables es estocástica. Esto es porque una vez que alguna variable de entrada presenta un comportamiento aleatorio, los resultados o variables de salida se tornan estimaciones y no resultados exactos.

Para modelar el comportamiento estocástico de estas variables, se utilizan funciones de distribución probabilísticas que pueden ser a) teóricas (distribución normal, exponencial, poisson, binomial, etc.) o b) empíricas (tabla de datos, de frecuencias, etc.). A continuación, veremos los cuatro pasos a seguir para llegar a estas funciones:

1. Recopilación de datos del sistema real

Esta etapa requiere de bastante tiempo, esfuerzo y una planificación detallada para aprovechar los recursos. Desafortunadamente, no siempre es factible su realización ya que puede que el sistema no exista aún, o bien porque el relevamiento necesario es muy costoso o peligroso, etc. En esos casos, suele recurrirse al conocimiento de expertos, o a registros históricos del sistema en estudio o de sistemas análogos.

2. Identificación de la distribución de probabilidad que mejor representa a la variable de entrada

Cuando los datos están disponibles, esta etapa se inicia con la construcción de un histograma. Se selecciona una familia de distribuciones en base al histograma y la naturaleza del fenómeno que representa cada variable.

3. Determinación de parámetros

Una vez que se seleccionó la familia de distribuciones, se deben determinar los valores de los correspondientes parámetros que optimizan el ajuste de la distribución a los datos con que se cuenta en la muestra recolectada.

4. Evaluación de la distribución y de los parámetros.

En esta etapa se evalúa la calidad de la representación de los datos a través de la distribución y sus parámetros característicos. Esta evaluación se puede hacer gráficamente o utilizando pruebas estadísticas como las pruebas chi-cuadrado y Kolmogorov-Smirnov.

Si no se puede encontrar una distribución teórica adecuada, será necesario utilizar una distribución empírica.

1. Recopilación de datos

La calidad de la información obtenida del relevamiento de datos depende de la calidad del trabajo realizado en el campo. Una buena organización de recursos y un cuidadoso trabajo de campo son vitales. Por esto, lo primero es la elaboración de un plan. Este plan nos permitirá aprovechar de la mejor manera posible los recursos con que disponemos para la obtención de los datos que interesan a la simulación del sistema. Este plan tendrá en cuenta en qué momento y en qué lugar obtendremos los datos relevantes, y qué cantidad mínima de datos vamos a necesitar para darle consistencia a nuestro trabajo. Es el objetivo de la simulación el que guiará las premisas del plan, y su elaboración debe seguir una serie de recomendaciones:

- Usar un formato estandarizado para registrar la información. Esto vale tanto para estudios en los que los datos serán obtenidos a partir de registros existentes (por ejemplo: historias clínicas, series estadísticas, etc.) así como para estudios que se hacen directamente en el campo.
- Algunas variables, al ser relevadas, tienen varias respuestas posibles no mutuamente excluyentes. En este caso, es necesario forzar la opción “si/no” para cada posible respuesta. Por ejemplo, durante la última semana consumió: pescado si/no; legumbres: si/no; carnes rojas: si/no; carnes de ave: si/no, etc.
- Las variables numéricas deberían ser registradas con la misma exactitud con que fueron obtenidas, sin redondear.
- No categorizar variables numéricas para registrarlas. La categorización suele realizarse una vez obtenida la totalidad de datos del relevamiento.
- Cuando el mismo sujeto es observado más de una vez, por ejemplo durante el control de embarazo o a lo largo de un ensayo, se obtienen medidas repetidas sobre el mismo individuo. No debe considerarse cada visita de un sujeto como un registro independiente.

Por otro lado, también es incorrecto tratar registros múltiples de un individuo como si fueran registros de distintos individuos. Este tipo de datos requiere de métodos estadísticos específicos que se conocen como técnicas para medidas repetidas.

- Deben hacerse arreglos de logística, como decidir sobre el tamaño y composición de los equipos de campo y su traslado al lugar de la encuesta. La cantidad de entrevistadores requeridos depende del tamaño de la muestra, de la cantidad de días que consuman las entrevistas y de la cantidad de encuestas que un entrevistador puede completar en un día.
- Será esencial acompañar el trabajo de campo y proporcionar retroalimentación a los entrevistadores antes de que se haya entrevistado una gran cantidad de casos. Para esto, es necesario que el trabajo de campo sea realizado a un ritmo que permita la identificación de errores sistemáticos, si hubiere, y que se proporcione retroalimentación a los equipos de trabajo antes que sea demasiado tarde.
- Se deberá probar en forma preliminar el instrumento de relevamiento de datos (cuestionario) para identificar potenciales problemas. Se recomienda realizar una prueba preliminar a informantes similares a los que van a ser entrevistados durante la encuesta.
- La recopilación de datos de alta calidad solo es posible si se asigna suficiente tiempo a la capacitación exhaustiva de los trabajadores de campo.
- Otro aspecto importante es la realización de un estudio piloto o preliminar. Este estudio se usa para demostrar que todos los procedimientos funcionan fluidamente y que se comprenden y siguen todos los protocolos. El estudio piloto debe llevarse a cabo, por lo menos, unos días antes del trabajo de campo real. Esto dará tiempo para corregir cualquier problema detectado durante el estudio piloto.
- Antes de obtener todos los datos, se deberá contar con los servicios para hacer la carga de datos y su tratamiento estadístico. Es muy útil hacer pruebas pilotos de ingreso y análisis de datos usando los cuestionarios del estudio piloto. Dependiendo del tamaño del relevamiento, es probable que sea necesario reclutar más o menos personal para el ingreso de datos.
- Como en todo proceso, pueden producirse errores en cualquier etapa. Por ejemplo:

- cuando se toman las mediciones;

- cuando se registran los datos originales (por ejemplo, en la historia clínica);

- cuando se transcribe de la fuente original a una planilla;

- cuando se transcriben los datos para armar la base.

