ISSN 0188-7297

INSTITUTO MEXICANO DEL TRANSPORTE

SCT
SECRETARÍA DE COMUNICACIONES Y TRANSPORTES

MODELOS DE PROBABILIDAD EN TRANSPORTE E INGENIERÍA. USOS COMUNES Y AJUSTE DE DATOS

Eric Moreno Quintero
Verónica Nieves Cruz

Publicación Técnica No. 545
Sanfandila, Qro, 2019

SECRETARÍA DE COMUNICACIONES Y TRANSPORTES

INSTITUTO MEXICANO DEL TRANSPORTE

Modelos de probabilidad en transporte e ingeniería. Usos comunes y ajuste de datos.

Publicación Técnica No. 545
Sanfandila, Qro, 2019

Esta investigación fue realizada en la Coordinación de Integración del Instituto Mexicano del Transporte, por el Dr. Eric Moreno Quintero con la colaboración de la Lic. en Matemáticas Aplicadas Verónica Nieves Cruz, de la Universidad Autónoma de Querétaro, durante su estancia de prácticas profesionales en este Instituto.

Esta investigación es el producto final del proyecto de investigación interna de clave TI-02/18, titulado “Modelos de Probabilidad en Transporte e Ingeniería. Usos comunes y ajuste de datos”.

Se agradecen las colaboraciones y sugerencias de: el Dr. Alfredo López, de la Coordinación de Ingeniería Portuaria y Sistemas Geoespaciales; del Dr. Francisco Carrión, de la Coordinación de Ingeniería Vehicular e Integridad Estructural; así como del M. en C. Juan Fernando Mendoza y del M. en I. José Antonio Gómez, de la Coordinación de Infraestructura, quienes amablemente compartieron algunos materiales de ingeniería de tránsito así como datos de prueba para los ensayos estadísticos y aportaron su experiencia de manejo de datos experimentales en sus respectivos trabajos de laboratorio.

Contenido

Índice de figuras

Índice de tablas

Sinopsis

Este trabajo revisa las distribuciones de probabilidad más comunes en problemas de transporte e ingeniería, y añade una revisión bibliográfica que identifica aplicaciones exitosas reportadas de estas distribuciones en la práctica ingenieril y del transporte.

El capítulo 1 discute la relevancia de tener modelos probabilistas adecuados para datos de tipo aleatorio usados en la práctica ingenieril y del transporte.

El capítulo 2 trata las distribuciones de probabilidad más comunes encontradas en problemas de ingeniería y transporte, indicando detalles prácticos para el cálculo de probabilidades.

El capítulo 3 muestra una amplia revisión de literatura de aplicaciones exitosas reportadas para distribuciones de probabilidad en campos diversos como: ingeniería de tránsito, hidrología, confiabilidad, accidentalidad y otros.

El capítulo 4 revisa las tres técnicas de bondad de ajuste más usadas en la práctica para evaluar modelos probabilistas de datos: Ji-cuadrada, Kolmogorov-Smirnov y Anderson-Darling. Aquí se desarrollan ejemplos numéricos detallados con Excel, para que el lector pueda replicarlos.

El capítulo 5 da algunas conclusiones y comenta el uso de software estadístico especializado y la correcta interpretación de resultados.

El apéndice al final da las referencias de todos los artículos citados en las diversas aplicaciones de distribuciones de probabilidad que se detallan en el Capítulo 3.

Abstract

This work reviews the more common probability distributions found both in transport and engineering problems, adding a literature review to identify successful applications reported for these distributions at the engineering and transport practice.

Chapter 1 discusses the relevance of having adequate probabilistic models for random data used in the engineering and transport practices.

Chaper 2 reviews the more common probability distributions found both in engineering and transport problems, giving some practical details for probability calculations.

Chapter 3 shows a wide literature review of successful applications reported for probabilitydistributions in several fields as: traffic engineering, hydrology, reliability, accident studies and several more.

Chapter 4 reviews the three more usual tecniques for goodness of fit to assess probabilistic models chosen for data: Chi-squared, Kolmogorov-Smirnov and Anderson-Darling. Here detailed numerical examples in Excel are shown, in order that readers can replicate them.

Chapter 5 give some conclusions and comments on the use of statistical software and the proper interpretation of results.

Appendix at the end gives references for all the papers reviewed on the several applications of probability distributions commented at Chapter 3.

Resumen ejecutivo

Este trabajo surge de la identificación de una carencia de información en la investigación y en la solución de problemas de transporte y de ingeniería, donde hay datos de tipo aleatorio, esto es: la elección de una distribución de probabilidad que represente adecuadamente a esos datos.

En algunos casos prácticos, puede haber cierta información del modelo probabilístico que conviene usar según el campo de origen de los datos aleatorios.

Así, por ejemplo, los trabajos que tratan con aforos vehiculares tradicionalmente se refieren a la distribución Poisson para representar conteos y a la distribución exponencial para representar tiempos entre llegadas de modo similar, los problemas relativos a tiempos de vida de equipos o tiempos de espera a la falla en un sistema, tradicionalmente han usado la distribución Weibull o la Gamma para representar estas variables.

En este par de ejemplos, si bien la práctica ha mostrado la conveniencia de usar esas distribuciones, ante un conjunto concreto de datos se requiere estimar los parámetros numéricos que definen las distribuciones, y agregar un criterio de validez estadística para tener un nivel de confianza en el uso de esos modelos.

Con este antecedente, el Capítulo 1 comenta la relevancia de tener un modelo probabilístico adecuado para datos de tipo aleatorio, observando que esto permite:

a) Tener un modelo de los datos con sustento probabilístico teórico firme.

b) Estimar confiablemente probabilidades de los eventos de interés y sus valores esperados.

c) Lograr resultados más realistas al usar técnicas de simulación

d) Tener una base estadística sólida en la argumentación de artículos a someter a revistas científicas.

Continuando con el Capítulo 2, se hace una revisión de las distribuciones de probabilidad básicas que suelen tratarse en aplicaciones de transporte e ingeniería, mostrando sus caracterizaciones más comunes y las fórmulas para la media y la varianza de cada una. En particular, se aclaran las dos versiones que existen tanto para la distribución Geométrica como para la Pascal, que frecuentemente aparecen separadas en los textos universitarios de estadística para ingeniería, y que pueden causar confusión en su interpretación.

En este capítulo también se comentan aspectos prácticos para el cálculo de probabilidades, desarrollando varios ejemplos numéricos a detalle, con histogramas de los datos de muestra y su interpretación, y mostrando las capacidades de cálculo estadístico que proporciona la versión estándar de Excel, y con las cuales se ahorra tiempo de cálculo.

El Capítulo 3 muestra una amplia revisión de literatura en la que se identifican aplicaciones de las distribuciones de probabilidad que se han reportado con éxito en diversos campos de ingeniería y transporte. Con esta identificación, el trabajo de investigación o de solución de problemas que requieren de un modelo probabilístico adecuado para datos aleatorios, puede tener una guía que aproveche la experiencia de otros investigadores para manejar convenientemente sus propios datos.

Así, en la Tabla 3.1 que muestra aplicaciones de distribuciones discretas, se observa que la distribución Poisson se ha usado para clasificar puntos viales de alta accidentalidad, o para estimar el número diario de fallas en una red de cómputo; que la distribución Binomial Negativa (Pascal) se ha utilizado para modelar número de pedidos en un inventario y también en conteo de accidentes viales; y que la distribución logarítmica también se ha utilizado para analizar frecuencias de accidentes viales.

Análogamente, en la Tabla 3.2 que muestra aplicaciones de distribuciones continuas, puede verse que la distribución Exponencial se ha utilizado para modelar retrasos en el sistema ferroviario británico, o en la modelación de líneas de espera en un aeropuerto; que la distribución Gamma se ha utilizado con éxito para modelar niveles de lluvia o también para modelar tasas de accidentes de aviación como función de las horas de vuelo de los pilotos; que la distribución Lognormal se ha utilizado para modelar tiempos de procesamiento de datos en un sistema de información, para modelar velocidades del viento y también para modelar índices de medición de compactación de suelos. Otros ejemplos de aplicaciones exitosas se muestran en la Tabla 3.2 para las distribuciones Weibull, Logística y Gumbel.

La identificación de aplicaciones concluye con un breve resumen de cuatro distribuciones especializadas en problemas de diseño de materiales en ingeniería:

a) **Distribución Rosin-Rammler**: relacionada con la Weibull, y que fue usada por Rosin y Rammler en 1933 para describir la distribución del tamaño de partículas de material granulado.
b) **Distribución Hiperbólica**: propuesta por Bagnold y Barndorff-Nielsen en 1980, para modelar tamaños de partículas de material sedimentado. Esta distribución es tal que la gráfica del logaritmo de su densidad resulta ser una hipérbola (Bagnold & Barndorff-Nielsen, 1980).
c) **Distribución Log-Laplace**: Esta distribución se propuso como alternativa a la Hiperbólica, de la cual se encontraron desempeños pobres en algunos análisis de sedimentos (Fieller, Gilbertson and Olbricht, 1984).

d) **Distribución Birnbaum-Sanders**: Surgió al modelar fallas en materiales debido a fracturas, en ensayos donde los materiales de prueba se someten a ciclos repetidos de esfuerzos. Se conoce también como distribución del tiempo a la fatiga (fatigue life distribution).

Como muestra de la variedad de aplicaciones de las distribuciones de probabilidad en problemas de ingeniería, en este capítulo se cita el siguiente fragmento de Dianty, Yahaya y Ahmad (2014):

> "*Las propiedades ingenieriles de suelos, colectados de distintos sitios designados para construir torres de comunicaciones, fueron obtenidos y analizados. […] La razón de vacíos, el peso unitario a granel y el peso unitario seco, siguieron una distribución Normal; el contenido de agua, el límite líquido y el índice de plasticidad siguieron una distribución Gamma; la especificación de la gravedad, la porosidad, la saturación, el ángulo interno de fricción y de cohesión se ajustaron a una distribución Weibull, y el peso unitario saturado siguió una distribución log-normal.*"

En el Capítulo 4 se inicia la discusión de las técnicas de bondad de ajuste para los datos. Primeramente, se examinan las técnicas usuales para estimar los parámetros numéricos que particularizan a la distribución de probabilidad elegida para los datos: a) el método de momentos, b) el método de máxima verosimilitud.

Luego de describir estos métodos y de mostrar ejemplos numéricos resueltos, se revisan las tres técnicas de bondad de ajuste más usuales en la práctica: a) Ji-cuadrada, b) Kolmogorov-Smirnov y c) Anderson-Datrling

En cada caso se muestra la secuencia de pasos a seguir, la referencia a las tablas de valores críticos, y se dan ejemplos numéricos desarrollados en Excel con los detalles necesarios para su replicación.

En el Capítulo 5 se muestran algunas conclusiones del trabajo, enfatizando la relevancia de tener representaciones probabilísticas adecuadas de los datos de tipo aleatorio, y el beneficio que representa tener un buen modelo probabilista básico empleando el mínimo de información de una muestra de datos.

Se prosigue con una sección de comentarios y sugerencias para utilizar software estadístico comercial e interpretar correctamente sus resultados, y se muestran ejemplos numéricos desarrollados con tres paquetes estadísticos de uso común: Minitab 14, JMP 9 y STATISTICA 4.3.

El Capítulo 5 finaliza con una sección que comenta dos casos que pueden aparecer en la práctica: 1) cuando se tienen varios posibles ajustes aceptables y 2) cuando no hay ajuste alguno en las distribuciones que se probaron.

En el primer caso se recomienda elegir la distribución con el mejor valor-p (p-value) para representar los datos, y en el segundo caso se recomienda construir una

distribución empírica, como en el ejemplo numérico que se desarrolla para tal propósito.

El anexo al final del trabajo incluye todas las referencias bibliográficas de los artículos referidos en los que se reportan aplicaciones diversas de las distribuciones de probabilidad, a fin de que el lector pueda consultar las fuentes primarias para su trabajo

1 Introducción

En el desarrollo de proyectos de transporte y en general de ingeniería, es frecuente encontrar datos que tienen carácter aleatorio y que deben tratarse adecuadamente por la relevancia que tienen en los objetivos de los proyectos.

Ejemplos de casos como: conteos por clase vehicular en un puente de cuota; tiempos de demora en las llegadas o salidas de aviones a un aeropuerto; número de accidentes en un crucero urbano; porcentajes de camiones sobrecargados en una ruta carretera; tiempos de tránsito de un embarque que va a la frontera, etc. son todos ejemplos de datos que tienen un comportamiento aleatorio. Son datos que si bien es posible tener una idea del rango de valores que podrían tomar, no permiten hacer un pronóstico exacto del siguiente valor que se observará en un muestreo de colecta.

La relevancia de estos datos es que con ellos se hacen estimaciones diversas de interés para un proyecto dado. Por ejemplo, en un puente de cuota el tener una estimación razonable de los conteos vehiculares es importante para pronosticar los recursos humanos y materiales que deben asignarse a fin de mantener un desempeño razonable del servicio en el puente, así como, el nivel de ingresos esperados con una aproximación razonable. De la misma manera, los datos de conteos vehiculares junto con los de los tiempos de atención en el cobro, permiten estimar los valores promedio de tiempos de espera y del tamaño de cola para los usuarios del puente.

En el contexto de la simulación de eventos discretos, los proyectos que utilizan esta técnica requieren de un modelo probabilístico adecuado para las variables que se utilizan en la simulación, de modo que la representación del sistema que haga la simulación sea lo más realista posible.

De esta manera, en un proyecto de simulación de flujos vehiculares en caminos sin control de semáforo, resulta más adecuado modelar los tiempos de llegada de los vehículos con una distribución exponencial, que utilizar otras distribuciones continuas. La adecuación de utilizar una modelo exponencial para los tiempos de llegada de los vehículos al sistema estudiado se verifica cuando el modelo de simulación reproduce correctamente la formación de pelotones de los vehículos circulantes y de colas en los cruceros, lo que en general no ocurre si se utiliza, por ejemplo, una distribución normal o una distribución uniforme para los tiempos entre la llegada de los vehículos.

Los datos de los ejemplos anteriores están asociados al concepto teórico de *variable aleatoria*. Una variable aleatoria asigna valores a cada resultado posible de un

suceso aleatorio, utilizando una función matemática que constituye su *ley de probabilidad*.

Es esta ley matemática la que permite estimar las probabilidades de los eventos de interés en un problema dado, y con la cual se calculan *valores esperados* que son relevantes para la situación estudiada.

Por ejemplo, si la probabilidad de encontrar un camión doblemente articulado con sobrecarga en una carretera se ha estimado en 37.2%, el valor esperado de sobrecargados en una muestra de 500 camiones es $500 \times 0.372 = 186$ aproximadamente; este número ya permite dimensionar el problema de los sobrecargados, en particular si se tiene alguna estimación de las toneladas extra que lleva cada camión con sobrecarga.

La circunstancia más común al abordar un problema de transporte o de ingeniería que tiene elementos de carácter aleatorio es que, aunque haya datos recientes disponibles, se requiere especificar una ley de probabilidad que represente estos datos razonablemente para que las estimaciones de probabilidades de eventos de interés y de valores esperados resulten confiables.

Si el interés fueran los accidentes viales severos/año en un crucero, ¿cuál es la probabilidad de que ocurran 5, 6 o 7 accidentes ahí? ¿Cuál ley de probabilidad caracteriza mejor a ese número de accidentes, una distribución Normal, una Binomial Negativa o una Poisson? Además, si ya se tiene una propuesta de distribución de probabilidad para los datos, por alguna razón teórica o práctica, ¿cuál es el criterio estadístico que debe usarse para evaluar la adecuación de los datos a esa distribución propuesta?

La cuestión anterior puede resolverse revisando la literatura del tema y las experiencias reportadas, para identificar las leyes de probabilidad que han tenido resultados aceptables en diversas aplicaciones.

Una vez especificada una ley de probabilidad para los datos, estos deben pasar por una prueba de *bondad de ajuste*, un criterio estadístico de aceptación general que indica qué tan buena es esta representación de los datos. Si para representar un conjunto de datos se proponen varias distribuciones de probabilidad, la ley de probabilidad que mejor caracteriza a esos datos es la que obtenga mejor desempeño en la prueba de bondad de ajuste.

De esta manera, se logra una base objetiva y con fundamento estadístico para proseguir con el estudio, usando un modelo probabilístico adecuado para los datos y del cual se obtienen estimaciones y valores esperados confiables.

En lo que sigue de este trabajo, el capítulo dos hace una revisión de las principales distribuciones de probabilidad que surgen en trabajos de transporte y de ingeniería en general.

En el Capítulo 3 se muestran aplicaciones diversas de las distribuciones de probabilidad que han sido reportadas en literatura y que permiten identificar la clase de problemas en los que las distribuciones referidas tienen mejores posibilidades de tener buenos ajustes.

El Capítulo 4 revisa las tres principales pruebas de bondad de ajuste usadas en la práctica: la prueba Ji-cuadrada, la Kolmogorov-Smirnov y la Anderson-Darling. Estas pruebas se explican a detalle con ejemplos de reportes de paquetes estadísticos comerciales típicos con datos de muestra. Asimismo, se muestran ejemplos numéricos resueltos de las técnicas de bondad de ajuste discutidas en el capítulo previo, utilizando un formato de hoja de cálculo de Excel, hace de que el usuario pueda reproducir estas técnicas por su cuenta.

Finalmente, en el Capítulo 5 de conclusiones se resume la experiencia de este trabajo y se señalan líneas futuras de desarrollo.

2 Distribuciones de probabilidad

Las distribuciones de probabilidad pueden ser discretas o continuas, dependiendo de la naturaleza de las variables aleatorias (v. a.) cuyo comportamiento aleatorio describen.

Las variables aleatorias discretas, generalmente se miden con números enteros, y en la práctica aparecen en situaciones donde hay conteo de un cierto evento de interés. Por ejemplo: el número de pasajeros en un autobús específico; el número de camiones articulados que cruzan una caseta de peaje entre las 10:00 am y las 11:00 am de un día laboral, el número de operadores sin licencia en regla en una muestra de 60 encuestas de camino, o el número de vehículos que usaron el estacionamiento de una terminal en un día determinado. En todos estos casos, la respuesta a la pregunta es un número entero, que suele variar cada vez que se vuelve a medir bajo la misma circunstancia.

Las variables aleatorias continuas, pueden, en principio, tomar cualquier valor positivo, cero o negativo, y en aplicaciones depende de lo que se esté midiendo. Por ejemplo, el tonelaje movido en camiones C2 que transportan materiales de construcción; el tiempo promedio en que se recorre una ruta carretera; los minutos de retraso de un vuelo que llega a un aeropuerto o el rendimiento de combustible mensual de un camión articulado. En estos casos, la respuesta en cada ejemplo es un número expresado con alguna fracción decimal, que también cambia cada vez que se repite la observación, ya que, los valores de interés a los casos ejemplificados cambian cada vez que se repite la observación o medición.