Usualmente no podemos saber si los datos son correctos, pero deberíamos asegurar que son plausibles, y detectar los errores más groseros. La consistencia de los datos se mejora identificando y, de ser posible, rectificando errores.

2. Identificación de la función de distribución de probabilidades

Con los datos compilados, se procede a la construcción de tablas de frecuencias e histogramas para buscar similitudes entre estos gráficos y alguna función de distribución de probabilidades teóricas. De esta forma, se podrá representar y reproducir una variable aleatoria de interés mediante alguna de estas funciones teóricas. El histograma es el más conocido de los gráficos para resumir un conjunto de datos numéricos. Construir manualmente un histograma es laborioso, pero la mayoría de los programas de computación estadísticos producen histogramas.

Histogramas

Se construye un histograma para mostrar gráficamente la forma de la distribución de los datos, por eso debemos atender los aspectos visuales de la representación.

Para construir un histograma es necesario previamente construir una tabla de frecuencias. A partir de una muestra de una variable numérica es posible construir una distribución de frecuencias, clasificando los datos en clases o categorías definidas por el investigador. Las clases o los intervalos de clase de una tabla de frecuencias deben ser mutuamente excluyentes y exhaustivos, es decir, cada dato debe caer en una y solo una clase, y todos los datos deben tener una clase a la cual pertenecen.

Para la construcción de una tabla de frecuencias se siguen los siguientes pasos:

1. Se determina el valor máximo x_{max} y el valor mínimo x_{min} de la muestra n de datos compilados.

2. Se divide el rango $r = (x_{max} - x_{min})$ de los datos en clases o intervalos de cantidad c, los que no necesariamente deben tener la misma longitud. La cantidad de intervalos o clases c puede ser definida arbitrariamente, aunque por lo general se suele adoptar $n^{½}$, y también se recomienda que c sea un número entre 6 y 15.

3. Se cuenta el número de observaciones que cae en cada clase y se determina la frecuencia en cada clase (es decir, la cantidad de veces que los datos caen entre los valores mínimo y máximo de cada clase).

4. Se calculan las frecuencias relativas, frecuencias acumuladas y frecuencias acumuladas relativas para cada intervalo:

- frecuencia f_i = número de casos que cae en el intervalo i
- frecuencia relativa porcentual $fr_i = (f_i / n) * 100$ = porcentaje de casos en el intervalo i
- frecuencia acumulada $fa_i = f_1 + f_2 + ... + f_i$ = suma de las frecuencias desde la primer categoría hasta la categoría i
- frecuencia acumulada relativa porcentual $far_i = (fa_i / n) * 100$ = suma de las frecuencias relativas desde la primer categoría hasta la categoría i

La Tabla 1 presenta las tasas de población rural cada 100 habitantes y la Tabla 2 muestra la tabla de frecuencias para los datos de la Tabla 1 (año 2010, Argentina, fuente: Instituto Geográfico Nacional). Se definieron arbitrariamente intervalos de longitud igual a 2.

Tabla 1: Tasas de Población Rural cada 100 habitantes. Año 2010. Fuente: Instituto Geográfico Nacional

Provincia	Tasa	Provincia	Tasa
Ciudad Autónoma Buenos Aires	0,00	Río Negro	12,95
T.del Fuego, Antártida e Islas	1,19	La Rioja	13,52
Buenos Aires	2,78	Entre Ríos	14,28
Santa Cruz	3,91	Chaco	15,41
Neuquén	8,39	La Pampa	16,82
Chubut	8,81	Corrientes	17,16
Santa Fe	9,15	Mendoza	19,13
Córdoba	10,34	Formosa	19,14
San Luís	11,33	Tucumán	19,19
Jujuy	12,59	Catamarca	22,87
San Juan	12,87	Misiones	26,24
Salta	12,89	Santiago del Estero	31,30

Tabla 2: Distribución de Frecuencias. Tasas de Población Rural por Provincia. Año 2010.

Intervalo	Frecuencia	Frecuencia Acumulada	Frecuencia Relativa	Frecuencia Relativa Acumulada
xmin xmax	f_i	fa_i	fr_i	fra_i
[0 , 2)	2	2	8,33	8,33
[2 , 4)	2	4	8,33	16,67
[4 , 6)	0	4	0,00	16,67
[6 , 8)	0	4	0,00	16,67
[8 , 10)	3	7	12,50	29,17
[10 , 12)	2	9	8,33	37,50
[12 , 14)	5	14	20,83	58,33
[14 , 16)	2	16	8,33	66,67
[16 , 18)	2	18	8,33	75,00
[18 , 20)	3	21	12,50	87,50
[20 , 22)	0	21	0,00	87,50
[22 , 24)	1	22	4,17	91,67
[24 , 26)	0	22	0,00	91,67
[26 , 28)	1	23	4,17	95,83
[28 , 30)	0	23	0,00	95,83
[30 , 32)	1	24	4,17	100,00

El intervalo [0, 2) indica el conjunto de números reales entre 0 y 2, incluye el 0 y excluye el 2.

Una vez que tenemos las tablas de frecuencia, pasamos a la construcción del gráfico o histograma. Se aclara antes que los intervalos pueden ser todos de la misma longitud (tal como lo muestra el ejemplo de la Tabla 2) o bien de distintas longitudes. Además, la forma del histograma depende del número de intervalos de clase que seleccionemos.

Existen distintas fórmulas que permiten calcular el número de clases apropiado para un conjunto de datos, en base al rango de datos y al número de datos. La decisión es una solución de compromiso. Empíricamente se sabe que la fórmula $\sqrt{n}$ (siendo n la cantidad total de datos de la muestra) puede ser una buena aproximación, y también se sabe empíricamente que entre 6 y 15 clases resulta ser una buena elección. Demasiados intervalos harán que caigan muy pocas observaciones en cada clase, por lo que las alturas de las barras variarán irregularmente.

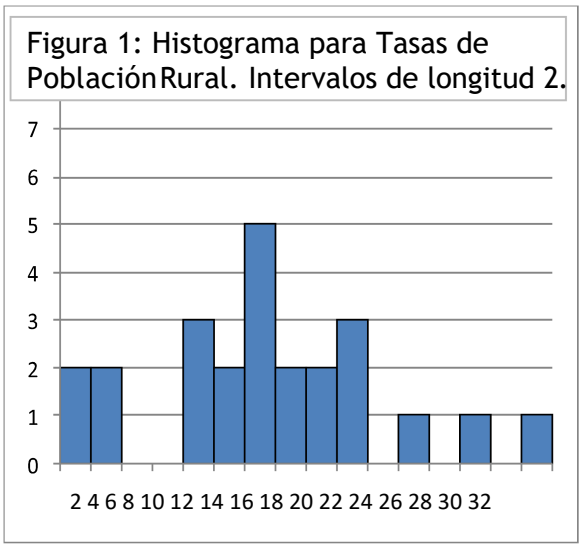
Muy pocas clases producen una gráfica más regular, pero demasiado agrupamiento puede hacer que se pierdan las características principales en la forma del gráfico.

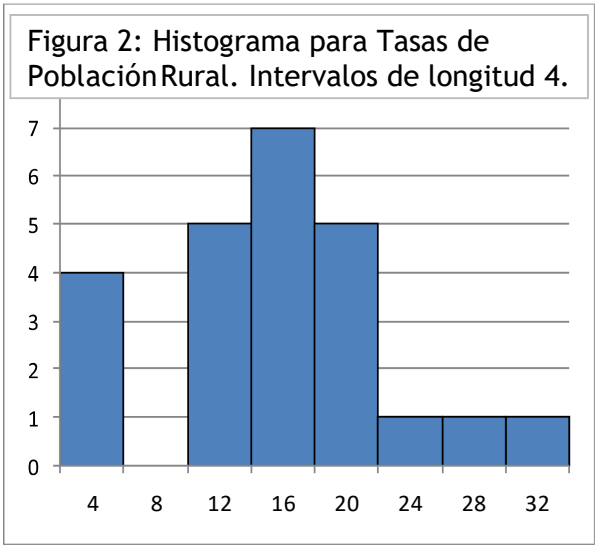
Intervalos de clase de la misma longitud

Se trazan dos ejes de coordenadas ortogonales. En el eje horizontal X se representan los valores de x en el eje vertical Y se representa alguna medida de frecuencia (frecuencia absoluta, frecuencia relativa o frecuencia relativa porcentual).

Indicamos en el eje horizontal los límites de los intervalos de clase. Asociamos a cada clase una columna cuya base cubre el intervalo de clase y cuya altura indica cuantos datos "caen" en un intervalo a través de la frecuencia o la frecuencia relativa de la clase. El gráfico se construye sin dejar espacio horizontal entre categorías, a menos que una clase esté vacía (es decir tenga altura cero).

Las Figuras 1 y 2 presentan dos histogramas para los datos de tasas de población rural de la Tabla 1. El primero tiene intervalos de longitud dos, mientras que el segundo tiene intervalos de longitud cuatro.





En los gráficos puede apreciarse que:

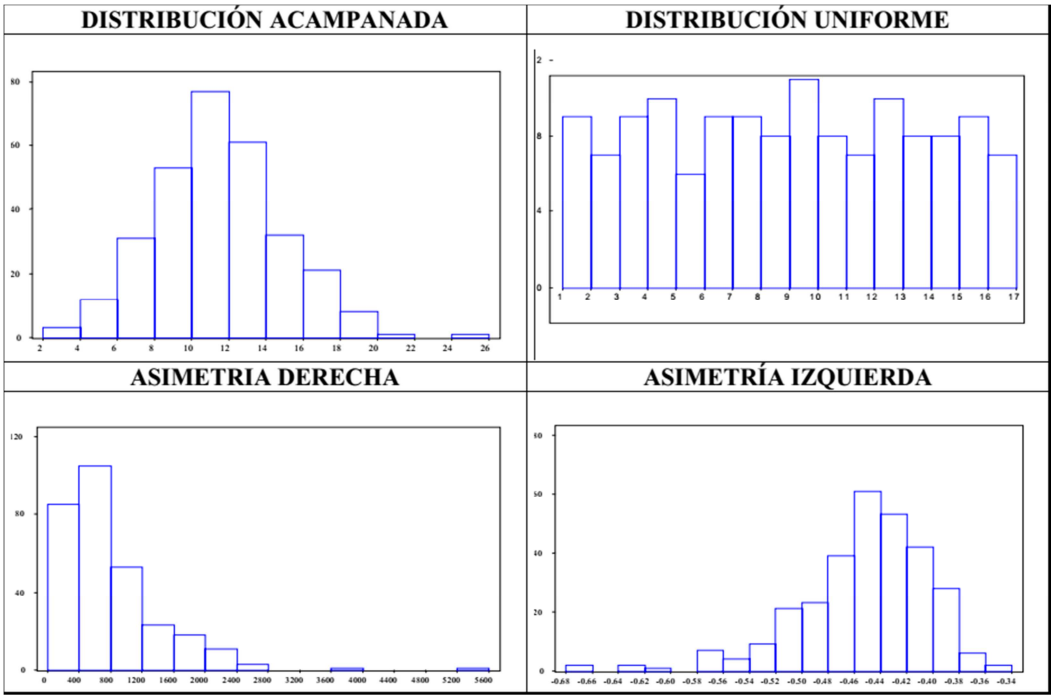
- La distribución es asimétrica, con mayor concentración de datos en los intervalos bajos/medios o en el segundo cuarto.