Es de interés saber qué tan frecuentemente aparecen ciertos valores, qué porcentaje de esos valores están entre ciertos límites especificados y también a largo plazo, cuál es el valor promedio que representa la medida general que nos interesa.

Estas cuestiones se pueden responder si se conoce una distribución de probabilidad para los valores observados o medidos, la cual permite calcular las probabilidades de ocurrencia de los posibles valores, y con ello estima el porcentaje de veces o el porcentaje del tiempo en que aparecen ciertos valores, y permite calcular el *valor esperado* de esas mediciones, que es justamente el promedio general de comportamiento de las observaciones de interés.

Enseguida se describen las distribuciones que con frecuencia se utilizan en la práctica ingenieril.

2.1 Las distribuciones básicas

Las distribuciones más comunes tratadas en transporte e ingeniería son las siguientes:

Distribuciones discretas

Para estas distribuciones, el ensayo de Bernoulli, es la base para otras distribuciones descritas con experimentos prototipo. El ensayo de Bernoulli hace una observación para detectar el evento de interés (un accidente, un retraso en un aeropuerto, un camión con sobrecarga, etc.) del cual se conoce la probabilidad de ocurrencia *p*.

Si el evento ocurre, la variable aleatoria (v.a.) asociada X = 1, y se dice que hubo un "éxito"; si no ocurre, X = 0 y hay un "fracaso" (con probabilidad *1 – p).* Entonces X tiene distribución Bernoulli de parámetro *p*. Algunas distribuciones discretas descritas con ensayos de Bernoulli son la Binomial, la Geométrica y la Pascal.

a) Distribución Binomial. Al repetir *n* veces un ensayo Bernoulli de parámetro *p*, la v. a. X = número de éxitos logrados tiene distribución Binomial de parámetros *n*, *p*. Los valores de X y su función de probabilidad son:

$$P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}\ ;\ \ k = 0, 1, 2, \dots, n;\ \ 0 \le p \le 1$$

b) **Distribución Geométrica**. Repitiendo un ensayo Bernoulli de parámetro *p*, hasta lograr éxito, se tiene la distribución Geométrica de parámetro *p*. Una v.a. X geométrica puede interpretarse de dos formas:

X_a = *número de fracasos previos* a la aparición del éxito, y

X_b = *número de repeticiones* que se necesitaron para lograr el éxito.

Es importante distinguir entre estas dos interpretaciones, en particular si se usa software comercial estadístico para calcular probabilidades geométricas. El rango de valores de X y su función de probabilidad son:

$$P[X_a = k] = \ p(1-p)^k\ ;\ \ k = 0, 1, 2, \dots, \infty;\ \ 0 \le p \le 1$$

$$P[X_b = k] = \ p(1-p)^{k-1}\ ;\ \ k = 1, 2, \dots, \infty;\ \ 0 \le p \le 1$$

c) **Distribución Pascal o Binomial Negativa**. Esta distribución generaliza la geométrica, repitiendo el ensayo Bernoulli hasta lograr un número dado *r* de éxitos. La v.a. asociada X tiene distribución Pascal (o Binomial Negativa) de parámetros *p* y *r*. Como con la geométrica, X hay dos interpretaciones:

X_a = *número de fracasos previos a la aparición del r-ésimo* éxito, y

X_b = *número de repeticiones requeridas para lograr el r-ésimo éxito*.

El rango de valores de X y su función de probabilidad son:

$$P[X_a = k] = \binom{r+k-1}{k} p^r (1-p)^k \; ; \;\; k = 0, 1, 2, \dots, \infty; \;\; 0 \leq p \leq 1$$

$$P[X_b = k] = \binom{k-1}{r-1} p^r (1-p)^{k-r} \; ; \;\; k = r, r+1, r+2, \dots, \infty; \;\; 0 \leq p \leq 1$$

Otra distribución discreta de interés es la Poisson, útil en situaciones de conteos:

d) **Distribución Poisson.** Esta distribución generaliza el proceso de conteo que de una Binomial. A diferencia de esta que repite *n* veces el ensayo y cuenta del total de éxitos logrados, la variable X Poisson cuenta el número de ocurrencias de un evento de interés (similar a "éxitos") en un período de tiempo dado, conociendo la tasa media de ocurrencias por unidad de tiempo, denotada por λ. El rango de valores de X y su función de probabilidad son:

$$P[X = k] = \frac{e^{-\lambda} \lambda^k}{k!} \; ; \;\; k = 0, 1, 2, \dots, \infty; \;\; \lambda > 0$$

e) Distribución Logarítmica. Esta distribución suele aparecer en la literatura como una derivación basada en el desarrollo de la serie de Maclaurin para el logaritmo natural de la variable $(1 - x)$:

$$-\ln(1-x) = x + \frac{x^2}{2} + \frac{x^3}{3} + \dots ; \text{para } -1 \leq x < 1$$

Con un parámetro $0 < \theta < 1$ y la igualdad de la serie de Maclaurin anterior, puede verificarse la siguiente ecuación, que puede considerarse como la suma de probabilidades discretas que define a la distribución logarítmica:

$$\sum_{k=1}^{\infty} \frac{-1}{\ln(1-\theta)} \frac{\theta^k}{k} = 1$$

Por lo que el rango de valores de X y la función de probabilidad son:

$$P[X = k] = \frac{-1}{\ln(1-\theta)} \frac{\theta^k}{k} \; k = 1, 2, \dots \infty; \;\; 0 < \theta < 1$$

Distribuciones continuas

En el caso de las v. a. continuas, las siguientes distribuciones de probabilidad aparecen con frecuencia en problemas de transporte e ingeniería:

a) Distribución Normal. Esta distribución, muy conocida surge de modo natural cuando los valores de interés observados muestran cierta simetría alrededor de un valor central, con histogramas de apariencia acampanada. Sus dos parámetros son la media μ y la desviación estándar σ. Cuando se obtienen muestras de algún dato de interés y se usa su promedio como referencia, los promedios obtenidos repitiendo el muestreo varias veces, suelen tener distribución normal. El rango de valores de X y su *función de densidad de probabilidad (fdp)* $f(x)$ son como sigue:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; \quad -\infty < x < \infty; \quad \sigma > 0$$

b) **Distribución Exponencial**. Esta distribución está naturalmente asociada a la distribución Poisson; su único parámetro suele denotarse por λ. Una variable X exponencial toma valores no-negativos y puede representar los tiempos que separan la aparición de ocurrencias de una distribución Poisson, como se muestra más adelante. El rango de valores de X y su función de densidad de probabilidad (fdp) son:

$$f(x) = \lambda e^{-\lambda x}; \quad x \geq 0; \quad \lambda > 0$$

c) **Distribución Gamma**. Esta distribución puede considerarse generalización de la distribución exponencial. Tiene dos parámetros, α de forma y β de escala. Cuando α es entero, se conoce como *Distribución Erlang,* la que permite modelar el tiempo de espera a la aparición de un número entero α de ocurrencias de una distribución Poisson de parámetro β. El rango de valores de X y su función de densidad de probabilidad (fdp) son:

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-(x/\beta)}; \quad x \geq 0; \quad \alpha, \beta > 0$$

Donde Γ indica la función Gamma, definida como: $\Gamma(x) = \int_0^{\infty} e^{-t}\, t^{x-1} dt$.

d) **Distribución Lognormal**. Esta distribución se aplica si la variable aleatoria X tiene la propiedad de que su logaritmo natural tiene una distribución normal de parámetros μ y σ. La variable X lognormal toma valores positivos, y su histograma puede mostrar formas sesgadas con colas alargadas. Esta distribución tiene diversas aplicaciones con buenos resultados. El rango de valores de X y su función de densidad de probabilidad (fdp) son:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2}\left(\frac{Ln(x)-\mu}{\sigma}\right)^2}; \quad x, \sigma > 0$$

e) **Distribución Weibull**. Esta distribución fue propuesta por el matemático sueco W. Weibull en 1951, en un desarrollo teórico basado en la forma general que debe tener una función de distribución acumulada para una variable aleatoria continua.

En un trabajo previo relacionado con la distribución del tamaño de partículas granuladas, Rossini y Ramler en 1933 usaron una distribución que resultó ser Weibull.

Esta distribución tiene dos parámetros, λ de forma y *k* de escala, y ha encontrado diversas aplicaciones también con buenos resultados.

El rango de valores de X y su función de densidad de probabilidad (fdp) son:

$$f(x) = \frac{\lambda}{k}\left(\frac{x}{k}\right)^{\lambda-1} e^{-(x/k)^{\lambda}};\ \ x > 0;\ \ \ \lambda, k > 0$$

f) **Distribución Logística**. En esta distribución, la función de distribución acumulada (FDA) tiene la forma de *curva logística*, utilizada en modelos de crecimiento, y por ello es que se le llama de esta manera.

Una de sus aplicaciones más conocida es la *regresión logística o logit*, para estudios de elecciones discretas en diversos campos. El rango de valores de X y su función de densidad de probabilidad (fdp) son:

$$f(x) = \frac{e^{-\left(\frac{x-a}{b}\right)}}{b\left[1 + e^{-\left(\frac{x-a}{b}\right)}\right]^2};\ -\infty < x < \infty;\ -\infty < a < \infty;\ b > 0$$

g) **Distribución de Gumbel**. Esta distribución se desarrolló para estudiar los valores más grandes (extremos) en muestras de datos y originalmente se aplicó para estimar niveles de inundaciones (Evans, 2000). El rango de valores de X y su función de densidad de probabilidad (fdp) son:

$$f(x) = \frac{1}{b} e^{-\left[\frac{x-a}{b} + e^{-\left(\frac{x-a}{b}\right)}\right]};\ -\infty < x < \infty;\ -\infty < a < \infty;\ b > 0$$

En la tabla 2.1 se muestra un resumen de las distribuciones discretas con las fórmulas de media y varianza; la tabla 2.2 muestra el correspondiente resumen para distribuciones continuas.

La media de cada distribución, es el valor esperado de la correspondiente variable aleatoria, y la varianza da una medida de dispersión de los posibles valores alrededor de esta media; su raíz cuadrada es la desviación típica.

Estas fórmulas de media y varianza serán de utilidad para estimar los parámetros numéricos para una distribución dada, a partir de los correspondientes momentos de la muestra de datos que se tenga, como se verá más adelante.

Tabla 2.1. Resumen de distribuciones discretas (elaboración propia)

Distribución	Función de probabilidad	Media	Varianza
Binomial	$P[X=k]=\binom{n}{k}p^k(1-p)^{n-k}$ $n>1; 0<p<1;\ k=0,1,2,\dots,n$	np	$np(1-p)$
Geométrica (núm. de fallas)	$P[X=k]=\ p(1-p)^k$ $0<p<1;\ k=0,1,2,\dots$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Geométrica (núm. de ensayos)	$P[X=k]=\ p(1-p)^{k-1}$ $0<p<1;\ k=1,2,\dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Pascal (núm. de fallas)	$P[X=k]=\binom{r+k-1}{k}p^r(1-p)^k$ $r\ >0; 0<p<1;\ \ k=0,1,2,\dots$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$
Pascal (núm. de ensayos)	$P[X=k]=\binom{k-1}{r-1}p^r(1-p)^{k-r}$ $r\ >0; 0<p<1;\ \ k=r,r+1\ \ r+2,\dots$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$
Poisson	$P[X=k]=\frac{\lambda^k e^{-\lambda}}{k!}$ $\lambda>0; k=0,1,2,\dots$	λ	λ
Logarítmica	$P[X=k]=\frac{A\theta^k}{k}$ $A=\left[\frac{-1}{\ln(1-\theta)}\right];\ 0<\theta<1;\ k=1,2,\dots,$	$\frac{A\theta}{1-\theta}$	$\frac{A\theta(1-A\theta)}{(1-\theta)^2}$

Tabla 2.2. Resumen de distribuciones continuas (elaboración propia)

Distribución	Función de densidad de probabilidad	Media	Varianza
Normal	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ $-\infty < x < \infty; -\infty < \mu < \infty;\ \sigma > 0$	μ	σ^2
Exponencial	$f(x) = \lambda e^{-\lambda x};\ x \geq 0;\ \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-(x/\beta)};$ α: forma, β: escala; $\alpha, \beta > 0; x \geq 0;$	$\alpha\beta$	$\alpha\beta^2$
Lognormal	$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2}\left(\frac{Ln(x)-\mu}{\sigma}\right)^2}$ $x, \sigma > 0;\ -\infty < \mu < \infty$	$e^{\left(\mu+\frac{\sigma^2}{2}\right)}$	$e^{2\sigma^2+2\mu} - e^{\sigma^2+2\mu}$
Weibull	$f(x) = \frac{\lambda}{k}\left(\frac{x}{k}\right)^{\lambda-1} e^{-(x/k)^{\lambda}}$ λ: forma, k: escala; $\lambda, k > 0;\ x > 0$	$k\Gamma\left(1+\frac{1}{\lambda}\right)$	$k^2\Gamma\left(1+\frac{2}{\lambda}\right) - \left[\Gamma\left(1+\frac{1}{\lambda}\right)\right]^2$
Logística	$f(x) = \frac{e^{-\left(\frac{x-a}{b}\right)}}{b\left[1+e^{-\left(\frac{x-a}{b}\right)}\right]^2}$ $-\infty < x < \infty;\ -\infty < a < \infty;\ b > 0$	a	$\frac{b^2\pi^2}{3}$
Gumbel	$f(x) = \frac{1}{b} e^{-\left[\frac{x-a}{b} + e^{-\left(\frac{x-a}{b}\right)}\right]}$ $-\infty < x < \infty;\ -\infty < a < \infty;\ b > 0$	$a + \gamma b$ $\gamma \approx 0.5772$	$\frac{b^2\pi^2}{6}$

2.2 Aspectos del cálculo de probabilidades

En un problema práctico, ya que se ha identificado una distribución de probabilidad adecuada para los datos que se tienen a la mano, es común que se quiera calcular la probabilidad de ocurrencia de un evento de interés. En estos cálculos es de utilidad la función de distribución acumulada (FDA) F que se define por la condición: $F(k) = P[X \leq k]$, que es la probabilidad de que la v.a. X no rebase el valor *k*.

Ejemplo 2.1 (adaptado de Schwar y Puy, 1975). En una intersección semaforizada con tiempo de ciclo de 60 seg, en promedio 200 vehículos/hora dan vuelta a la izquierda. Sabiendo que cada ciclo acomoda 3 vueltas a la izquierda, se quiere estimar el porcentaje de los ciclos en los que hay demora.

Si X = número de vueltas a la izquierda, puede modelarse como variable Poisson y cada ciclo en promedio tiene:

$$\frac{60seg}{ciclo} \times \frac{200\frac{veh}{hora}}{3600\frac{seg}{hora}} = \lambda = 3.33\ veh/ciclo$$

Entonces, la probabilidad de que *k* vehículos den vuelta a la izquierda es:

$$P[X = k] = \frac{3.33^{\,k} e^{-3.33}}{k!}$$

Y la demora ocurre cuando X > 3; la probabilidad que interesa es:

$$P[X \geq 4] = \sum_{n=4}^{\infty} \left(\frac{3.33^{\,n} e^{-3.33}}{n!} \right) = \frac{3.33^{\,4} e^{-3.33}}{4!} + \frac{3.33^{\,5} e^{-3.33}}{5!} + \frac{3.33^{\,6} e^{-3.33}}{6!} + \cdots$$

Este cálculo se simplifica tomando la diferencia entre 1 y el evento complemento:

$$P[X \geq 4] = 1 - \sum_{n=0}^{3} \left(\frac{3.33^{\,n} e^{-3.33}}{n!} \right) =$$

$$1 - \left[\frac{3.33^{\,0} e^{-3.33}}{0!} + \frac{3.33^{\,1} e^{-3.33}}{1!} + \frac{3.33^{\,2} e^{-3.33}}{2!} + \frac{3.33^{\,3} e^{-3.33}}{3!} \right] \approx 0.4269$$

Esto indica que aproximadamente el 42.7% del tiempo, hay demoras en la vuelta a la izquierda.

Ejemplo 2.2. La figura 2.1 es una muestra cruces de vehículos en la frontera México-Estados Unidos con datos de 2013. El histograma sugiere que X = número de cruces mensuales luce como una Normal, que se estima con una media μ = 34,848 y una desviación estándar σ = 2,566; y aunque X es un número entero, se obtiene una representación razonable.

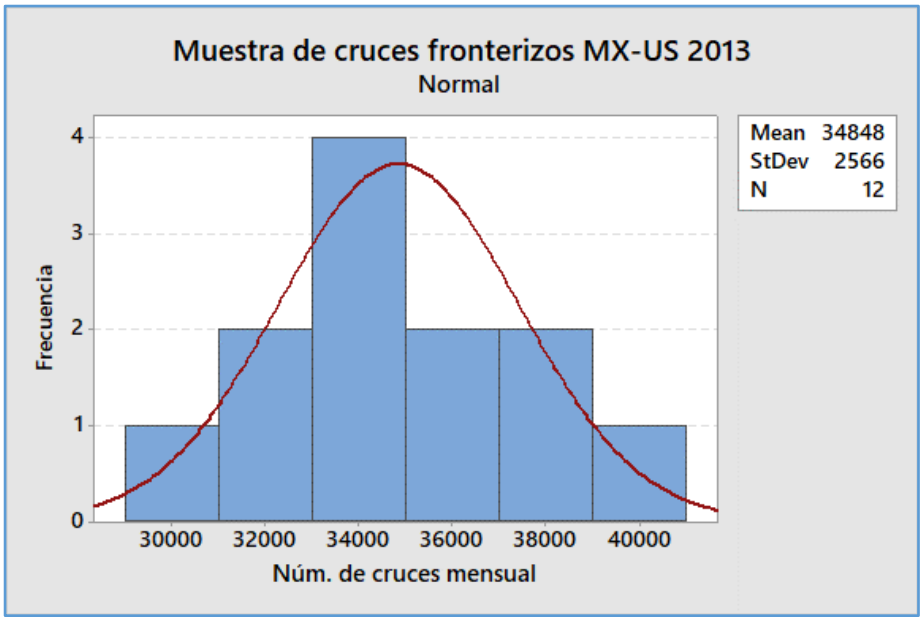


Figura 2.1. Histograma y ajuste de curva normal de cruces fronterizos. (elaboración propia)

Si X representa el comportamiento típico de los cruces hacia los E.U.A., y digamos que interesa saber qué porcentaje del tiempo, el número de cruces rebasa los 36,000 vehículos, la respuesta está en la probabilidad de que X > 36,000.

Para obtener esa probabilidad se aplica la siguiente transformación:

$$P[X > 36000] = P\left[\frac{X-34848}{2566} > \frac{36000-34848}{2566}\right] = P[\varphi > 0.4489] = 0.3267$$

Donde φ corresponde al argumento de la densidad de probabilidad Normal estándar de media 0 y desviación estándar 1, y la probabilidad mostrada se localizó en una tabla para dicha distribución. Entonces, aproximadamente el 32.7% del tiempo, el número de cruces mensuales serán mayores de 36,000.