- Se observan algunos intervalos con frecuencias más altas que el resto. Tal vez podríamos pensar en dos agrupamientos.

- En ambos histogramas observamos un único pico (o moda).

Es importante remarcar qué características del gráfico que no se mantienen al modificar la definición de los intervalos de clase pueden ser consideradas como artificiales. En esos casos, conviene investigar un poco más para detectar la forma más apropiada del histograma que represente fielmente a la variable en cuestión.

Los histogramas siguientes (Figura 3) representan distintas formas posibles para la distribución de los datos. Los dos primeros (distribuciones acampanada y uniforme) muestran distribuciones aproximadamente simétricas, mientras que los dos últimos muestran distribuciones claramente asimétricas. La distribución acampanada se asemeja a la Distribución Normal mientras que la distribución uniforme se identifica con su Distribución homónima, ambas funciones de distribución teóricas. Para los histogramas asimétricos, la asimetría inspira distribuciones que pueden ser Exponencial, Poisson ó Weibull, entre otras.



(Figura 3)

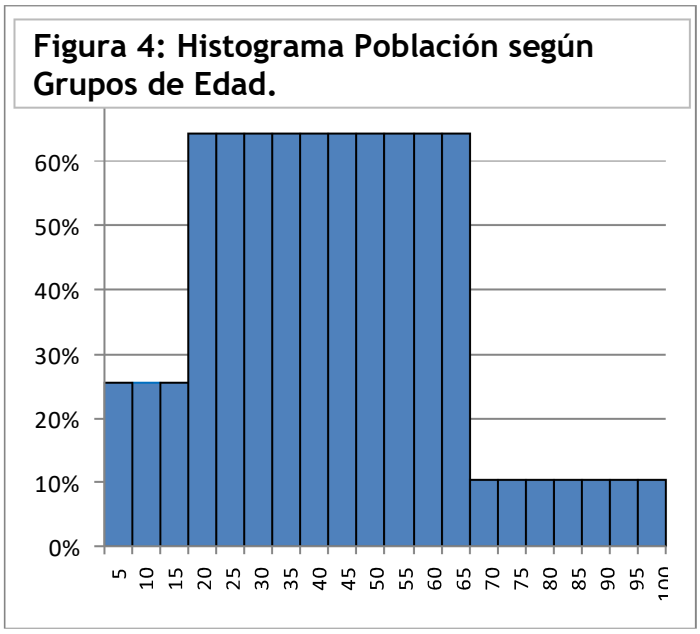
El histograma debería representar la frecuencia asociada a cada clase en el área de la barra y no en su altura. Cuando las clases son todas de la misma longitud, representar la frecuencia en la altura es equivalente a representarla en el área, ya que en todas las barras el área y la altura son proporcionales. Esta es la principal ventaja de mantener todos los intervalos de la misma longitud, pero no siempre puede hacerse.

Intervalos de clase todos de diferente longitud

En ocasiones, es necesario construir histogramas con intervalos de clase de distinto tamaño, por ejemplo, cuando se toma información de datos sociales o económicos. En estos casos, la altura de la barra debe ser tal que el área de la barra sea proporcional a la frecuencia. Consideraremos este tipo de histogramas. La Tabla 3 presenta los datos de población argentina según grupos de edad relevados en el Censo de 2010 y notificados por el INDEC. Nótese que los intervalos de edad tienen diferente longitud. Cuando (erróneamente) se construye un histograma considerando como altura de la barra la frecuencia relativa se obtiene la gráfica siguiente (Figura 4). Se aclara que la última categoría de edad se truncó arbitrariamente en 100 años para poder representarla.

Tabla 3: Distribución de Frecuencias. Población por Grupos de Edad. Año 2010. Fuente: INDEC

Intervalo	Frecuencia	Frecuencia Relativa
	f_i	fr_i
0-14	10.222.317	25,48%
15-64	25.790.131	64,29%
65 y más	4.104.648	10,23%



A partir de este gráfico, concluiríamos que la proporción de casos es notablemente mayor en el grupo de 15 a 64 años, que en los grupos de menores de 15 años o mayores de 64. El problema es que en la imagen parece haber más notificaciones de gente de 15 a 64 que de cualquier otro grupo de edad. Entonces, ¿cómo construimos el histograma teniendo en cuenta que los intervalos de clase son de distinta longitud? La barra debe tener una altura tal que el área (base x altura) sea igual a la frecuencia (o a la frecuencia relativa). Es decir:

altura de la barra = frecuencia en el intervalo / longitud del intervalo

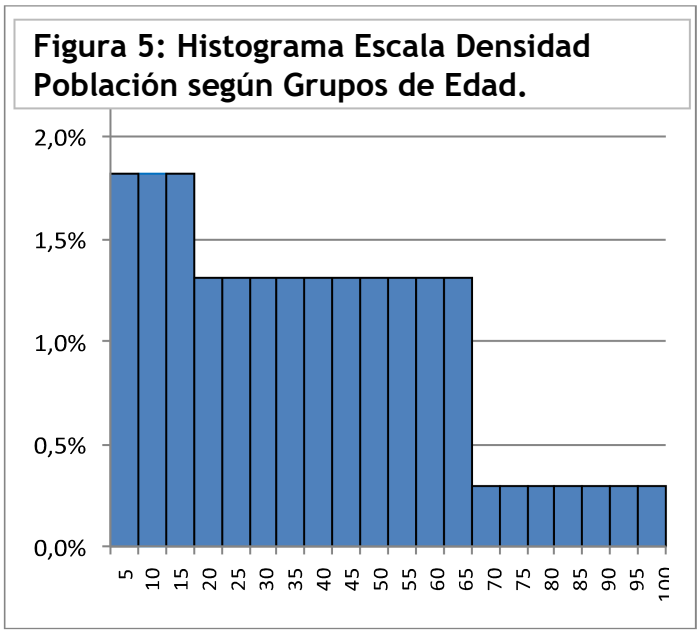
De este modo el área de la barra coincide con la frecuencia en el intervalo:

A = base * altura = (long. intervalo) * (frec. en intervalo) / (long intervalo) = frec

La altura de la barra definida de este modo se denomina “escala densidad” porque indica el número de datos por unidad de la variable. La última columna de la Tabla 4 muestra la escala densidad para los datos de la Tabla 3, y la Figura 5 muestra el histograma que se obtiene usando la escala densidad.

Tabla 4: Distribución de Frecuencias. Población por Grupos de Edad. Año 2010. Fuente: INDEC

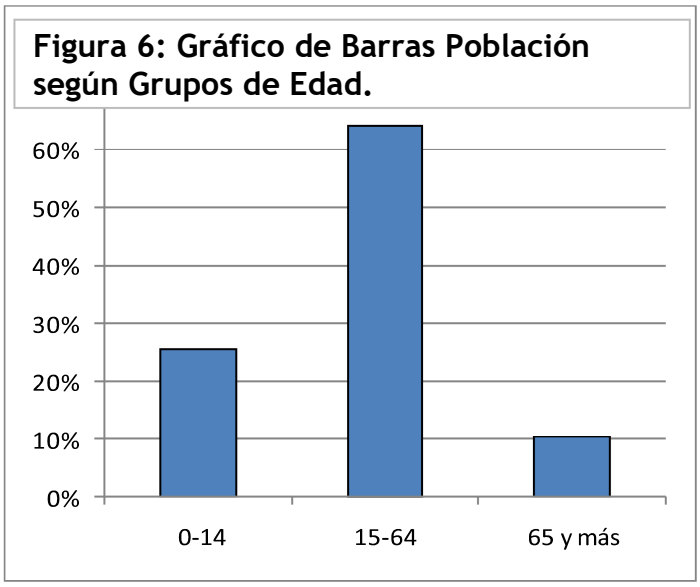
Intervalo	Frecuencia	Frecuencia Relativa	Escala Densidad
	f_i	fr_i	fr_i
0-14	10.222.317	25,48%	1,82%
15-64	25.790.131	64,29%	1,31%
65 y más	4.104.648	10,23%	0,29%



En este histograma, el porcentaje de casos para cada grupo está representado en el área de la barra. Se muestra que una gran proporción de casos ocurre en menores de 15 años, y que la proporción desciende con el aumento de edad. Notar la diferencia en la forma del gráfico de las Figuras 4 y 5.

Una práctica común al manejar datos como los del ejemplo es tratar los datos como categóricos y representarlos en un gráfico de barras como el de la Figura 6.

Cabe resaltar que, en este ejemplo de los grupos de edad de la población, el gráfico de barras da una impresión engañosa de la distribución de casos.



Las diferencias entre un gráfico de barras y un histograma son:

- El gráfico de barras no tiene en cuenta el hecho de que los intervalos de clase tienen distinta longitud.
- El gráfico de barras representa el porcentaje en la altura de la barra. Mientras que en un histograma el porcentaje se representa en el área de la barra.
- En el gráfico de barras, las barras se representan separadas para indicar que no hay continuidad entre las categorías. En un histograma barras adyacentes deben estar en contacto indicando que la variable es continua.

¿Cuándo usar cada uno de ellos? ¿Cuál de las dos representaciones es adecuada?

- Depende de lo que se pretenda mostrar con los datos.
- Cuando la variable que define los grupos es categórica, corresponde usar un gráfico de barras.
- Cuando la variable que define las categorías es numérica, en general lo que interesa es estudiar la distribución de casos en los distintos valores de la variable aleatoria, por lo tanto es preferible el histograma ya que la escala del eje horizontal respeta la escala de la variable de interés.
- Para variables numéricas discretas con pocos valores posibles puede utilizarse un gráfico de barras.

¿A qué función corresponde?

El siguiente paso es determinar la familia de funciones de distribución que se probará para representar el conjunto de datos en estudio. Para ello se cuenta con la forma del histograma y también se cuenta con la naturaleza del proceso. O sea que por un lado apreciamos visualmente a qué función se asemeja nuestro Histograma, pero por otro lado –y no menos importante-, debemos investigar para qué tipos de fenómenos se suelen usar los distintos tipos de distribuciones y observar si alguno de estos casos se identifica con la variable aleatoria que estamos tratando. Por ejemplo, si nuestra variable aleatoria fuera "estatura de una población de individuos", es sabido que ésta y otras características fisiológicas del ser humano son muy bien representadas por la función Normal. Pues bien, entonces nuestro primer intento será representar el conjunto de datos con la función teórica Normal.

Es así que cada una de las funciones de distribución teóricas conocidas tienen fenómenos a los cuales acostumbran representar muy fielmente:

- **Binomial**: Modela el número de éxitos en n pruebas independientes con probabilidad de éxito p; por ejemplo, el número de disquetes defectuosos en un lote de n. Es una distribución de probabilidad discreta que mide el número de éxitos en una secuencia de n ensayos independientes de Bernoulli (experimentos dicotómico, esto es, con resultados posibles "éxito" ó "fracaso").
- **Binomial negativa:** Modela el número de pruebas requeridas para lograr k éxitos; por ejemplo, el número de disquetes que deberían ser revisados para encontrar k defectuosos.
- **Poisson**: Modela el número de eventos independientes que ocurren en una cantidad fija de tiempo o espacio; por ejemplo, el número de clientes que llegan a un centro comercial durante una hora, o el número de defectos encontrados en 30 m2 de una lámina de metal.
- **Exponencial:** Modela el tiempo entre eventos independientes, o tiempos de procesos sin memoria donde no se puede inferir el tiempo final del proceso a partir del tiempo transcurrido; por ejemplo, tiempos entre arribos de clientes. Cuando el tiempo entre eventos sigue una distribución exponencial, el número de eventos en un intervalo fijo de tiempo sigue la distribución Poisson.