Las funciones estadísticas de Excel

Los dos ejemplos previos ilustran que el cálculo de probabilidades en un caso concreto puede implicar varias operaciones, algunas algo laboriosas y que requieren atención para su ejecución.

Para facilitar estos cálculos es conveniente utilizar el software Excel de Microsoft el cual ofrece ya varias rutinas de cálculo de probabilidades de algunas distribuciones, en el estilo de hoja de cálculo.

De este modo, la probabilidad calculada del ejemplo 2.1 se efectúa en Excel con la instrucción siguiente:

= 1 – POISSON.DIST(3,3.333,1)

En este cálculo a 1 se le resta la probabilidad $P[X \leq 3]$ dando el resultado 0.4269.

La figura 2.2 ilustra el significado de los argumentos de la función POISSON.DIST:

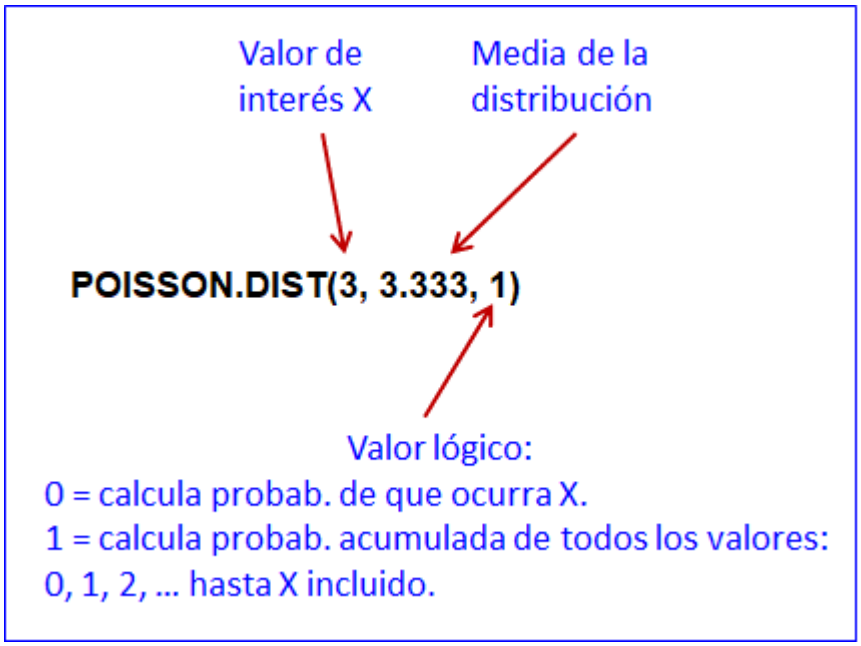


Figura 2.2. Sintaxis de función estadística en Excel. (elaboración propia)

La sintaxis de las funciones estadísticas de Excel, es semejante en todos los casos.

La tabla 2.3 resume las distribuciones de probabilidad que están disponibles en Excel, con algunas notas para un correcto uso de ellas. En cada caso, *el valor lógico se denota por* ***acu.***

Para distribuciones discretas, con *acu* = 0 se calcula la probabilidad de ocurrencia del valor *k* de interés; con *acu* = 1 se calcula la probabilidad acumulada de los posibles valores hasta el valor *k* incluido.

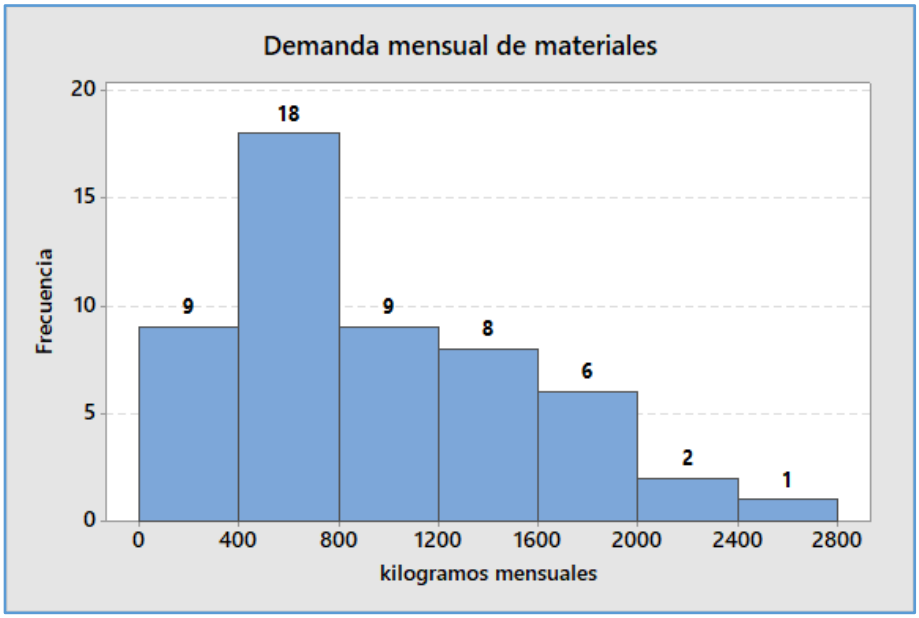
Para distribuciones continuas, con *acu* = 0 se calcula la función de densidad de probabilidad (fdp) justo en el valor *x* de interés; con *acu* = 1 se calcula la probabilidad acumulada de los posibles valores hasta el valor *x* incluido.

Tabla 2.3. Funciones estadísticas en Excel (elaboración propia)

Distribución	Parámetros y variable	Función de Excel
Binomial	n entero $> 1; 0 < p < 1;\ k = 0, 1, 2, \dots, n$	**DISTR.BINOM.N(k, n, p, acu)**
Geométrica *(núm. de fallas)*	$0 < p < 1;\ k = 1$ Se resuelve como Pascal con un solo éxito.	**NEGBINOM.DIST(k,1,p,acu)** *Excel sólo tiene la versión para la v.a. = núm. de fallas*.
Pascal *(núm. de fallas)*	$r > 0; 0 < p < 1;\ \ k = 0, 1, 2, \dots$	**NEGBINOM.DIST(k,r,p,acu)**
Poisson	$Media: \lambda > 0; k = 0, 1, 2, \dots$	**POISSON.DIST(k, Media, acu)**
Normal	$-\infty < x < \infty$; Media: μ; Desv. estándar: σ $-\infty < \mu < \infty;\ \sigma > 0$	**DISTR.NORM.N(x, μ, σ, acu)**
Normal Estándar	$-\infty < x < \infty;$ Media $= 0$; Desv. estándar $= 1$	**DISTR.NORM.ESTAND(x)**
Exponencial	$x \geq 0;$ Parámetro: $\lambda > 0$	**DISTR.EXP.N(x, λ, acu)**
Gamma	α: forma, β: escala; $\alpha, \beta > 0; x \geq 0;$	**DISTR.GAMMA.N(x,α,β,acu)**
Lognormal	$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2}\left(\frac{Ln(x)-\mu}{\sigma}\right)^2}$ $x, \sigma > 0;\ -\infty < \mu < \infty$	**DISTR.LOGNORM(x, μ, σ, acu)**
Weibull	$f(x) = \frac{\lambda}{k}\left(\frac{x}{k}\right)^{\lambda-1} e^{-(x/k)^{\lambda}}$ λ: forma, k: escala; $\lambda, k > 0;\ x > 0$	**DISTR.WEIBULL(x, λ,k, acu)**

Algunos ejemplos adicionales ilustran el uso de las funciones de Excel.

Ejemplo 2.3. Datos de la demanda mensual *X* (kg) de materiales para un tipo de transformador que produce una empresa se muestran en el siguiente histograma.



La forma sesgada del histograma sugirió un ajuste estadístico para una distribución Gamma, resultando adecuados los siguientes parámetros (Gutiérrez-González, E. et al, 2013):

$$\alpha\ (\text{forma}) = 0.638; \quad \beta\ (\text{escala}) = 1358.123$$

Con estos valores, la demanda esperada, o demanda mensual media resulta ser:

$$\alpha \times \beta = 0.638 \times 1358.123 = 866.482\ \text{kg}$$

Si se desea estimar la probabilidad de que la demanda X no supere, por ejemplo, los 1,500 kg, el cálculo con Excel es:

$$P[\ X \leq 1500] = \text{DISTR.GAMMA.N}(1500,0.638,1358.123,1) = 0.8126.$$

Es decir, con probabilidad 81.26% la demanda no supera los 1,500 kg. Esto también puede interpretarse como que la demanda no supera ese valor el 81.26% del tiempo.

Ejemplo 2.4. Los siguientes datos de vuelos de Aeroméxico llegando al Aeropuerto Internacional de la Ciudad de México (AICM) se registraron en septiembre/2012 en el horario de las 06:00am a las 11:59 am.

LLEGADAS DE AEROMÉXICO AL AICM - SEP/2012							
VUELO	PROGRAMADO	HORA_REAL	DESTINO	VUELO	PROGRAMADO	HORA_REAL	DESTINO
AMX 646	06:19	06:08	LOS ANGELES	AMX 484	10:01	10:01	LAS VEGAS,NEV
AMX 702	06:25	06:17	HERMOSILLO	AMX 019	10:10	10:15	LOS ANGELES
AMX 170	06:36	06:18	TIJUANA	AMX 402	10:13	10:41	NEW YORK
AMX 408	07:01	07:01	NEW YORK	AMX 464	10:21	10:12	HERMOSILLO
AMX 690	07:53	07:37	CANCUN, Q.R.	AMX 9320	10:36	09:48	ZIHUATANEJO
AMX 498	08:00	07:54	MONTERREY NL.	AMX 1806	10:36	11:32	GUADALAJARA
AMX 539	08:05	08:13	CANCUN, Q.R.	AMX 515	10:39	10:40	MERIDA, YUC.
AMX 597	08:55	08:52	CANCUN, Q.R.	AMX 907	11:04	10:57	CANCUN, Q.R.
AMX 412	09:23	09:02	MIAMI	AMX 531	11:18	11:19	MERIDA, YUC.
AMX 184	09:30	09:00	TIJUANA	AMX 482	11:27	11:27	LAS VEGAS,NEV
AMX 434	09:31	09:21	ORLANDO	AMX 507	11:43	11:33	VILLAHERMOSA
AMX 030	09:36	09:48	BUENOS AIRES	AMX 436	11:46	11:42	ORLANDO
AMX 686	09:40	09:42	CHICAGO				
AMX 914	09:56	10:00	LA HABANA, CU				
AMX 451	09:56	10:12	LA HABANA, CU				

El conteo de X = número de llegadas por hora (6:00 a 6:59, 7:00 a 7:59, etc.) se ajustó a una distribución Poisson de media λ = 4.5 llegadas/hora.

Para estimar la probabilidad de que en un periodo cualquiera de una hora entre las 06:00 am y las 11:59 am lleguen 6 o más vuelos de Aeroméxico, el cálculo es:

$$P[X \geq 6] = P[X > 5] = =1 – \text{POISSON.DIST}(5,4.5,1) = 0.2971$$

Si se desea estimar la probabilidad de que en un periodo determinado de una hora lleguen *exactamente* seis vuelos, el cálculo es:

$$\text{POISSON.DIST}(6,4.5,0) = 0.1281.$$

Estas estimaciones pueden interpretarse como que en aproximadamente el 29.71% de las veces, el número de llegadas por hora en ese horario excede a cinco; mientras que en alrededor de 12.81% del tiempo, se tendrán exactamente seis llegadas en un periodo cualquiera de una hora del mismo horario.

Ejemplo 2.5 (adaptado de Schwar y Puy, 1975). Una flotilla de 20 camiones llega a una terminal a cargar sus embarques entre las 09:00 am y las 10:00 am. Cada camión permanece en la terminal sólo 6 minutos, y la probabilidad de que un camión llegue en cualquier instante de ese periodo es uniforme.

Si se quiere tener una probabilidad de alrededor del 95% de que ningún camión haga fila de espera, ¿cuántas plataformas de carga debería tener la terminal?

Considerando el uso de la terminal por un camión cualquiera como un "éxito" y dado que esto ocurre en 6 minutos de los 60 disponibles del horario de carga, la probabilidad de éxito es 6/60 = 1/10. Como hay 20 camiones en el proceso, el número X de camiones en la terminal puede modelarse con una distribución Binomial de $n = 20$ *y* $p = 0.10$.

Para la condición de que con probabilidad 0.95 o equivalentemente, el 95% de las veces, los camiones no esperen en fila, se requiere hallar el valor *k* que cumple:

$$P[\,X \leq k] = 0.95$$

Usando la instrucción DISTR.BINOM.N(*k*, 20, 0.1, 1) se calcula la probabilidad acumulada de la distribución Binom (20, 0.10) hasta el valor *k* en la siguiente tabla:

Valor *k*	P[X ≤ *k*]
0	0.12158
1	0.39175
2	0.67693
3	0.86705
4	0.95683
5	0.98875
6	0.99761

La tabla indica que la probabilidad de tener en la terminal desde cero hasta $k = 3$ camiones es 0.86705, y que para $k = 4$ camiones, la probabilidad es de 0.95683. Por tanto, se recomienda tener cuatro plataformas de carga en la terminal. De esta manera en aproximadamente el 95.68% del tiempo los camiones no tienen que esperar.

3 Aplicaciones comunes de las distribuciones

Cuando se tiene un problema real con datos de tipo aleatorio, y se busca una representación adecuada para ellos, surge la necesidad de proponer alguna distribución de probabilidad que se acomode a la situación.

Las descripciones de los experimentos prototipo que se usan en los textos universitarios para motivar algunas distribuciones discretas, como es el caso de la Binomial, que repite *n* experimentos donde el "éxito" ocurre con probabilidad *p*, o la Geométrica, donde se repite el experimento tantas veces como se requiera para lograr el "éxito" pueden servir de guía en algunos casos en los que la situación es análoga al experimento prototipo.

Por ejemplo, en los aforos vehiculares que se hacen en periodos determinados del día, si la tasa promedio de conteo no varía mucho en el periodo, y los intervalos de observación son adecuados (10 a 15 min), se tienen condiciones cercanas a los supuestos con los que se desarrolla la distribución Poisson. Por eso, esta distribución es una buena propuesta para modelar datos de conteo, y se ha referido como un buen modelo de llegadas en el ámbito de la ingeniería del transporte.

En otros casos, la sugerencia de utilizar cierta distribución para los datos que se tienen viene de la experiencia reportada por otros investigadores que la han usado con datos de la misma naturaleza. Así, por ejemplo, en los estudios de confiabilidad de sistemas o equipos industriales, donde se colectan tiempos de vida de componentes o tiempos a la falla de los equipos, las distribuciones Exponencial o Weibull han demostrado resultar con buenos ajustes a los datos.

En este capítulo se hace un resumen de experiencias exitosas de aplicación de algunas distribuciones de probabilidad, a fin de identificar las posibilidades de su uso, según la naturaleza o tipo de datos que se trata de ajustar.

3.1 Distribuciones discretas

La tabla 3.1 se muestran aplicaciones reportadas de diversas fuentes para algunas de las distribuciones discretas de probabilidad mencionadas en el capítulo anterior.

En el Anexo al final de este trabajo están las referencias de los artículos citados para los interesados en consultar a los autores.

Tabla 3.1. Aplicaciones de distribuciones discretas (elaboración propia)

Distribución	Resultado / Aplicaciones
Binomial	Calcula la probabilidad de lograr *k* éxitos al repetir *n* ensayos independientes, cada uno con probabilidad *p* de éxito.
Se ha utilizado en *Control de Calidad*, y problemas de *Muestreo y encuestas.* • Friedman et al (1983), usan un modelo binomial para evaluar la calidad de datos de análisis químicos en laboratorios ambientales. • Lee et al (2010) la usaron para identificar orígenes de tipos de grano en un sistema de rastreo para las agroindustrias de granos. • Sexauer, McBee & Bloch (2011) usan un modelo binomial para calcular la probabilidad de que un transformador eléctrico tenga sobrecarga. • Kozik (2014) reporta el uso de la distribución binomial para interpretar el espectro 31-P NMR del ácido Alfa-dodecafosfotúngstico, resultante de una combinación de 13 espectros de isótopos componentes.	
Poisson	Calcula la probabilidad de tener *exactamente k* ocurrencias de un evento (análogo al "éxito" descrito en la Binomial) en un lapso de observación dado, si los eventos ocurren a una tasa media λ; esta tasa λ coincide con la media y la varianza Poisson.
Se aplica para *conteo* de eventos de interés cuando se puede en principio, contar cualquier número de eventos (a diferencia de la Binomial, que cuenta a lo más *n*). El lapso de observación suele ser tiempo (p. ej. aforo vehicular en un crucero), pero también aplica en espacios físicos (p. ej. núm. de defectos por metro lineal en un cable, o el número de defectos por m^2 en telas para asientos automotrices). • Gerlough, D. L. (1955) reporta aplicaciones de la distribución Poisson en: análisis de tasas de llegadas de vehículos a un punto dado; estudios de espaciamiento entre vehículos; cálculo de probabilidades de encontrar un lugar libre de estacionamiento y el estudio de cierto tipo de accidentes. • Al-Gahmdi, A.S. (2000) desarrolla una metodología para clasificar puntos viales de alta accidentalidad con base en una distribución Poisson. • Letkowski, J. (2012) reporta varias estimaciones con base Poisson, como: el número de clientes por hora llegando a un autolavado; el número diario de fallas en una red de cómputo; el número	

de reparaciones requeridas por cada 10 millas de carretera, o el número de fugas por cada 100 millas de oleoducto.

- Baldeh, M. et al. (2016) aplican la distribución Poisson al estudio del proceso de desbordamiento de aguas pluviales.

Binomial negativa	Calcula la probabilidad de tener *k* fallas antes de acumular *r* éxitos al repetir ensayos independientes con probabilidad *p* de éxito. Si *r* = 1, es la distribución **Geométrica**

Se ha usado en problemas de conteo análogos a los de la Poisson, en los que la varianza resulta bastante mayor que su media (rasgo conocido como sobredispersión).

- Taylor, C.J. (1961). Aplica una distribución binomial negativa para modelar el número de pedidos en un control de inventarios.
- Simon, L. J. (1962). Reporta modelos de binomial negativa para problemas de control de calidad de defectos en lotes de producción y también en conteo de accidentes viales.
- Poch, M. and Mannering, F. (1996). Modelan accidentes viales en intersecciones con base en una binomial negativa.
- Zamani H. & Ismail N. (2010). Usan la binomial negativa combinada con una distribución de Lindley para modelar el número de reclamaciones a aseguradoras.
- Fahidy, T. Z. (2012). Aplica la binomial negativa en varios experimentos de electroquímica, para el conteo de ensayos exitosos que contienen porcentajes especificados de producto.

Logarítmica	Calcula la probabilidad de contar *k* apariciones de un evento de interés, con base en el desarrollo en serie de potencias de la función logaritmo.