- **Normal o Gaussiana:** Es una distribución de probabilidad de variable continua permite modelar numerosos fenómenos naturales, sociales y psicológicos, caracteres morfológicos de individuos como la estatura, el peso, etc, caracteres fisiológicos como el efecto de un fármaco, caracteres sociológicos como el consumo de cierto producto por un mismo grupo de individuos, caracteres psicológicos como el coeficiente intelectual, etc. También es una función que modela la distribución de un proceso que puede representarse como la suma de varios procesos; por ejemplo, el tiempo de ensamblaje de un automóvil puede representarse como la suma de los tiempos de ensamblaje de las distintas partes.
- **Lognormal:** Modela la distribución de un proceso que puede representarse como el producto de varios procesos; por ejemplo, la velocidad de retorno de una inversión con interés compuesto es igual al producto de los retornos de todos los periodos considerados.
- **Gamma:** Se utiliza para modelar valores de datos positivos que sean asimétricos a la derecha y mayores que 0. La distribución gamma se utiliza comúnmente en estudios de supervivencia de fiabilidad. Por ejemplo, la distribución gamma puede describir el tiempo que transcurre para que falle un componente eléctrico. La mayoría de los componentes eléctricos de un tipo particular fallará aproximadamente en el mismo momento, pero unos pocos tardarán más en fallar.
- **Beta:** Se utiliza para modelar variables limitadas a un intervalo entre 0 y 1. Suele utilizarse para modelar eventos que se definen por valores mínimos y máximos. La escala de la distribución beta suele modificarse para modelar el tiempo hasta la culminación de una tarea.
- **Erlang:** Modeyla los tiempos de los procesos que pueden representarse como la suma de varios procesos con distribución exponencial; por ejemplo, una red de computadora falla cuando una computadora y dos computadoras de respaldo fallan sucesivamente, y cada una de ellas tiene tiempos entre fallas exponencialmente distribuidos. La distribución Erlang es un caso especial de la distribución gamma y se utiliza frecuentemente en aplicaciones de teorías de colas.
- **Weibull**: Modela los tiempos entre fallas de un componente; por ejemplo, los tiempos entre fallas de un disco rígido. La distribución exponencial es un caso particular de la distribución Weibull.
- **Uniforme continua y discreta**: Modelan procesos completamente inciertos, ya que todos son igualmente probables.

- **Triangular:** Modela procesos de los cuales sólo se conocen los valores mínimos, máximos y más probables; por ejemplo, cuando se conoce la duración mínima, máxima y más probable de la prueba de un producto. Se emplea mucho en Economía y en aquellos problemas en los cuales se conocen muy pocos datos o no se dispone de datos confiables. Nos permite estimar las duraciones de las actividades de un proyecto usando las tres estimaciones : optimista, muy pesimista, y pesimista. Se denomina triangular "equilátero" cuando viene definida por dos parámetros, que representan el valor mínimo y el valor máximo de la variable. Se denomina triangular "general" cuando viene dada por tres parámetros, que representan el valor mínimo y el valor máximo de la variable, y el valor del punto en el que el triángulo toma su altura máxima (en este caso el triángulo no es necesariamente equilátero).
- **Empírica:** Modela los procesos para los cuales no se pudo encontrar una distribución teórica apropiada.

Una vez que se ha seleccionado una familia de distribuciones, el próximo paso es la estimación de sus parámetros correspondientes.

¿A qué nos referimos con los parámetros?

Cuando se desea determinar información acerca de una característica particular de una población (por ejemplo, la media), generalmente se toma una muestra aleatoria de esa población porque no es factible ni conveniente medir toda la población. Utilizando esa muestra, se calculan "características" de la muestra correspondiente, que se usa para resumir información acerca de "características" desconocidas de la población. Estas características de interés de la población se conocen como parámetros y las características correspondientes de la muestra es la estimación de estos parámetros.

Los parámetros generalmente se representan con letras griegas -por ejemplo, la media de la población se representa con la letra griega mu (μ) y la desviación estándar de la población, con la letra griega sigma (σ)-. Los parámetros son constantes fijas, es decir, no varían como las variables. Sin embargo, sus valores por lo general se desconocen, porque es poco factible medir una población entera.

Cada distribución es definida totalmente por varios parámetros específicos, generalmente entre uno y tres. Los valores de los parámetros determinan la ubicación y la forma de la curva en la gráfica de distribución y cada combinación única de valores de parámetros produce una curva de distribución única.

Por ejemplo, una distribución normal es definida por dos parámetros: μ que no es otra cosa que la media muestral Xm y σ^2 que es la desviación estándar. Si se especifican estos parámetros, se conoce con precisión toda la distribución.

Para una distribución Poisson, se necesita el parámetro α, cuya primera aproximación es la media muestral Xm. Para una distribución Exponencial, el parámetro correspondiente es λ cuya primera aproximación es la inversa de la media muestral: 1/Xm.

Los parámetros son medidas descriptivas de toda una población. Sin embargo, sus valores por lo general se desconocen, porque es poco factible medir una población entera. Por eso, se puede tomar una muestra aleatoria de la población para obtener estimaciones de los parámetros. Un objetivo del análisis estadístico es obtener estimaciones de los parámetros de la población, junto con la cantidad de error asociada con estas estimaciones.