Aunque la inspiración de su función de probabilidad luce como un interés puramente matemático, se ha aplicado en problemas de conteo donde puede obtenerse cualquier número mayor a cero.

- Andreassen, D.C. (1986). Logra un buen ajuste a una distribución logarítmica al analizar frecuencias de accidentes en Australia.
- Deni, S.D. and Jemain, A.A. (2009). Combinan una distribución logarítmica con una geométrica en un estudio de secuencias de días lluviosos y días secos en Malasia.
- TutorVista.com, un servicio de tutoría universitaria en línea, reporta como aplicaciones de la distribución logarítmica, la modelación de reclamaciones en aseguradoras y el conteo del número de productos que adquieren los consumidores en cierto periodo de observación.

3.2 Distribuciones continuas

Enseguida se muestran aplicaciones reportadas de algunas distribuciones continuas, referidas previamente con referencias de los artículos citados.

Tabla 3.2. Aplicaciones de distribuciones continuas. (elaboración propia)

Distribución	Resultado / Aplicaciones
Normal	Su función de densidad se caracteriza por dos parámetros: media μ y desviación estándar σ; tiene forma acampanada, y es simétrica respecto a su media.
Es una distribución muy usada; suele aplicarse a datos con histogramas de forma acampanada y simétrica respecto a la media. Aproxima bien a la Binomial y a la Poisson cuando en estas es difícil el cálculo práctico. Su relevancia en buena parte se debe al Teorema Central del Límite, que indica que los promedios muestrales casi siempre tienen comportamiento Normal. Muchos experimentos producen datos Normales: los errores de medición, las longitudes de componentes industriales producidos en masa, el peso de personas del mismo sexo y grupo de edad (p. ej. pasajeros de autobús), etc. (Upton, G. & Cook, I, 2002). • El Departamento de Transporte del Estado de Washington, E.U.A., WSDOT (1994) reporta el uso de la Normal para caracterizar comportamiento de: datos de densidad Proctor, datos de esfuerzos a compresión de Concreto de Cemento Portland y datos de mezclas de concreto. • Boll, J. et al, (1997). Usan una distribución normal para estudiar flujos de agua y material disuelto en diversos sitios de medición en los estados norteamericanos de Nueva York y Delaware. • Hashim, I.H. (2011), usa la Normal con datos de velocidades en caminos rurales de dos carriles en Egipto. • Ruimin, L., Chai, H. and Tang, J. (2013) usan una Normal para caracterizar la distribución de tiempos de recorrido en distintos tipos de vialidades urbanas con datos de placas de automotores en Beijing. • Büchel, B. & Corman, F. (2018), usan modelos Normales para describir tiempos de viaje en autobuses públicos, para datos agregados por sección, por ruta y para periodos dentro y fuera de pico.	
Exponencial	Se aplica con datos continuos no negativos donde el histograma muestra una tendencia decreciente, indicando que es más probable obtener valores

	pequeños de la variable aleatoria, que valores grandes. Es el modelo básico para el tiempo de espera a que ocurra un evento de interés.

Se relaciona con la Poisson; si un proceso de conteo (variable discreta) obtiene datos Poisson de parámetro λ, *el tiempo entre ocurrencias consecutivas del evento que se cuenta* (variable continua) es Exponencial de media $1/\lambda$. Por eso, la Exponencial se recomienda como modelo básico de tiempos entre llegadas en la literatura de ingeniería del transporte.

- Drâgulescu, A. and Yakovenko, V.M. (2001) utilizan una Exponencial para modelar la distribución del ingreso de la población en los Estados Unidos.
- Harrris, M. (2006), propone una distribución Exponencial como modelo de los tiempos de retraso en el sistema ferroviario británico.
- Mehri, H., Djemel, T. and Kammoun, H. (2006) modelan las líneas de espera en un aeropuerto como una cola donde los tiempos de servicio y entre llegadas son exponenciales.
- Dauxois, J.Y., Jomhoori, S. ad Yousefzadeh, F. (2014) utilizan una exponencial para modelar los tiempos de demora en procesos de mantenimiento industriales.

Gamma	La Gamma se considera una generalización de la Exponencial de media $1/\lambda$ ya que mientras la Exponencial modela el tiempo de espera del evento de interés, la Gamma modela el tiempo de espera a que ocurra el *k-ésimo* evento.

Es útil en situaciones donde el tiempo relevante es el de espera a que ocurra un número entero de eventos. Se ha aplicado en problemas de confiabilidad, en teoría de colas y para modelar precipitación pluvial diaria en varias regiones del mundo.

- Singh, A., Singh, A. K. and Iaci, R.J. (2002), usan una Gamma para mejorar la estimación del límite superior del intervalo de 95% de confianza para la media de las concentraciones de exposición de contaminantes en los Estados Unidos.
- Krishnamoorty, K. (2006) reporta el uso de la Gamma para modelar niveles de lluvia, considerando que la precipitación se da solamente cuando las moléculas de agua se adhieren a partículas de polvo de masa suficiente, y el tiempo para que pase eso es análogo al tiempo de espera a que ocurra un cierto número de eventos de interés.
- Husak, G.J., Michaelsen, J. and Funk, C. (2006) usan una distribución Gamma para modelar niveles de lluvia en África.

<table>
<tr><td colspan="2">Knecht, W.R. (2015) usa una distribución Gamma para modelar la tasa de accidentes de aviación como función del número total de horas de vuelo de los pilotos.</td></tr>
<tr><td>Lognormal</td><td>La distribución Lognormal surge naturalmente en problemas donde la variable aleatoria de interés sólo tiene valores positivos y el histograma tiene un sesgo notable a la derecha. Puede ser una alternativa al uso de la Gamma.</td></tr>
<tr><td colspan="2">En procesos aleatorios donde la v. a. de interés X resulta del multiplicar un gran número de otras v. a. positivas independientes, la Lognormal suele ser una buena representación para X.
Limpert, E; Stahel, W and Abbt, M. (2001) reportan diversas aplicaciones de la Lognormal, por ejemplo: Geología (concentraciones de elementos y su radioactividad); Medio ambiente (distribución de partículas, productos químicos y organismos en el medio ambiente) y Ciencias de la atmósfera (distribución de tamaños de aerosoles y nubes).
Cabrera, J.B.D. et al. (2004) usan la Lognormal en un modelo para tiempos de procesamiento de datos en un sistema de información computarizado.
Krishnamoorty, K. (2006) reporta el uso de la Lognormal en varias aplicaciones como: datos GPS de posición de vehículos seleccionados; velocidades de los vientos; datos de exposición a contaminantes y tamaños de gotas de lluvia.
Yuan, J; Goverde, R.M.P. and Hansen, I.A. (2006) aplican una Lognormal para modelar tiempos de arribo de trenes a plataforma y de aproximación a la señalización en el ferrocarril holandés.
Jiao, T; Wen, X and Wang, X. (2013) usan una distribución Lognormal para modelar valores de índices de medición de compactación de suelos.</td></tr>
<tr><td>Weibull</td><td>Es una de las distribuciones más utilizadas en estudios de confiabilidad de equipos y de sistemas. Es una alternativa a la Exponencial, la Gamma y a la Lognormal para modelar tiempos de espera a la ocurrencia de un evento de interés.</td></tr>
<tr><td colspan="2">Esta distribución fue propuesta por Waloddi Weibull en 1951, y se corresponde con la distribución que utilizaron Rosin y Rammler en 1933 para describir el tamaño de partículas de materiales granulados.
Brown, W.K. and Wohletz, K.H. (1995) obtienen una distribución Weibull con base en principios físicos en los procesos de obtención de material granulado y verifican la relación de la Weibull con los trabajos previos de Rossin-Rammler.</td></tr>
</table>

- Zobeck, T.M; Gill, T.E and Popham, T.W. (1999) usan una distribución Weibull para describir el tamaño de partículas de polvo en el aire.
- Jiang, R. and Murthy, D.N.P. (2011) revisan el uso de distribuciones Weibull en estudios de confiabilidad para modelar tiempos de vida, edades de reemplazo y vida residual en equipos y componentes.
- Razali, A.M. et al (2008), modelan datos de velocidades de vientos con una distribución Weibull.

Logística	Esta distribución es llamada así por la forma de su función de distribución acumulada, que es la curva logística. La curva de densidad se parece a la Normal, pero con colas más largas, y puede ser una alternativa para la Normal.

Se ha utilizado en regresión logística, con modelos Logit para estudios de elección discreta de usuarios enfrentados a alternativas (p. ej. decidir entre autobús y tren, bajo diferentes ofertas de precio y tiempo).

- Van Beek, P. (1978) utiliza una distribución Logística para modelar una política de revisión continua para control de inventarios.
- Rajkai, K. et al. (1996) modelan con una distribución Logística características de retención de agua a partir de datos de densidad de suelos en Suecia.
- Cooray, K. (2005) analiza datos de tiempos de vida con una distribución Logística modificada que le permite un manejo mejorado de colas muy largas en la distribución de los datos.
- Krishnamoorty, K. (2006) describe el uso de la Logística en comparaciones con la Weibull para estudios de velocidades de viento.
- Bowling, S.R. et al (2009), desarrollan un método de aproximación a la función de distribución acumulada $\Phi(x)$ de la Normal, usando una distribución Logística.

Gumbel	Llamada también distribución de valores extremos, surgió de los trabajos de Julius Gumbel (1935) en estudios de valores extremos (el máximo o el mínimo) de datos muestrales.

Se ha usado en diversos campos donde interesa estimar probabilidades de que la v. a. de interés alcance o rebase valores de referencia dados (p. ej. en hidrología para niveles de lluvia o sequía; en el ambiente industrial la estimación del mínimo tiempo de ocurrencia a una falla en un equipo).

- Önoz, B. and Bayazit, M. (1995), reportan el uso de la distribución Gumbel como la de mejor desempeño para modelar niveles máximos de inundaciones con datos de varias regiones del mundo.
- Xu, Y.L. (1995) utiliza una distribución Gumbel para modelar las características de fatiga por presiones de viento en un edificio de una universidad de Texas.
- Jánosikova, L. and Slavik, M. (2014) modelan la llegada de pasajeros a paradas de autobús de transporte público en la República Eslovaca utilizando una distribución Gumbel.
- Rehman, K; Burton, P.W. and Weatherill, G.A. (2018) utilizan una distribución Gumbel y el método de simulación Monte Carlo en el análisis de riesgo sísmico en Pakistán.

3.3 Otras distribuciones más especializadas

Durante el desarrollo del presente trabajo, se hallaron referencias a otras distribuciones para problemas muy particulares, y que no aparecen en referencias típicas de aplicaciones de probabilidad. En esta sección se describen brevemente estas distribuciones y el campo de aplicación en el que surgen.

Distribución Rosin-Rammler

Esta distribución, comentada previamente por su relación con la Weibull fue usada por Rosin y Rammler en 1933 para describir la distribución del tamaño de partículas de material granulado. La función de densidad es como sigue:

$$f(x;\ P_{80}, m) = \begin{cases} 1 - e^{\ln(0.2)\left(\frac{x}{P_{80}}\right)^m} & x\ \geq 0 \\ 0 & x < 0 \end{cases}$$

donde x es el tamaño de partícula; P_{80} es el percentil 80 de la distribución del tamaño de partículas y m describe la dispersión de la distribución. (Wikipedia, 2018a)

Distribución Hiperbólica

Esta distribución fue propuesta por Bagnold y Barndorff-Nielsen en 1980, para modelar tamaños de partículas de material sedimentado (Wikipedia, 2018a). La distribución hiperbólica es tal que la gráfica del logaritmo de la densidad resulta ser una hipérbola (Bagnold & Barndorff-Nielsen, 1980). La función de densidad es como sigue:

$$f(x) = \frac{\gamma}{2\alpha\delta K_1(\delta\gamma)} e^{-\sqrt[\alpha]{\delta^2 + (x-\mu)^2} + \beta(x-\mu)}$$

con los parámetros α, β (de asimetría), δ (de escala) son todos reales, $\gamma = \sqrt{\alpha^2 - \beta^2}$ y K_1 es la función de Bessel modificada de segunda clase. (Wikipedia, 2018b).

La literatura reporta que esta distribución ha tenido el problema de generar soluciones múltiples para ciertos valores de los coeficientes que la definen, razón por la cual la distribución Log-Laplace ha sido propuesta como alternativa de mejor desempeño.

Distribución Log-Laplace

Esta distribución fue propuesta como una alternativa a la distribución hiperbólica, de la cual se encontró un desempeño pobre en algunos análisis de sedimentos (Fieller, Gilbertson and Olbricht, 1984).

La función de densidad es como sigue: (Wikipedia, 2018c).

$$f(x \mid \mu, b) = \frac{1}{2bx} e^{-\left(\frac{\ln x - \mu}{b}\right)}$$

Distribución Birnbaum-Sanders

Esta distribución surgió en la modelación de fallas en materiales debido a fracturas, en experimentos donde los materiales de prueba se someten a ciclos repetidos de esfuerzos.

Se conoce también como *distribución del tiempo a la fatiga* (fatigue life distribution). La función de densidad es como sigue:

$$f(x) = \frac{\sqrt{\frac{x-\mu}{\beta}} + \sqrt{\frac{\beta}{x-\mu}}}{2\gamma(x-\mu)} \varphi\left(\frac{\sqrt{\frac{x-\mu}{\beta}} - \sqrt{\frac{\beta}{x-\mu}}}{\gamma}\right)$$

donde los parámetros son: γ(forma); μ (localización); β (escala), y φ es la función de densidad de probabilidad de la Normal estándar. (Wikipedia, 2018d).

Una aplicación de la Birnbaum-Sanders en el estudio de transporte de sedimentos por corrientes de agua, es reportada por Turowski (2010).

Todas estas distribuciones que se han referido en el presente capítulo han aparecido en la práctica ingenieril y en temas de transporte.

La utilidad de cada una de ellas se ha determinado por los usuarios mediante pruebas estadísticas de bondad de ajuste a los datos empleados, con lo que se han obtenido modelos adecuados para estimar probabilidades y valores esperados.

Como ejemplo concreto de utilización de los diversos modelos de probabilidad en el estudio de suelos, Dianty, Yahaya y Ahmad (2014) reportan:

> "*Las propiedades ingenieriles de suelos, colectados de distintos sitios designados para construir torres de comunicaciones, fueron obtenidos y analizados. […] La razón de vacíos, el peso unitario a granel y el peso unitario seco, siguieron una distribución Normal; el contenido de agua, el límite líquido y el índice de plasticidad siguieron una distribución Gamma; la especificación de la gravedad, la porosidad, la saturación, el ángulo interno de fricción y de cohesión se ajustaron a una distribución Weibull, y el peso unitario saturado siguió una distribución log-normal."*

Este ejemplo ilustra la variedad de aplicaciones que tienen las distribuciones de probabilidad y la utilidad de tener un esquema general para elegir opciones a ensayar con los datos que se obtengan en la práctica.

Para decidir cuál de los modelos probabilísticos es el mejor para los datos que se tienen, es necesario hacer pruebas estadísticas de bondad de ajuste, las cuales se discuten en el siguiente capítulo.

4 Las técnicas de bondad de ajuste

Estas técnicas son pruebas estadísticas diseñadas para dar un criterio de aceptación o rechazo del modelo probabilístico que se haya propuesto para un conjunto de datos dado. Para elegir una distribución de probabilidad que pueda representar adecuadamente al conjunto de datos de interés, en la práctica se suele considerar lo siguiente:

1. Observar la forma de distribución que sugiere el histograma de los datos. Por ejemplo, con datos continuos, si el histograma tiene forma acampanada y simétrica respecto su parte central, las distribuciones Normal o Logística pueden ser candidatas; en cambio, si los datos son discretos, la Binomial o la Poisson pudieran funcionar.

2. Considerar información histórica de las distribuciones que se han utilizado en diversas áreas. Por ejemplo, en estudios de confiabilidad y riesgo en sistemas o equipos, las distribuciones Exponencial y Weibull son bastante usadas; en estudios de Hidrología (niveles de lluvia, caudales de ríos, etc.) la Lognormal y la Gamma se han reportado con éxito. En estudios de conteos de accidentes y aforos vehiculares, las distribuciones Poisson y Binomial Negativa han tenido también mucho uso. Las referencias de aplicaciones del capítulo anterior pueden guiar la elección.

3. Razones teóricas para algunas distribuciones. Por ejemplo, se ha demostrado que en el conteo de eventos con distribución Poisson, los tiempos que transcurren entre eventos tienen distribución Exponencial; también, al hacer muestreos repetidos de una medición y observar los promedios de cada muestra, el Teorema del Límite Central asegura que el comportamiento de los promedios resulta Normalmente distribuido.

Una vez propuesta una distribución de probabilidad para los datos, se requiere:

a) Estimar los parámetros para la distribución, utilizando los datos de la muestra; por ejemplo, si la propuesta es una Poisson o una Exponencial, hay que estimar su parámetro λ; si la propuesta es para una Lognormal., hay que estimar los valores de μ y σ que la caracterizan, etc.

b) Realizar la prueba de bondad de ajuste para los datos, suponiendo que provienen de la distribución propuesta, y decidir si se acepta o se rechaza.

Los detalles de estas tareas se discuten en las siguientes secciones.

4.1 Estimación de parámetros

La estimación de los parámetros que pueden caracterizar a la distribución elegida para los datos de interés es el primer paso para buscar el ajuste. Dos de los métodos que se usan comúnmente para estimar los parámetros de una distribución son: a) el método de momentos y b) el método de máxima verosimilitud.

En el método de momentos se igualan los momentos teóricos de la distribución propuesta con los momentos muestrales calculados con los datos por ajustar, para derivar de las ecuaciones resultantes las estimaciones de los parámetros. Este método es sencillo de plantear, aunque en ocasiones las ecuaciones resultantes deben resolverse por métodos numéricos, y si la muestra de datos es pequeña, puede producir resultados fuera del rango que se espera para los parámetros.

El método de máxima verosimilitud consiste en encontrar los parámetros de la distribución propuesta que logran optimizar la probabilidad de haber obtenido justamente la muestra de datos que se pretende ajustar. Este método es más preciso que el de momentos, pero es más complicado de aplicar, pues puede requerir métodos numéricos de optimización.

Método de momentos

El momento de orden *k* de una variable aleatoria X se define como el valor esperado de X^k, denotado por: $\mu'_k = E(X^k)$. Análogamente, el momento muestral de orden *k* del conjunto de datos $X_1, X_2, \dots, X_n$ se define por: $m'_k = \frac{1}{n}\sum_{j=1}^{n} X_j^k$. Así, resulta que $\mu'_1 = E(X^1) = E(X) = \mu$ = media de la distribución, y $m'_1 = \frac{1}{n}\sum_{j=1}^{n} X_j = \bar{X}$ = promedio muestral.

Si la v. a. X tiene un solo parámetro, al igualar los momentos, se puede estimar ese parámetro. Para ejemplificar, si se tiene la siguiente muestra de datos y se piensa que son de una distribución exponencial, X~ Expo(λ):

0.5911	0.4492	1.9330	1.7006	4.6715	5.1281
0.6877	3.5171	1.8068	3.0269	5.4919	1.2789
3.3220	1.2005	0.3136	6.2130	6.8401	0.0394
0.4528	6.2765	4.8028	0.0168	1.0646	3.2254
8.1609	4.0749	1.7330	0.1035	0.8789	0.6952

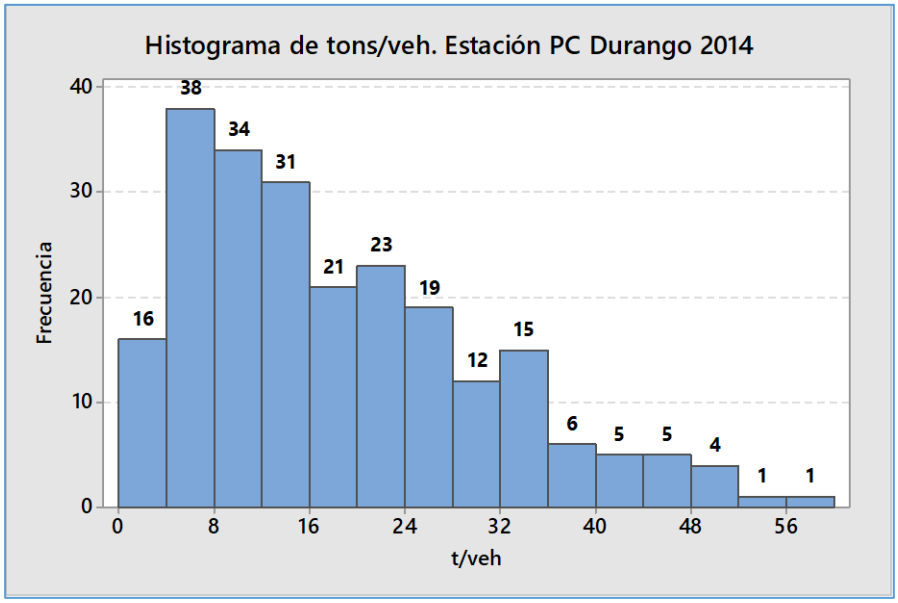
El primer momento de X, E(X) = media de X = $1/\lambda$ mientras que el primer momento muestral obtenido de los datos es el promedio $\bar{X}$ = 2.656, así que:

$$\frac{1}{\lambda} = 2.656, \text{con lo que } \lambda = \frac{1}{2.656} \approx 0.3765$$

La densidad de probabilidad propuesta para ajustar los datos es entonces:

$$f(x) = 0.3765\, e^{-0.3765x}$$

En otro ejemplo, el histograma siguiente muestra valores medios de toneladas movidas por camión, estimados con datos de la estación de muestreo PC Durango en 2014 (datos de EECAN-IMT en: https://www.imt.mx/micrositios/seguridad-y-operacion-del-transporte/estadisticas/consulta-del-eecan.html).



La forma del histograma sugiere una distribución Gamma para representar los datos. Los dos primeros momentos de una distribución Gamma (α, β) son:

$$\mu_1' = E(X) = \alpha\beta$$

Y como $\sigma^2 = E(X^2) - E(X)^2$, entonces

$$\mu_2' = E(X^2) = \sigma^2 + E(X)^2 = \alpha\beta^2 + (\alpha\beta)^2 = \alpha(\alpha+1)\beta^2$$

Del cálculo de momentos muestrales (tamaño de muestra = 231) resultó:

$$m_1' = \frac{1}{231}\sum_{j=1}^{231} X_j = 18.27 \quad m_2' = \frac{1}{231}\sum_{j=1}^{231} {X_j}^2 = 487.41$$

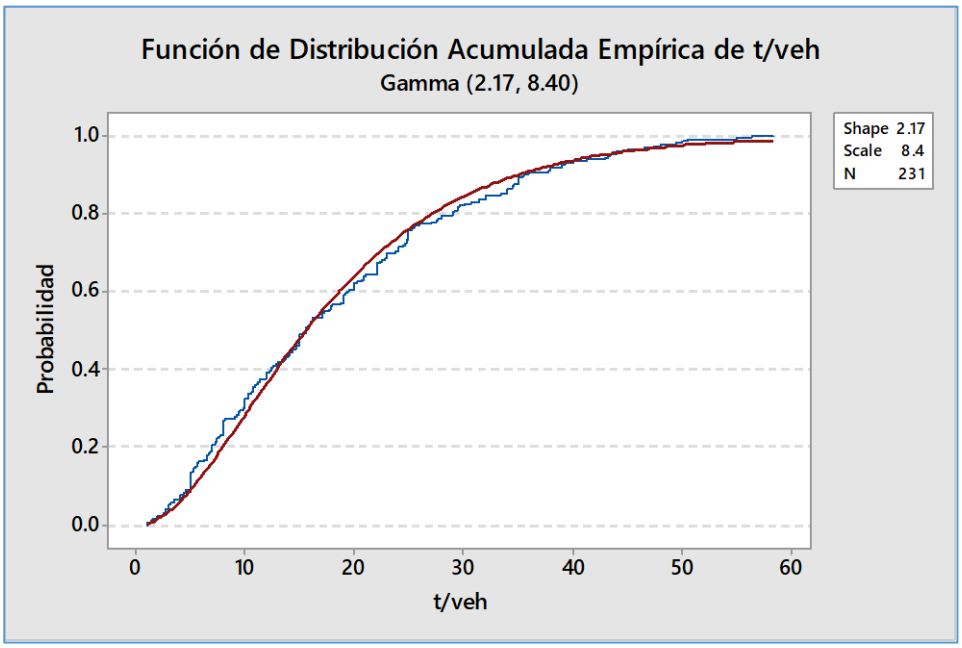
Así que el sistema de ecuaciones para estimar α y β es:

$$\alpha\beta = 18.27$$

$$\alpha(\alpha+1)\beta^2 = 487.41$$

Resolviendo estas ecuaciones simultáneas: $\alpha \approx 2.174, \beta \approx 8.404$.

La gráfica siguiente muestra la aproximación de la FDA teórica de la distribución Gamma (2.174, 8.404) propuesta con la FDA empírica calculada con los datos.



El paso siguiente para la Gamma propuesta es aplicar una prueba de bondad de ajuste para decidir la calidad estadística de su representación.

Método de Máxima Verosimilitud

Este método busca los parámetros de la distribución propuesta que maximizan la probabilidad de tener justo esa muestra de datos. En la literatura suele abreviarse con las siglas MLE (Maximum Likelihood Method) por su significado en inglés.

Por ejemplo, si se obtuvo una muestra con los datos: $X_1 = 0, X_2 = 1,\ X_3 = 1,\ X_4 = 3$ independientes, y se desea ajustarlos a una distribución Poisson de parámetro λ, la probabilidad de haber obtenido justamente esa muestra de datos depende del parámetro λ como sigue:

$$P[\text{obtener esa muestra}] = P[X_1 = 0] \times P[X_2 = 1] \times P[X_3 = 1] \times P[X_4 = 3] =$$

$$e^{-\lambda} \times \lambda e^{-\lambda} \times \lambda e^{-\lambda} \times \frac{1}{6}\lambda^3 e^{-\lambda}$$

La expresión previa es una función de λ y se llama *Función de verosimilitud* de la muestra de datos, suele denotarse con la letra *L*:

$$L(\lambda) = e^{-\lambda} \times \lambda e^{-\lambda} \times \lambda e^{-\lambda} \times \frac{1}{6}\lambda^3 e^{-\lambda}$$

El valor de λ, que maximiza el valor de $L(\lambda)$ es el que se propone para la distribución de los datos, por lo general se encuentra con técnicas de cálculo diferencial.

En este ejemplo, la expresión simplificada para la función de verosimilitud es:

$$L(\lambda) = \frac{1}{6}\,\lambda^5 e^{-4\lambda}$$

Resolviendo con la rutina de optimización en línea de *Wolfram Alpha* (https://www.wolframalpha.com/examples/mathematics/applied-mathematics/optimization/) se tiene el siguiente resultado:

Input interpretation:

maximize	function	$\lambda^5\left(\frac{1}{6}\exp(-4\lambda)\right)$
	domain	$\lambda > 0$

Global maximum:

$$\max\left\{\frac{1}{6}\lambda^5\exp(-4\lambda)\,\middle|\,\lambda>0\right\} \approx 0.0034271 \text{ at } \lambda = 1.25$$

Con lo que se verifica la existencia de máximo, y la distribución propuesta para los datos es Poisson con función de probabilidad:

$$P[X=k] = \frac{(1.25)^{\,k} e^{-1.25}}{k!}$$

El método de máxima verosimilitud da resultados más precisos que el método de momentos, pero debe resolver un problema de optimización, que cuando hay más de un parámetro, puede no tener soluciones explícitas, y requerir de técnicas numéricas. Algunas distribuciones tienen fórmulas explícitas para estimar sus parámetros con base en el método de máxima verosimilitud, las cuales se conocen como *Estimadores de Máxima Verosimilitud* (MLE por sus siglas en inglés).

La tabla 4.1 muestra fórmulas explícitas para estimar parámetros en algunas distribuciones; se usa tanto el método de momentos como el de Máxima Verosimilitud (MLE).

Los momentos muestrales calculados a partir de los datos X_j son:

$$m_1' = \frac{1}{n}\sum_{j=1}^{n} X_j$$

$$m_2' = \frac{1}{n}\sum_{j=1}^{n} {X_j}^2$$

Tabla 4.1 Estimadores de parámetros de algunas distribuciones

Distribución	Parámetros	Estimadores	Método
Binomial	$n > 1$ $0 < p < 1$	$n = \frac{m_1'^2}{m_1' + m_1'^2 - m_2'}$ $p = \frac{m_1' + m_1'^2 - m_2'}{m_1'}$	Momentos
Geométrica *(No. de fallas)*	$0 < p < 1$	$p = \frac{1}{1 + m_1'}$	MLE
Geométrica *(No. ensayos)*	$0 < p < 1$	$p = \frac{1}{m_1'}$	MLE
Pascal *(No. de fallas*	$r > 0$ $0 < p < 1$	$r = \frac{m_1'^2}{m_2' - m_1'^2 - m_1'}$ $p = \frac{m_1'}{m_2' - m_1'^2}$	Momentos
Pascal *(No. ensayos)*	$r > 0$ $0 < p < 1$	$r = \frac{m_1'^2}{m_2' + m_1' - m_1'^2}$ $p = \frac{m_1'}{m_2' + m_1' - m_1'^2}$	Momentos
Poisson	$\lambda > 0$	$\lambda = m_1'$	MLE
Normal	$-\infty < \mu < \infty$ $\sigma > 0$	$\mu = m_1';\ \sigma = \sqrt[2]{m_2' - m_1'^2}$	MLE
Exponencial	$\lambda > 0$	$\lambda = \frac{1}{m_1'}$	MLE
Gamma	$\alpha, \beta > 0$	$\alpha = \frac{m_1'^2}{m_2' - m_1'^2};\ \beta = \frac{m_2' - m_1'^2}{m_1'}$	Momentos
Lognormal	$x, \sigma > 0$ $-\infty < \mu < \infty$	De los datos X_i se calcula $Y_i = LnX_i$ $\mu = E(Y) = m_{1Y}'$ $\sigma = \sqrt[2]{E(Y^2) - E(Y)^2} = \sqrt[2]{m_{2Y}' - m_{1Y}'^2}$	MLE

4.2 Bondad de ajuste Ji-cuadrada (χ^2)

Esta técnica se basa en comparar la frecuencia observada en los datos contra la frecuencia esperada según la distribución de probabilidad propuesta para los datos. Las diferencias entre la frecuencia observada en los datos y la que se esperaría con la distribución propuesta se suman, y esta suma se aplica en una prueba de hipótesis con una distribución Ji-cuadrada, para decidir si la distribución propuesta es adecuada o no, usualmente con un nivel de significación del 5%.

Se recomienda tener al menos N= 30 de tamaño de muestra. Con muestras más pequeñas, la prueba puede funcionar, pero si hay pocos datos, los resultados son menos confiables.

Partiendo de una muestra de N datos, la prueba se desarrolla con los siguientes pasos.

1. Se colocan los N datos en una tabla de frecuencias con $m \approx \sqrt{N}$ intervalos y se obtiene la FRECUENCIA OBSERVADA FO_j en cada intervalo *j*

2. Se propone la distribución de probabilidad para los datos, (con base en el histograma de los datos, por razones técnicas o históricas, etc.), se estiman sus parámetros a partir de la muestra de datos y se anota el valor *k = número de parámetros necesarios para determinar la distribución*.

3. Con la distribución propuesta se calcula la FRECUENCIA ESPERADA FE_j de cada intervalo *j,* donde $FE_j = Probab.[el\ dato\ esté\ en\ intervalo\ j] \times N$

4. Se calcula el estimador:

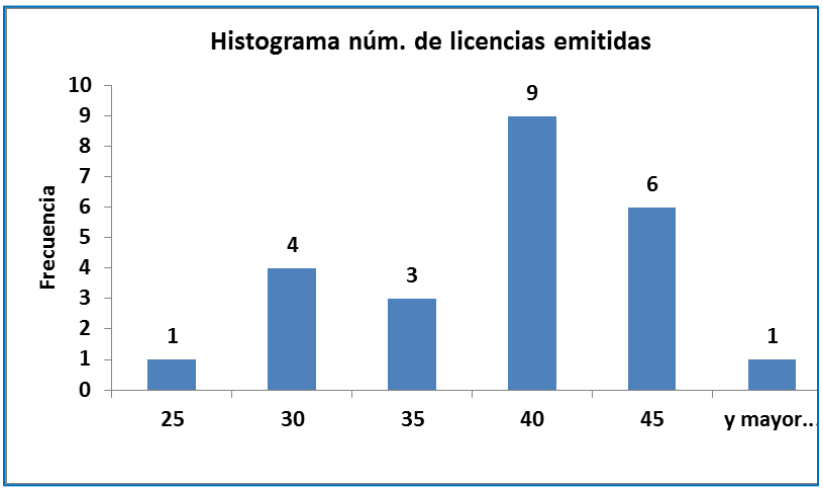
$$C = \sum_{j=1}^{m} \frac{\left(FE_j - FO_j\right)^2}{FE_j}$$

5. Si *C* es menor o igual al valor crítico de una χ^2 con $m - k - 1$ grados de libertad, y a un nivel de significación del α%, entonces no se puede rechazar la hipótesis de que los datos corresponden a la distribución propuesta.

La prueba Ji-cuadrada puede usarse tanto para distribuciones discretas como continuas. Los ejemplos enseguida ilustran la aplicación de esta técnica.

Ejemplo 4.1. (variable discreta) La emisión de licencias de operador para camiones de carga en un Centro SCT durante 24 semanas (datos SCT), y su histograma correspondiente se muestran enseguida. Interesa conocer la distribución de probabilidad que describe adecuadamente el número de licencias que se emiten cada semana.

n	licencias	n	licencias
1	37	13	39
2	44	14	31
3	37	15	40
4	30	16	27
5	44	17	38
6	39	18	43
7	34	19	23
8	40	20	43
9	30	21	58
10	37	22	42
11	29	23	41
12	36	24	35



Como la variable número de licencias es entera, y el histograma muestra una forma parecida a una distribución Poisson, se propone esta ley de probabilidad para los datos. La media muestral es $m_1' = 37.38$, que es el parámetro λ para la Poisson.

Con los datos supuestos de una distribución Poisson de parámetro λ = 37.38, la siguiente tabla muestra los cálculos. El número de intervalos se aproxima por $\sqrt{24} \approx 4.89$ por lo que se toman 5 intervalos. El número de parámetros estimados para esta distribución es *k = 1*

	Intervalo	FrecObs	FrecEsp	Ji-cuadrada
1	hasta 25	1	0.505	0.486
2	26 a 30	4	2.578	0.784
3	31 a 35	3	6.255	1.694
4	36 a 40	9	7.519	0.292
5	41 y más	7	7.143	0.003
			Suma =	**3.258**
	Grad_libertad		Ji-cuad 95%	
	3		**7.815**	

La frecuencia esperada se calcula con las probabilidades Poisson (37.38) y la función de Excel *POISSON.DIST()*.

Así, del primer intervalo P[Núm_Lics. ≤ 25] = POISSON.DIST(25,37.38,1) = 0.0210. Entonces la frecuencia esperada en las 24 mediciones es aproximadamente igual a 0.0210 × 24 ≈ 0.505. El resto de valores de frecuencia esperada se obtuvieron de modo análogo.

Los grados de libertad para la χ^2 se calculan como: $m - k - 1 = 5 - 1 - 1 = 3$.

Con el nivel de significación que suele usarse α = 5%, el valor crítico $\chi^2_{3,0.95}$ se obtuvo con la función de Excel: *INV.CHICUAD(0.95,3)* = 7.815.

Este valor crítico excede al valor de la prueba: 3.258, y se acepta la hipótesis de que los datos de emisiones de licencias de operador son Poisson (37.38).

En la figura 4.1 se ve la distribución Poisson teórica, indicando con la parte sombreada que el núm. de licencias semanales estará entre 25 y 48, en el 94.8% ≈ 95% de veces.