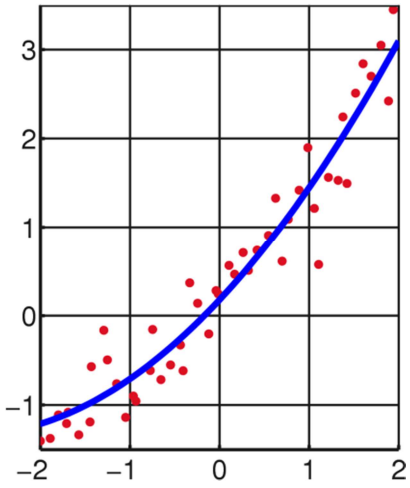
Dado que el parámetro estimado es un resumen de información acerca de un parámetro obtenido a partir de la muestra, el valor de un parámetro estimado depende de la muestra específica que fue extraída de la población. Sus valores cambian aleatoriamente de una muestra aleatoria a la siguiente, por lo que un parámetro estimado es una cantidad aleatoria (variable).

Por ejemplo, cuando tomamos una muestra de una población con distribución normal, la media de la muestra es un parámetro estimado. El valor de la media de la muestra basado en la muestra en cuestión es una estimación de la media de la población. Este valor estimado cambiará aleatoriamente si se toma otra muestra de la misma población normal.

Ejemplo de estimaciones de parámetros: supongamos que usted trabaja para un fabricante de bujías que estudia un problema en la separación entre electrodos. Sería demasiado costoso medir cada bujía que se fabrica. En lugar de ello, toma una muestra aleatoria de 100 bujías y mide la separación en milímetros. La media de la muestra es 9.2. Esta es la estimación de la media de la población (μ). Igualmente crea un intervalo de confianza de 95% para μ que es (8.8, 9.6). Esto significa que puede estar 95% seguro de que el valor verdadero de la separación promedio de todas las bujías se encuentra entre 8.8 y 9.6.

3. Ajuste de parámetros - cuadrados mínimos

La primera estimación o aproximación que hacemos de los parámetros de la distribución elegida no tiene por qué ser la mejor estimación. De hecho, es muy baja la chance de que la primera estimación sea la mejor. Entonces, a esa primera estimación será preciso ajustarla, mejorarla y, en el mejor de los casos, optimizarla.



(Figura 7: Método de mínimos cuadrados)

Cuadrados Mínimos es una técnica de análisis numérico enmarcada dentro de la optimización matemática, en la que, dados un conjunto de pares ordenados y una familia de funciones, se intenta encontrar la función continua, dentro de dicha familia, que mejor se aproxime a los datos (un "mejor ajuste"), de acuerdo con el criterio de mínimo error cuadrático. En nuestro caso, el conjunto de pares ordenados está representado por los intervalos definidos para los valores de la variable aleatoria en cuestión y sus respectivos frecuencia relativa (o probabilidad) asociados.

Cuadrados Mínimos busca minimizar la suma de los cuadrados de las diferencias en las ordenadas entre los puntos generados por la función elegida y los correspondientes valores en los datos. La técnica se usa comúnmente en el ajuste de curvas.

Dados n puntos con coordenadas x e y en un plano, y sea fj(x) un conjunto de funciones linealmente independientes, se quiere encontrar una función f(x) que sea combinación lineal de las funciones base, de modo que

$$f(x) = \sum c_j f_j$$

Por tanto, se trata de hallar los coeficientes cj que hagan que la función aproximante f(x) sea la mejor aproximación para los puntos dados (x, y). El criterio de "mejor aproximación" puede variar, pero en general se basa en aquél que minimice una "acumulación" del error individual (en cada punto) sobre el conjunto total. El error con signo positivo (o negativo) de la función f(x) en cada punto, se define como:

$$e = y - f(x)$$

Como se intenta medir y minimizar el error en todo el conjunto de la aproximación, dicho error (el error "total" sobre el conjunto de puntos considerado) se define como:

$$\sum [y - f_{(x)}]^2$$

donde el término $[y - f_{(x)}]$ se eleva al cuadrado para independizar al error del signo y evitar “compensaciones” entre errores de distinto signo. Para alcanzar este objetivo, se utiliza el hecho que la función f debe poder describirse como una combinación lineal de una base de funciones. Los coeficientes de la combinación lineal serán los parámetros que queremos determinar.

Por ejemplo, supongamos que f es una función cuadrática: $f_{(x)=}ax^2+bx+c$

Por tanto es una combinación lineal de las funciones $f_{1(x)} = x^2$, $f_{2(x)} = x$ y $f_{3(x)} = 1$

Se buscan los valores de los coeficientes a, b y c, de modo que minimicen la suma:

$$\sum [y - f(x)]^2 = \sum [y - ax2+bx+c]^2$$

A las funciones que multiplican a los coeficientes buscados, que en este caso son x2, x y 1, se les conoce con el nombre de funciones base de la aproximación. Para ese caso general

se deduce a continuación la fórmula de la mejor aproximación discreta (i.e. para un conjunto finito de puntos), lineal y según el criterio del error cuadrático medio, que es la llamada aproximación lineal por mínimos cuadrados.

Ejemplo de construcción de histograma y ajuste de parámetros

Considere la Tabla 5 como un conjunto de datos que fue obtenido con un generador de números aleatorios con distribución Exponencial con tiempo medio λ igual a 1.

Tabla 5: Datos Exponenciales $\lambda = 1$

1.9342	1.7925	4.3410	0.2236	0.8880
4.1705	0.0215	0.1490	0.3645	1.7085
1.5877	0.1212	0.2290	0.2933	0.6903
1.0927	0.0366	0.8559	0.2684	1.0873
2.0793	3.0274	0.0356	0.2632	0.0913

Para estos datos, la construcción de la tabla de frecuencias y posterior histograma parte de los siguientes cálculos:

x_{min} = 25

x_{max} = 0.0215

r = 4.3196

c = 5

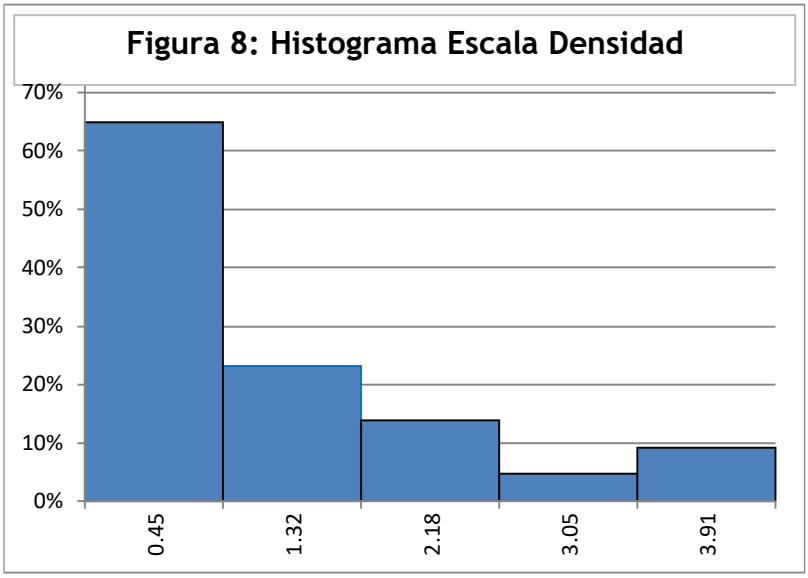
b = 0.8639

x_m = 1.0941

λ_o = 0.9140

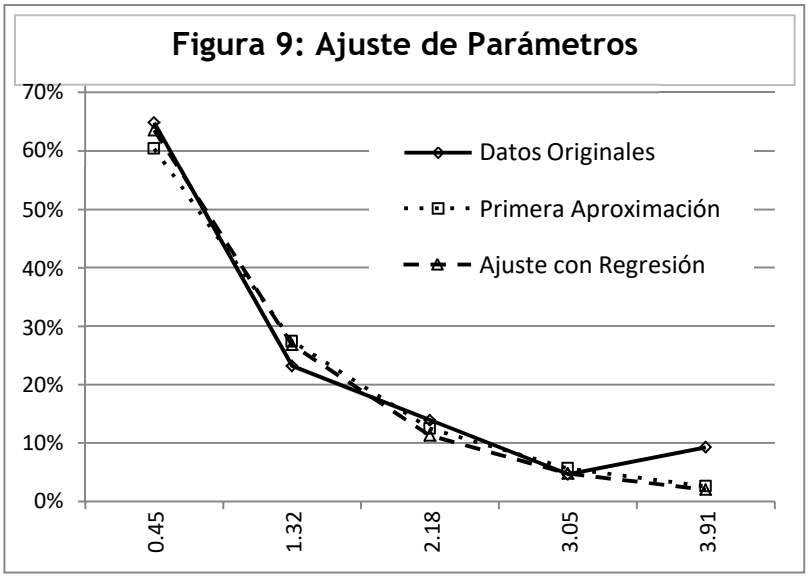
Con estos cálculos, se construye la Tabla 6 que muestra los intervalos, las frecuencias absolutas, relativas y acumuladas correspondientes y también la frecuencia "normalizada" (usando la escala densidad) para luego, en la Figura 8, presentar el histograma correspondiente.

Tabla 6: Intervalos y Distribución de Frecuencias.							
Intervalo				Frecuencia	Frecuencia Relativa	Escala Densidad	Acumulada
i	x_{min}	x_{max}	x_m	f_i	fr_i	fr_i	FDA
1	0.0215	0.8854	0.45	0	0	0	0
2	0.8854	1.7493	1.32	2	0.4	0.46300893	0.4
3	1.7493	2.6132	2.18	2	0.4	0.46300893	0.8
4	2.6132	3.4771	3.05	0	0	0	0.8
5	3.4771	4.3410	3.91	1	0.2	0.23150447	1
				5	1		



En lo que respecta a estimación de parámetros, para el conjunto de datos de la Tabla 5, su valor medio x_m es igual a 1.0941; por lo tanto $\lambda \approx 0.914$, algo alejado de su verdadero valor de 1.

Ahora el problema se reduce a determinar el valor del parámetro λ que hace mínima la suma de los errores al cuadrado. Este problema se resuelve fácilmente con una planilla de cálculo, y el resultado es $\lambda \approx 0.9996$. La Figura 9 muestra a) los valores originales; b) el ajuste realizado con la primera aproximación; y c) el ajuste realizado por regresión o mínimos cuadrados.



4. Evaluación del ajuste

Por lo general, la función de distribución de probabilidades elegida y luego ajustada aún tendrá diferencias con el histograma. Entonces, para evaluar si realmente la distribución propuesta representa al conjunto de datos, o en otras palabras, si las diferencias observadas pueden ser menospreciadas, se pueden utilizar métodos estadísticos como por ejemplo:

- **Kolmogorov-Smirnov**: decide en base a la máxima desviación entre la distribución acumulada teórica y la distribución acumulada empírica.
- **Chi-cuadrado**: Decide en base a la suma de errores al cuadrado que surgen de comparar el histograma con la distribución teórica.

Estas pruebas son una buena guía para evaluar una distribución. Sin embargo, ya que no existe una distribución teórica que ajuste perfectamente a los datos del mundo real, no se deberían tomar los resultados de estas pruebas en forma categórica. Es muy importante comprender el efecto del tamaño de la muestra. Si la muestra es pequeña, las pruebas aceptarán cualquier distribución. Por el contrario, si la muestra es grande, las pruebas rechazarán a todas las distribuciones propuestas. Por lo tanto, estas pruebas son sólo un elemento más a tener en cuenta durante la evaluación.

Estas pruebas, si bien se mencionan, no son objeto del estudio del presente texto.