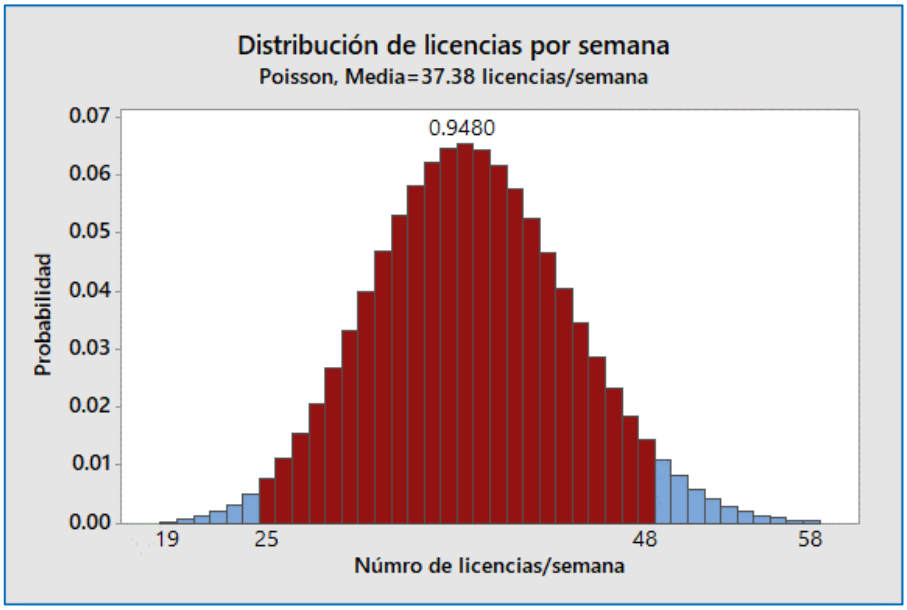
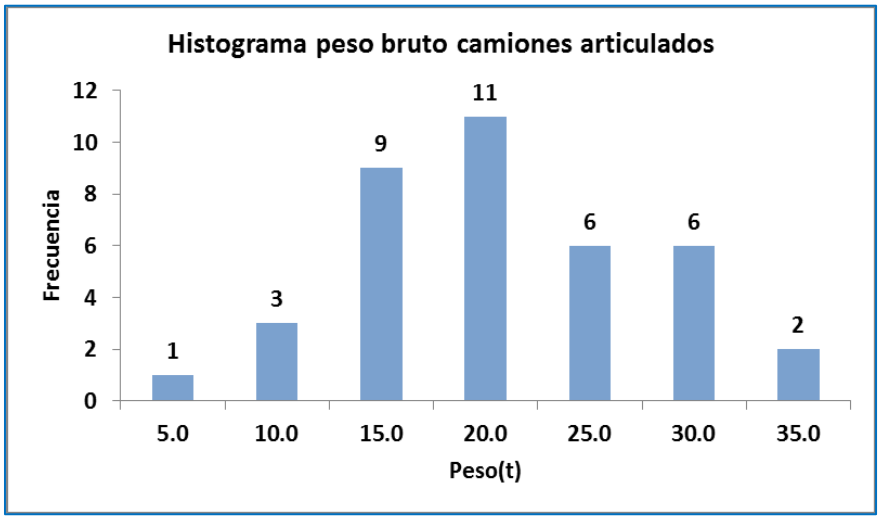


Figura 4.1. Ajuste Poisson; muestra de número semanal de licencias de operador. (elaboración propia)

Ejemplo 4.2. (variable continua) De encuestas de camino, se colectaron pesos brutos de camiones articulados en los siguientes 38 registros (datos SCT); se muestra el histograma de datos.

En este caso interesa conocer el comportamiento probabilístico del peso bruto de estos camiones.

n	Tons	n	Tons
1	16.4	20	29.6
2	7.0	21	14.2
3	26.6	22	14.1
4	20.9	23	11.7
5	16.8	24	27.7
6	18.7	25	22.4
7	18.5	26	16.0
8	4.6	27	6.5
9	10.6	28	26.2
10	10.4	29	27.9
11	8.0	30	11.9
12	16.2	31	17.2
13	13.3	32	30.9
14	27.5	33	15.2
15	16.5	34	23.3
16	13.4	35	11.9
17	15.1	36	31.0
18	15.5	37	20.6
19	22.3	38	23.6



Como el peso bruto es variable continua, y el histograma de los datos sugiere forma acampanada para la densidad, se propone una distribución Normal para el ajuste.

El número de parámetros para determinar esta distribución son dos: la media μ y la desviación estándar σ; por lo que *k = 2*. Los momentos muestrales son:

$$m_1' = 17.90; \quad m_2' = 369.76$$

La estimación de parámetros es:

$$\mu = m_1' \approx 17.90; \quad \sigma = \sqrt[2]{m_2' - m_1'^{\,2}} = \sqrt[2]{369.76 - 17.90^2} \approx 7.02$$

La distribución propuesta para los datos resulta Normal de media 17.90 y desviación estándar 7.02.

El número de intervalos se aproxima por $\sqrt{38} \approx 6.16$; tomando *m* = 7 intervalos, la tabla de cálculos es como sigue.

La frecuencia esperada se calcula con las probabilidades Normal (17.90, 7.02) y la función de Excel *DISTR.NORM.N()*. Por ejemplo, la frecuencia esperada del intervalo de 5 a 10 se calcula como:

(DISTR.NORM.N(10,17.90,7.02,1) – DISTR.NORM.N(5,17.90,7.02,1))*38 = 3.695.

Los grados de libertad para la χ^2 se calculan como: $m - k - 1 = 7 - 2 - 1 = 4$.

	Intervalo	FrecObs	FrecEsp	Ji-cuadrada
1	hasta 5	1	1.260	0.054
2	5 a 10	3	3.695	0.131
3	10a 15	9	7.960	0.136
4	15 a 20	11	10.550	0.019
5	20 a 25	6	8.604	0.788
6	25 a 30	6	4.316	0.657
7	30 a 35	2	1.331	0.336
	SUMAS	**38**		**2.120**
	Grad_libertad		Ji-cuad 95%	
	4		**9.488**	

Con un nivel de significación α= 5%, el valor crítico $\chi^2_{4,0.95}$ es 9.488, que es mayor que el valor de la prueba: 2.12, por lo que se acepta la hipótesis sobre la distribución de los datos de peso bruto vehicular.

La distribución teórica para los datos se muestra en la figura 4.2, indicando el intervalo del 95% de confianza, con tonelajes entre 4.141 y 31.66 toneladas.

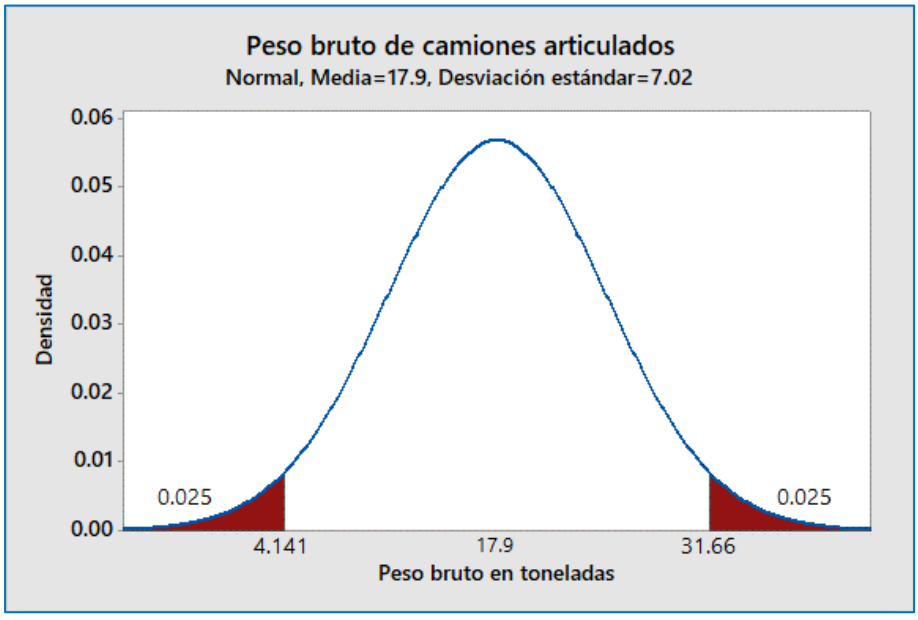
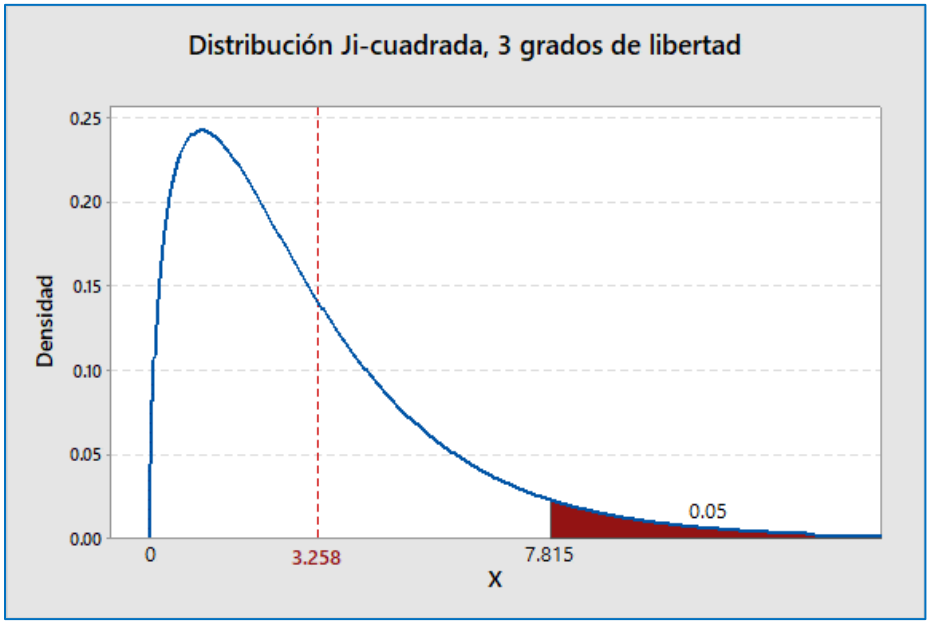


Figura 4.2. Ajuste Normal; muestra de pesos brutos de camiones articulados. (elaboración propia)

4.2.1 El valor p (p-value) en la prueba Ji-cuadrada

En los ejemplos de bondad de ajuste Ji-cuadrada, el procedimiento general calcula en el estadístico de prueba “C” y lo compara con el valor crítico para la prueba; cuando “C” es mayor al valor crítico, se rechaza la hipótesis de que los datos siguen la distribución propuesta, y en caso contrario no se rechaza.

El ejemplo 4.1 usa la distribución Ji-cuadrada con 3 grados de libertad, con un valor crítico de 7.815 para el nivel de significación α = 5% de la prueba. La gráfica siguiente muestra la distribución Ji-cuadrada con 3 g.l. con “C”= 3.258 y el valor crítico = 7.815, que deja a su derecha un área bajo la curva de 5%.



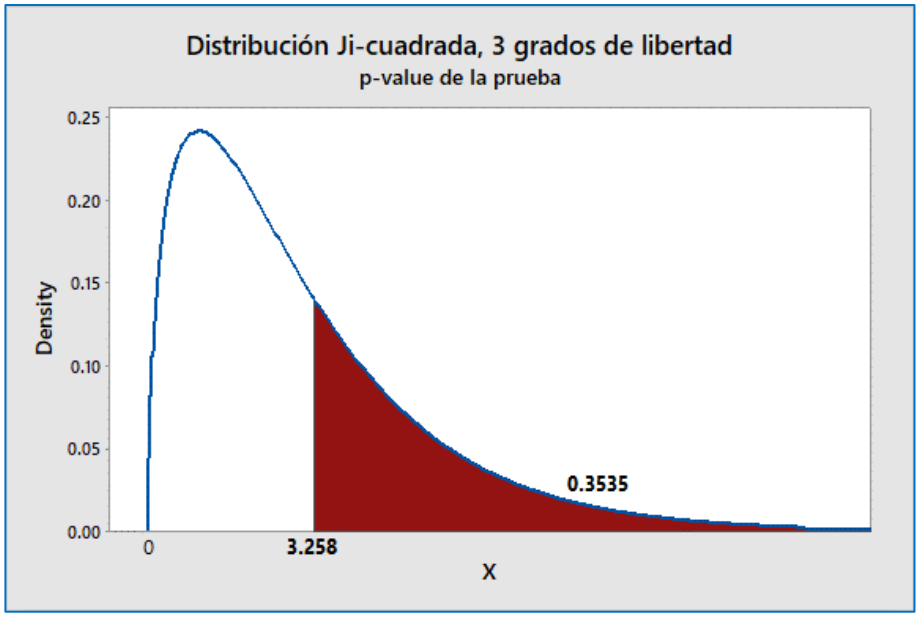
La aceptación de la hipótesis de que los datos siguen una distribución Poisson es clara pues el valor “C” está a la izquierda del valor crítico.

Una forma de cuantificar qué tanto se aleja “C” del crítico es estimar el área bajo la curva que deja a la derecha “C”.

Esta área es el llamado *valor-p* (*p-value* en la literatura inglesa) y da directamente una medida de la calidad del ajuste. Cuando el *p-value* de la prueba supera al 5%, se acepta el ajuste a la distribución propuesta. El valor-p para el ejemplo 4.1 resulta igual 0.3535, como se muestra en la gráfica que sigue.

Este valor-p para el ejemplo 4.1 también puede obtenerse con la siguiente instrucción de Excel, que calcula el área bajo la Ji-cuadrada a la derecha del valor dado (3.258, en este caso):

DISTR.CHICUAD.CD(H15,3) = 0.353531



4.3 Bondad de ajuste Kolmogorov-Smirnov

En esta técnica de ajuste se compara la Función de Distribución Acumulada (FDA) empírica obtenida de los datos, con la FDA teórica de acuerdo a la distribución de probabilidad propuesta.

Una limitación de esta prueba es que sólo debe usarse para distribuciones continuas. Se recomienda tener al menos N= 30 de tamaño de muestra. Con muestras más pequeñas, la prueba puede funcionar, pero si hay pocos datos, los resultados son menos confiables.

Partiendo de una muestra de N datos, y con una distribución propuesta como modelo para los datos, la prueba se desarrolla con los siguientes pasos.

1) Se colocan los *N* datos *en orden creciente* en una tabla: $X_1, X_2, \dots, X_N$ y se estiman los parámetros de la distribución propuesta para los datos.

2) Se calcula la PROBABILIDAD ACUMULADA OBSERVADA PAO_j de cada dato X_j

3) Con la distribución teórica propuesta se calcula la PROBABILIDAD ACUMULADA ESPERADA PAE_j para cada dato X_j.

4) Se calcula el valor absoluto $\left|PAO_j - PAE_j\right|$ para cada dato X_j y se busca la máxima diferencia hallada $MD = Max\left\{\left|PAO_j - PAE_j\right|, j = 1, 2, \dots, N\right\}$ en la tabla de datos.

5) El estimador MD se compara con el valor crítico *de la Tabla K-S* con *N* datos y un nivel de significación del α%.
   Si MD es menor o igual que este valor crítico, entonces no se puede rechazar la hipótesis de que los datos corresponden a la distribución propuesta.

Los valores críticos para varios niveles de significación de la prueba Kolmogorov-Smirnov se muestra en la Tabla 4.2 que se muestra enseguida:

Esta tabla suele encontrarse en textos de estadística o en sitios de temas estadísticos en Internet, como es el caso de *Real Statistics Using Excel* en: http://www.real-statistics.com/statistics-tables/kolmogorov-smirnov-table/.

Tabla 4.2. Valores críticos. Prueba Kolmogorov-Smirnov. (Frías, M.P., 2018)

			NIVEL DE SIGNIFICACIÓN α					
N	**20.0%**	**10.0%**	**5.0%**	**2.0%**	**1.0%**	**0.5%**	**0.2%**	**0.1%**
1	0.90000	0.95000	0.97500	0.99000	0.99500	0.99750	0.99900	0.99950
2	0.68337	0.77639	0.84189	0.90000	0.92929	0.95000	0.96838	0.97764
3	0.56481	0.63604	0.70760	0.78456	0.82900	0.86428	0.90000	0.92065
4	0.49265	0.56522	0.62394	0.68887	0.73424	0.77639	0.82217	0.85047
5	0.44698	0.50945	0.56328	0.62718	0.66853	0.70543	0.75000	0.78137
6	0.41037	0.46799	0.51926	0.57741	0.61661	0.65287	0.69571	0.72479
7	0.38148	0.43607	0.48342	0.53844	0.57581	0.60975	0.65071	0.67930
8	0.35831	0.40962	0.45427	0.50654	0.54179	0.57429	0.61368	0.64098
9	0.33910	0.38746	0.43001	0.47960	0.51332	0.54443	0.58210	0.60846
10	0.32260	0.36866	0.40925	0.45562	0.48893	0.51872	0.55500	0.58042
11	0.30829	0.35242	0.39122	0.43670	0.46770	0.49539	0.53135	0.55588
12	0.29577	0.33815	0.37543	0.41918	0.44905	0.47672	0.51047	0.53422
13	0.28470	0.32549	0.36143	0.40362	0.43247	0.45921	0.49189	0.51490
14	0.27481	0.31417	0.34890	0.38970	0.41762	0.44352	0.47520	0.49753
15	0.26589	0.30397	0.33750	0.37713	0.40420	0.42934	0.45611	0.48182
16	0.25778	0.29472	0.32733	0.36571	0.39201	0.41644	0.44637	0.46750
17	0.25039	0.28627	0.31796	0.35528	0.38086	0.40464	0.43380	0.45540
18	0.24360	0.27851	0.30936	0.34569	0.37062	0.39380	0.42224	0.44234
19	0.23735	0.27136	0.30143	0.33685	0.36117	0.38379	0.41156	0.43119
20	0.23156	0.26473	0.29408	0.32866	0.35241	0.37451	0.40165	0.42085
21	0.22517	0.25858	0.28724	0.32104	0.34426	0.36588	0.39243	0.41122
22	0.22115	0.25283	0.28087	0.31394	0.33666	0.35782	0.38382	0.40223
23	0.21646	0.24746	0.27491	0.30728	0.32954	0.35027	0.37575	0.39380
24	0.21205	0.24242	0.26931	0.30104	0.32286	0.34318	0.36787	0.38588
25	0.20790	0.23768	0.26404	0.29518	0.31657	0.33651	0.36104	0.37743
26	0.20399	0.23320	0.25908	0.28962	0.30963	0.33022	0.35431	0.37139
27	0.20030	0.22898	0.25438	0.28438	0.30502	0.32425	0.34794	0.36473
28	0.19680	0.22497	0.24993	0.27942	0.29971	0.31862	0.34190	0.35842
29	0.19348	0.22117	0.24571	0.27471	0.29466	0.31327	0.33617	0.35242
30	0.19032	0.21756	0.24170	0.27023	0.28986	0.30818	0.33072	0.34672
31	0.18732	0.21412	0.23788	0.26596	0.28529	0.30333	0.32553	0.34129
32	0.18445	0.21085	0.23424	0.26189	0.28094	0.29870	0.32058	0.33611
33	0.18171	0.20771	0.23076	0.25801	0.27577	0.29428	0.31584	0.33115
34	0.17909	0.21472	0.22743	0.25429	0.27271	0.29005	0.31131	0.32641
35	0.17659	0.20185	0.22425	0.25073	0.26897	0.28600	0.30597	0.32187
36	0.17418	0.19910	0.22119	0.24732	0.26532	0.28211	0.30281	0.31751
37	0.17188	0.19646	0.21826	0.24404	0.26180	0.27838	0.29882	0.31333
38	0.16966	0.19392	0.21544	0.24089	0.25843	0.27483	0.29498	0.30931
39	0.16753	0.19148	0.21273	0.23785	0.25518	0.27135	0.29125	0.30544
40	0.16547	0.18913	0.21012	0.23494	0.25205	0.26803	0.28772	0.30171
41	0.16349	0.18687	0.20760	0.23213	0.24904	0.26482	0.28429	0.29811
42	0.16158	0.18468	0.20517	0.22941	0.24613	0.26173	0.28097	0.29465
43	0.15974	0.18257	0.20283	0.22679	0.24332	0.25875	0.27778	0.29130
44	0.15795	0.18051	0.20056	0.22426	0.24060	0.25587	0.27468	0.28806
45	0.15623	0.17856	0.19837	0.22181	0.23798	0.25308	0.27169	0.28493
46	0.15457	0.17665	0.19625	0.21944	0.23544	0.25038	0.26880	0.28190
47	0.15295	0.17481	0.19420	0.21715	0.23298	0.24776	0.26600	0.27896
48	0.15139	0.17301	0.19221	0.21493	0.23059	0.24523	0.26328	0.27611
49	0.14987	0.17128	0.19028	0.21281	0.22832	0.24281	0.26069	0.27339
50	0.14840	0.16959	0.18841	0.21068	0.22604	0.24039	0.25809	0.27067
$n > 50$	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$	$\frac{1.73}{\sqrt{n}}$	$\frac{1.85}{\sqrt{n}}$	$\frac{1.95}{\sqrt{n}}$

Ejemplo 4.3. (variable continua) La muestra siguiente es de estimaciones de vehículos-kilómetro de camiones de carga, con datos de la estación de encuesta origen-destino PC Durango en 2014. El análisis de estas encuestas se desarrolla en el Estudio Estadístico de Campo del Autotransporte Nacional (EECAN) que realiza anualmente el IMT; disponible en:

https://www.imt.mx/micrositios/seguridad-y-operacion-del-transporte/estadisticas/consulta-del-eecan.html

	Muestra de veh-km. Estación PC Durango 2014			
8700	17784	21465	33580	17472
91166	25192	14150	17244	715
7700	12792	19968	24264	4316
34304	60444	40590	24264	10296
71548	66420	11886	11925	7735

El histograma de los datos sugiere una densidad con sesgo a la derecha, como se ve en la figura 4.3, lo que sugiere proponer una Gamma (α, β).

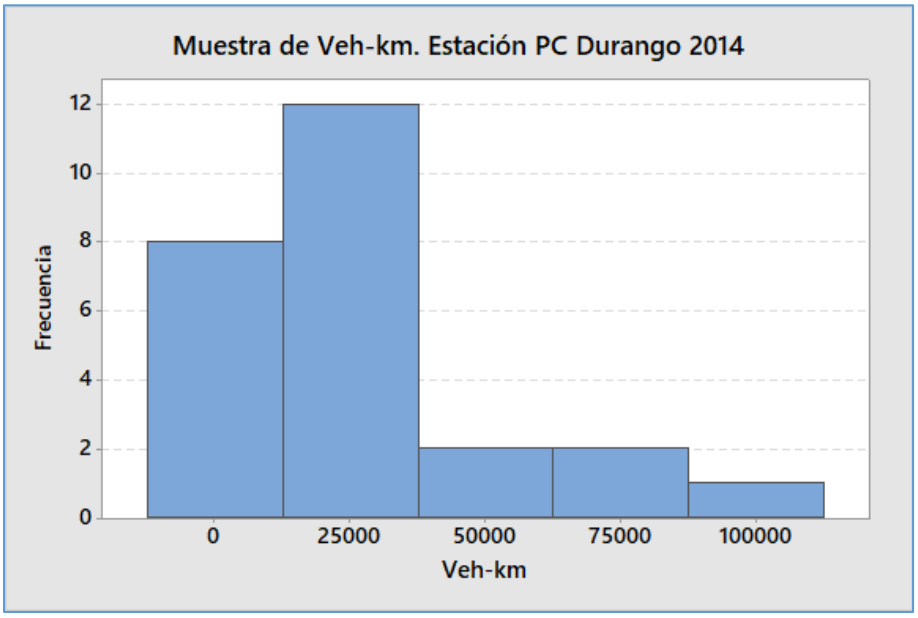


Figura 4.3. Histograma de veh-km. Estación PC Durango (elaboración propia)

Los momentos muestrales calculados a partir de los datos son:

$$m_1' = 26{,}236.8; \quad m_2' = 1{,}200{,}206{,}404.0$$

La correspondiente estimación de parámetros es:

$$\alpha = \frac{{m_1'}^2}{m_2' - {m_1'}^2} = \frac{26{,}236.8^2}{1{,}200{,}206{,}404.0 - 26{,}236.8^2} \approx 1.345$$

$$\beta = \frac{m_2' - {m_1'}^2}{m_1'} = \frac{1{,}200{,}206{,}404.0 - 26{,}236.8^2}{26{,}236.8} \approx 19{,}508.352$$

La distribución propuesta resulta Gamma con parámetro de forma $\alpha \approx 1.345$ y parámetro de escala $\beta = \approx 19{,}508.352$.

La tabla de los tonelajes en orden creciente, junto con sus valores PAO, PAE y valores absolutos de las diferencias es como sigue:

VehKm en orden	Frec.	Pr_Obs	PAO	PAE	Dif_ABS
715	1	0.0400	0.0400	0.0096	0.0304
4316	1	0.0400	0.0800	0.0968	0.0168
7700	1	0.0400	0.1200	0.1916	0.0716
7735	1	0.0400	0.1600	0.1925	0.0325
8700	1	0.0400	0.2000	0.2196	0.0196
10296	1	0.0400	0.2400	0.2635	0.0235
11886	1	0.0400	0.2800	0.3061	0.0261
11925	1	0.0400	0.3200	0.3071	0.0129
12792	1	0.0400	0.3600	0.3297	0.0303
14150	1	0.0400	0.4000	0.3642	0.0358
17244	1	0.0400	0.4400	0.4380	0.0020
17472	1	0.0400	0.4800	0.4431	0.0369
17784	1	0.0400	0.5200	0.4502	0.0698
19968	1	0.0400	0.5600	0.4973	0.0627
21465	1	0.0400	0.6000	0.5277	0.0723
24264	2	0.0800	0.6800	0.5804	0.0996
25192	1	0.0400	0.7200	0.5967	0.1233
33580	1	0.0400	0.7600	0.7202	0.0398
34304	1	0.0400	0.8000	0.7291	0.0709
40590	1	0.0400	0.8400	0.7956	0.0444
60444	1	0.0400	0.8800	0.9182	0.0382
66420	1	0.0400	0.9200	0.9382	0.0182
71548	1	0.0400	0.9600	0.9515	0.0085
91166	1	0.0400	1.0000	0.9810	0.0190

El cálculo de la PAE usa la función de Excel DISTR.GAMMA.N (x, Alfa, Beta, 1); por ejemplo, para el tonelaje x = 25192, el valor de PAE calculado en Excel resulta:

= DISTR.GAMMA.N(25192,1.345,19508.32,1) = 0.596634.

El estimador MD es el máximo valor de las diferencias absolutas entre distribución acumulada empírica (PAO) y teórica (PAE), que resulta igual a 0.1233, y aparece marcado con amarillo en la tabla para 25192 veh-km.

En la tabla para la prueba K-S (Tabla 4.2) con n = 25 datos, y nivel de significación α = 5%, el valor crítico es 0.26404, mayor que el estimador MD, por lo que se acepta el ajuste de la distribución Gamma propuesta para los datos.

4.4 Bondad de ajuste Anderson-Darling

Esta prueba es una versión mejorada de la prueba K-S, ya que detecta mejor las diferencias entre las distribuciones acumuladas empírica y teórica en las colas de las gráficas. Se recomienda usar solamente para distribuciones continuas. Los pasos de la prueba son:

1) Con la muestra de *N* datos se estiman los parámetros de la distribución propuesta, y se realizan los dos ordenamientos de datos siguientes.

2) Los datos de *menor a mayor*: $Y_j, j = 1, 2, \dots, N;\ \ Y_k \ \leq \ \ Y_{k+1}$

3) Los datos de *mayor a menor* $Y_{N-j+1}, j = 1, 2, \dots, N;\ \ Y_k \ \geq \ \ Y_{k+1}$.

6) Se calcula la PROBABILIDAD ACUMULADA ESPERADA PAE_j para cada dato Y_j.y la PROBABILIDAD ACUMULADA ESPERADA PAE_{N+1-j} para cada dato Y_{N+1-j}.para cada para cada valor Y_{n+j-1} con base en la distribución de probabilidad teórica propuesta para la prueba.

4) Se calcula el estadístico de la prueba como:

$$A_N = -\left[N + \frac{1}{N}\sum_{j=1}^{N}(2j-1)\left[LnPAE_j + Ln\left(1 - \ PAE_{N-j+1}\right)\right]\right]$$

5) Se ajusta el estadístico de prueba conforme a la distribución de probabilidad teórica propuesta, según la tabla prueba Anderson-Darling, y se determina el valor crítico de la prueba.

6) Si el estadístico de prueba es menor que el valor crítico del punto anterior, no se puede rechazar la hipótesis nula; en caso contrario, se rechaza.

La tabla de valores críticos y ajustes de la prueba A-D es como sigue:

Tabla 4.3. Valores críticos. Prueba Anderson-Darling. (Stephens, 1979; Real Statistics, 2018).

Distribución	Estadístico ajustado	α = 0.10	α = 0.05	α = 0.025	α = 0.01
Parámetros conocidos, N ≥ 5	A_N^2	1.933	2.492	3.070	3.853
Parámetros estimados a partir de los datos					
Normal y Lognormal	$A_N^2\left(1+\frac{3}{4N}+\frac{9}{4N^2}\right)$	0.631	0.752	0.873	1.035
Exponencial	$A_N^2\left(1+\frac{3}{10N}\right)$	1.062	1.321	1.591	1.959
Weibull	$A_n^2\left(1+\frac{1}{5\sqrt{n}}\right)$	0.637	0.757	0.877	1.038

Ejemplo 4.4. Los 50 datos siguientes son pesos medios de embarques ferroviarios de automotores nuevos, en toneladas/carro colectados en 2016.

Embarques ferroviarios (t/carro). Automotores nuevos				
17.34	24.02	25.62	36.53	25.62
19.30	24.86	25.62	39.68	25.88
20.00	25.62	26.51	20.00	15.15
20.00	14.47	13.50	20.00	21.79
20.00	33.35	13.60	20.00	18.22
21.79	11.19	13.19	11.62	29.76
21.79	34.20	16.49	22.05	31.97
21.79	17.36	33.00	24.11	13.08
22.05	39.68	17.44	12.30	20.00
22.05	20.00	14.12	20.00	20.00

El histograma de datos de la figura 4.4, indica una distribución con sesgo hacia la derecha, lo que puede sugerir una distribución Lognormal.

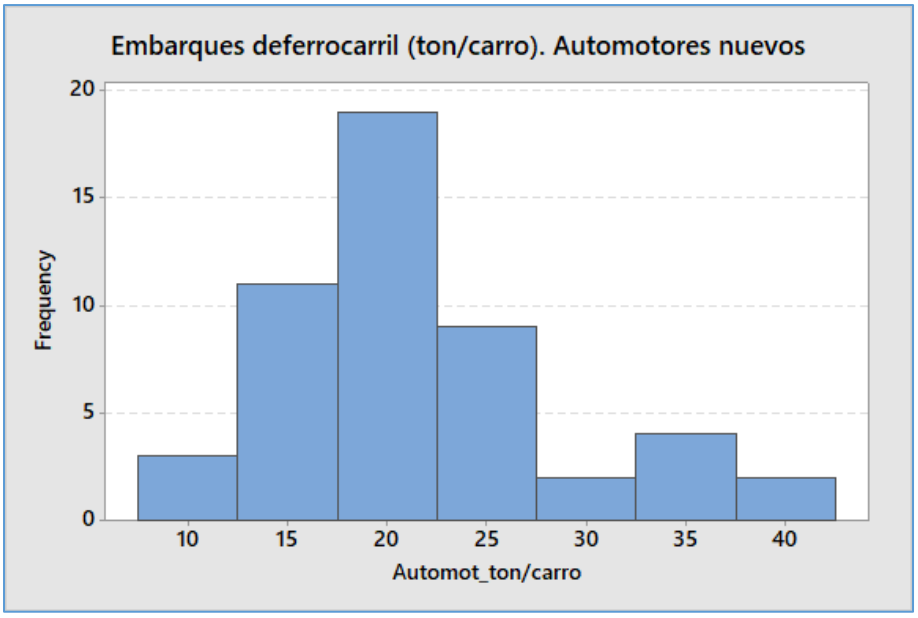


Figura 4.4. Histograma de ton/carro. Embarques ferroviarios de automotores. (elaboración propia)

Los momentos muestrales se calculan con los datos X_i y los valores $Y_i = LnX_i$ como indica la tabla 4.1:

$$m'_{1Y} \approx 3.0401 \; ; \; m'_{2Y} = 9.3398$$

Y la correspondiente estimación de parámetros para una Lognormal resulta:

$$\mu = E(Y) = m'_{1Y} \approx 3.0401$$

$$\sigma = \sqrt[2]{E(Y^2) - E(Y)^2} = \sqrt[2]{m'_{2Y} - {m'_{1Y}}^2} = \sqrt[2]{9.3398 - 3.0401^2} \approx 0.3124$$

La tabla con los tonelajes X_j, los tonelajes $Y_j$en orden creciente, los tonelajes Y_{50-j+1} en orden decreciente, los valores PAE_J, $1 - PAE_{50-J+1}$ y el cálculo para la prueba Anderson-Darling se muestra en la tabla siguiente, desarrollada en Excel.

					A	B	C
N	t/carro	Y_j	Y_{50-j+1}	$2j-1$	PAE_J	$1-PAE_{50-j+1}$	$(2j-1)[LnA+LnB]$
1	17.34	11.19	39.68	1	0.023	0.020	-7.691
2	19.30	11.62	39.68	3	0.030	0.020	-22.229
3	20.00	12.30	36.53	5	0.045	0.037	-32.019
4	20.00	13.08	34.20	7	0.067	0.058	-38.938
5	20.00	13.19	33.35	9	0.070	0.068	-48.138
6	21.79	13.50	33.00	11	0.081	0.072	-56.620
7	21.79	13.60	31.97	13	0.084	0.087	-63.879
8	21.79	14.12	29.76	15	0.105	0.129	-64.570
9	22.05	14.47	26.51	17	0.119	0.224	-61.601
10	22.05	15.15	25.88	19	0.151	0.247	-62.433
11	24.02	16.49	25.62	21	0.224	0.258	-59.896
12	24.86	17.34	25.62	23	0.275	0.258	-60.908
13	25.62	17.36	25.62	25	0.276	0.258	-66.079
14	14.47	17.44	25.62	27	0.281	0.258	-70.909
15	33.35	18.22	24.86	29	0.330	0.290	-68.090
16	11.19	19.30	24.11	31	0.399	0.324	-63.439
17	34.20	20.00	24.02	33	0.444	0.328	-63.576
18	17.36	20.00	22.05	35	0.444	0.433	-57.781
19	39.68	20.00	22.05	37	0.444	0.433	-61.083
20	20.00	20.00	22.05	39	0.444	0.433	-64.384
21	25.62	20.00	21.79	41	0.444	0.447	-66.325
22	25.62	20.00	21.79	43	0.444	0.447	-69.561
23	26.51	20.00	21.79	45	0.444	0.447	-72.796
24	13.50	20.00	21.79	47	0.444	0.447	-76.032
25	13.60	20.00	20.00	49	0.444	0.556	-68.557
26	13.19	20.00	20.00	51	0.444	0.556	-71.356
27	16.49	21.79	20.00	53	0.553	0.556	-62.484
28	33.00	21.79	20.00	55	0.553	0.556	-64.842
29	17.44	21.79	20.00	57	0.553	0.556	-67.199
30	14.12	21.79	20.00	59	0.553	0.556	-69.557
31	36.53	22.05	20.00	61	0.567	0.556	-70.325
32	39.68	22.05	20.00	63	0.567	0.556	-72.631
33	20.00	22.05	20.00	65	0.567	0.556	-74.937
34	20.00	24.02	20.00	67	0.672	0.556	-65.943
35	20.00	24.11	19.30	69	0.676	0.601	-62.123
36	11.62	24.86	18.22	71	0.710	0.670	-52.709
37	22.05	25.62	17.44	73	0.742	0.719	-45.810
38	24.11	25.62	17.36	75	0.742	0.724	-46.576
39	12.30	25.62	17.34	77	0.742	0.725	-47.672
40	20.00	25.62	16.49	79	0.742	0.776	-43.575
41	25.62	25.88	15.15	81	0.753	0.849	-36.296
42	25.88	26.51	14.47	83	0.776	0.881	-31.558
43	15.15	29.76	14.12	85	0.871	0.895	-21.143
44	21.79	31.97	13.60	87	0.913	0.916	-15.593
45	18.22	33.00	13.50	89	0.928	0.919	-14.146
46	29.76	33.35	13.19	91	0.932	0.930	-13.006
47	31.97	34.20	13.08	93	0.942	0.933	-11.932
48	13.08	36.53	12.30	95	0.963	0.955	-7.929
49	20.00	39.68	11.62	97	0.980	0.970	-4.934
50	20.00	39.68	11.19	99	0.980	0.977	-4.286

El correspondiente estimador de la prueba resulta:

$$A^2_{50} = -\left[50 + \frac{1}{50}\sum \text{Columna "C"}\right] = 0.5219$$

El valor crítico de la prueba Anderson-Darling con un nivel de significación α = 5% para *n = 50* datos se obtiene de la tabla 4.3 como sigue.

Puesto que fue necesario estimar los parámetros μ y σ de la Lognormal, se aplica el factor de corrección para el estimador:

$$A^* = A^2_{50}\left(1 + \frac{3}{4\times 50} + \frac{9}{4\times 50^2}\right) \approx 0.53019$$

El valor crítico de la tabla 4.4 para un nivel de significación $\alpha = 0.05$ es 0.752, mayor al valor A^* obtenido en la prueba, por lo que se acepta el ajuste de los datos a una Lognormal de parámetros: $\mu = 3.0401$; $\sigma = 0.3124$. La figura 4.5 muestra el histograma con la curva teórica ajustada.

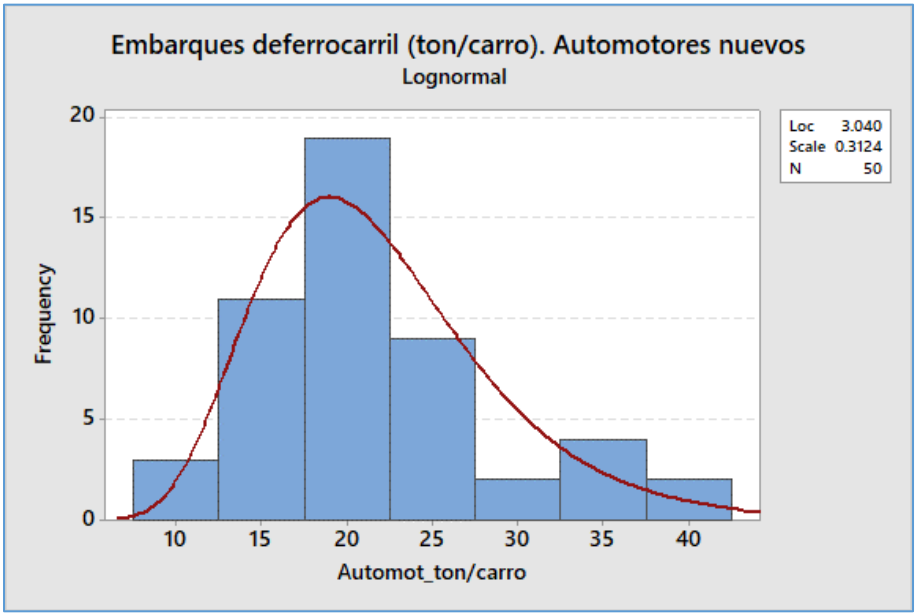


Figura 4.5. Ajuste Lognormal. Embarques ferroviarios de automotores. (elaboración propia)

5 Conclusiones

La presencia de datos de tipo aleatorio en proyectos de transporte e ingeniería plantea el problema de la caracterización probabilística adecuada de estos datos.

La caracterización adecuada de los datos tiene relevancia ya que:

a) Permite calcular probabilidades de los eventos que interesan al problema de donde vienen los datos, con un buen nivel de confianza.

b) Permite calcular valores esperados de variables de interés, así como, dimensionar variables involucradas que garanticen cierto nivel de probabilidad para una situación deseada (como en el ejemplo 2.5 de camiones con espera acotada).

c) Proporciona un marco estadístico adecuado y defendible para propósitos de publicación de artículos sometidos a revistas científicas o técnicas de prestigio.

d) En las aplicaciones de simulación, tener una representación probabilista confiable de los datos permite tener resultados más cercanos a la realidad.

Un ejemplo de este enfoque en Ingeniería de Tránsito, lo refieren Gerlow and Huber (1974):

> "*En el diseño de nueva infraestructura vial o nuevos planes de control es necesario pronosticar el desempeño del tráfico respecto a alguna característica particular, y frecuentemente es deseable ser capaces de hacer una predicción con una cantidad limitada de información disponible o supuesta. Por ejemplo, al diseñar un sistema de control para peatones, podría requerirse predecir la frecuencia de tiempos entre llegadas mayores a diez segundos; al diseñar la operación de una vuelta a la izquierda podría requerirse pronosticar cuántas veces por hora el número de autos que llegan durante un ciclo el semáforo superarán a cuatro vehículos. Los modelos de distribuciones estadísticas pueden habilitar al ingeniero de tráfico para hacer predicciones como las señaladas, contando con una cantidad mínima de información.*

La caracterización probabilista de los datos suele obtenerse partiendo de un histograma que muestra su comportamiento frecuencial para sugerir alguna forma de distribución conocida, o partiendo de información del uso reportado en la literatura de ciertas distribuciones para ciertas aplicaciones (p. ej. en confiabilidad el uso de la Weibull y la Exponencial; en problemas de aforos o conteos, el uso de Poisson, Binomial Negativa, etc.).

Para que la caracterización probabilista de los datos esté completa, y con sustento estadístico adecuado, se requieren dos pasos básicos: a) estimar los parámetros concretos que identifican a la distribución para los datos disponibles; b) hacer una prueba de bondad de ajuste que determine la calidad estadística de la representación con un nivel de confianza (usualmente el 95%) o equivalente a un nivel de significación (usualmente el 5%).

5.1 Uso de software estadístico

El uso de software estadístico para realizar el ajuste de distribuciones es una práctica común en la solución de problemas de transporte e ingeniería. La ventaja que se tiene con este tipo de programas es que se ahorra mucho tiempo de cálculo y se pueden obtener rápidamente buenas representaciones gráficas y reportes adecuados para mostrar un ajuste adecuado. Sin embargo, para aprovechar estas ventajas de los paquetes estadísticos es conveniente revisar y familiarizarse con la información que resulta a la salida de sus procesos.

En esta sección se reproducen algunos de los ejemplos del capítulo 4, resueltos con tres paquetes de uso común en la práctica: Minitab, JMP y STATISTICA.

Ajuste con Minitab

Con los datos del ejemplo 4.1, el resultado de ajuste con la prueba Ji-cuadrada con el paquete estadístico Minitab se muestra enseguida.

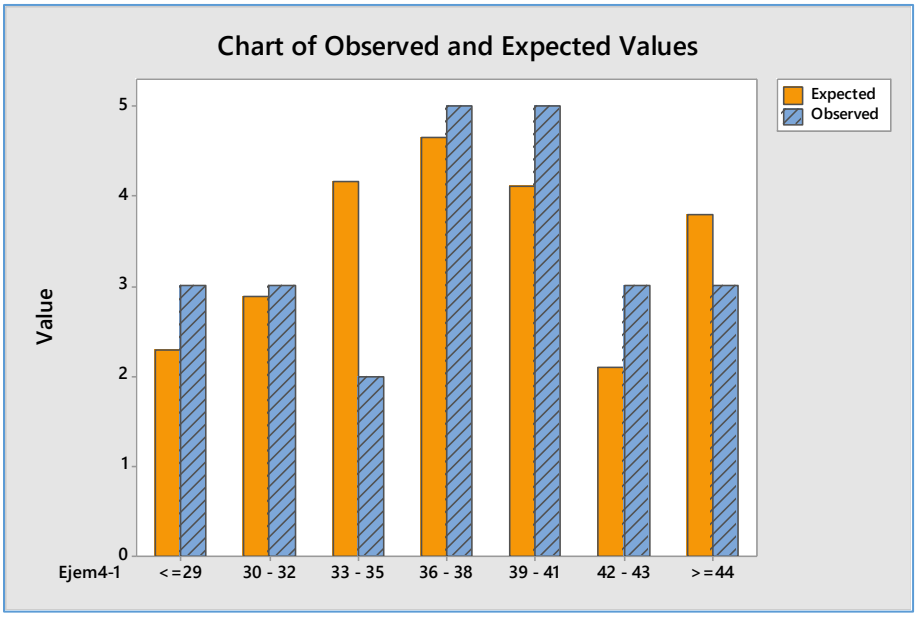
Goodness-of-Fit Test for Poisson Distribution

```
Data column: Ejem4-1

Poisson mean for Ejem4-1 = 37.375

                          Poisson              Contribution
Ejem4-1  Observed  Probability  Expected     to Chi-Sq
<=29            3     0.095193   2.28463       0.22400
30 - 32         3     0.120222   2.88533       0.00456
33 - 35         2     0.173660   4.16784       1.12757
36 - 38         5     0.194215   4.66115       0.02463
39 - 41         5     0.171525   4.11661       0.18957
42 - 43         3     0.087222   2.09334       0.39269
>=44            3     0.157962   3.79110       0.16508

 N  N*  DF   Chi-Sq  P-Value
24   0   5  2.12810    0.831
```

Lo que se lee en este reporte es primeramente, la estimación de la media de la distribución Poisson = 37.375; enseguida se muestra la tabla de intervalos considerados con sus valores de frecuencia observada, probabilidad Poisson, frecuencia esperada y la contribución a la suma del estadístico C (Ji-cuadrada).

Se reportan los grados de libertad (Degree of Freedom) como DF = 5, y se muestra el estadístico C = 2.12810 con el valor-p de la prueba: 0.831, por lo que se acepta la hipótesis Poisson.

Minitab también ofrece la siguiente gráfica que muestra las diferencias entre valores observados y valores esperados en el conjunto de datos:



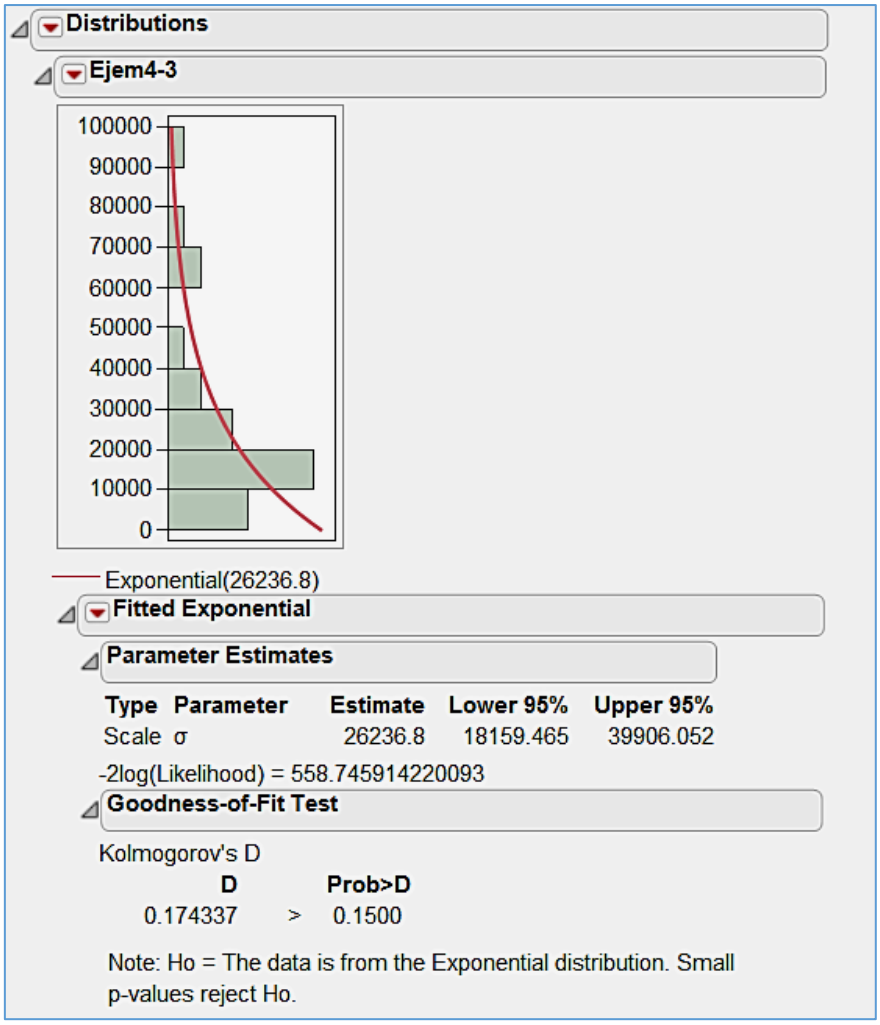
Ajuste con JMP

Con los datos del ejemplo 4.3, se hace un ajuste a una distribución Exponencial con el software JMP 9.0. El reporte se muestra enseguida.

El parámetro estimado para la exponencial es la media = 26236.8; también da un intervalo de confianza del 95% para dicha estimación: [18159.465, 39906.052].

Al final de la sección “Parameter Estimates” indica el valor del logaritmo de la función de verosimilitud (Likelihood) y luego se muestra el estadístico Kolmogorov-Smirnov igual a 0.174337, que tiene asociado un valor-p de 0.1500, con lo cual se aceptaría también la hipótesis de una distribución exponencial como la calculada.

En el histograma que aparece (formato vertical), se muestra el ajuste de la curva exponencial teórica.



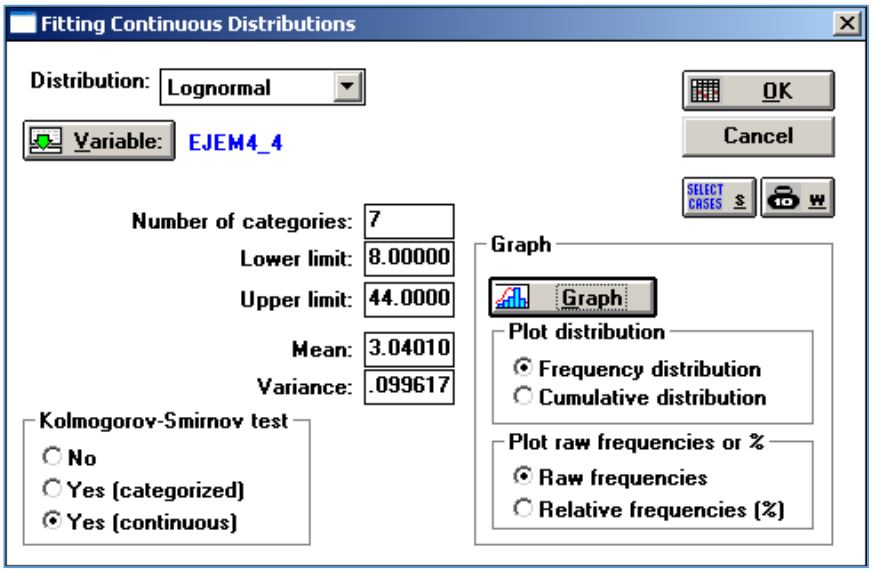
Ajuste con STATISTICA

Con los datos del ejemplo 4.4, se muestra el ajuste a una distribución Lognormal con el software STATISTICA 4.3. La primera parte del reporte resume las características de los datos y da la estimación de media (Mean: 3.04010) y varianza (Variance: 0.099617) para la disribución Lognormal que se ajusta.

Enseguida se muestra una tabla de intervalos para los datos, indicando frecuencia observada y su acumulada; porcentaje observado y su acumulado; frecuencia esperada y su acumulada; porcentaje esperado y su acumulado y finalmente las diferencias de observado contra esperado.

Se da la prueba de Kolmogorov-Smirnov, con el estadístico D = 0.1240919 indicando en su valor-p la leyenda “n.s.” (no significativa *-not-significant-*) lo que quiere decir que el valor-p no es menor al 5%. También se muestra la prueba Ji-cuadrada para 1 grado de libertad, con el estadístico de prueba C = 0.7615446 y un

valor-p asociado de 0.3828521, mayor al 5% de significación de la prueba, por lo que no se rechaza la hipótesis de datos Lognormales.

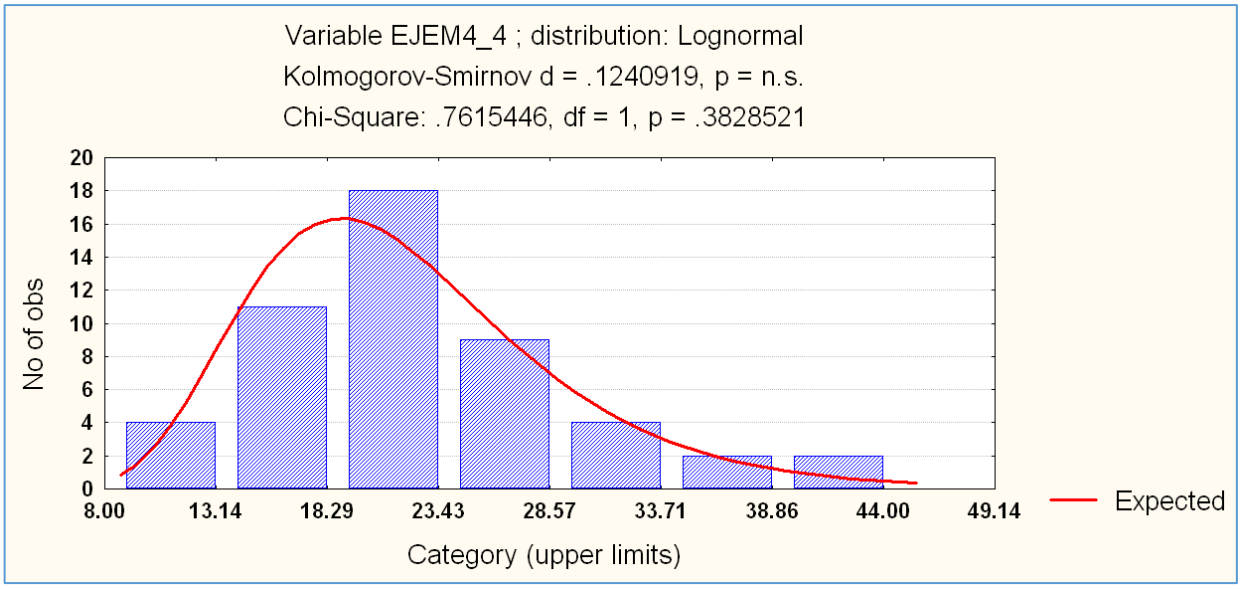


STATISTICA: Nonparametric Statistics & Distribution Fitting - [Variable EJEM4_4 ; distribution: Lognormal (new.sta)]

File Edit View Analysis Graphs Options Macro Window Help

NONPAR STATS — Kolmogorov-Smirnov d = .1240919, p = n.s.
Chi-Square: .7615446, df = 1, p = .3828521

Upper Boundary	observed freq-cy	cumulatv observed	percent observed	cumul. % observed	expected freq-cy	cumulatv expected	percent expected	cumul. % expected	observd-expected
<=13.143	4	4	8.00000	8.0000	3.53330	3.53330	7.06660	7.0666	.46670
18.2857	11	15	22.00000	30.0000	13.24633	16.77963	26.49266	33.5593	-2.24633
23.4286	18	33	36.00000	66.0000	15.26243	32.04205	30.52485	64.0841	2.73757
28.5714	9	42	18.00000	84.0000	9.89706	41.93912	19.79413	83.8782	-.89706
33.7143	4	46	8.00000	92.0000	4.80938	46.74850	9.61876	93.4970	-.80938
38.8571	2	48	4.00000	96.0000	2.01226	48.76076	4.02452	97.5215	-.01226
Infinity	2	50	4.00000	100.0000	1.23924	50.00000	2.47848	100.0000	.76076



5.2 El caso de varios ajustes o de ninguno

En la práctica, dado un conjunto de datos, no es raro que se puedan obtener ajustes aceptables con distintas distribuciones de probabilidad. En otros casos, también puede ocurrir que al probar todas las distribuciones que parecieran plausibles para representar los datos, o al examinar todas las opciones que ofrece un paquete estadístico, en ningún caso se obtenga un ajuste aceptable. Estos dos casos se comentan enseguida.

Varios ajustes posibles

En este caso, se recomienda elegir el modelo probabilista que tenga el mejor valor-p, ya que es el que se adentra mucho más en la región de aceptación de la hipótesis que se está probando.

El ejemplo siguiente muestra el ajuste de Minitab para los datos del ejemplo 4.4, considerando tres posibles ajustes: Lognormal, Exponencial y Normal.

Distribution ID Plot for Ejem4-3

```
Goodness of Fit Test

Distribution     AD       P
Gamma         0.371  >0.250
Exponential   0.574   0.387
Normal        1.671  <0.005


ML Estimates of Distribution Parameters

Distribution     Location    Shape        Scale  Threshold
Gamma                     1.37619  19124.51356
Exponential                        26318.99981
Normal*       26319.00000          23583.21450

* Scale: Adjusted ML estimate
```

El reporte de bondad de ajuste de Minitab muestra las tres distribuciones probadas, junto con su estadístico de prueba Anderson-Darling (indicado por la sigla AD) y su correspondiente valor-p (indicado por *p*), donde el valor más alto 0.387 es para la Exponencial, seguido del valor-p 0.250 para la Gamma; ambos mayores al 5% de significación de las pruebas.

De esto se puede entonces elegir como mejor representación para los datos del ejemplo 4-3, una distribución exponencial.

La distribución Normal aparece con el valor AD 1.671, y un valor-p 0.005, lo que indica que se rechaza la hipótesis de normalidad de los datos.

Finalmente, el reporte indica los parámetros estimados para cada distribución que se ha probado, dando para la Exponencial el parámetro de escala (media) 26318.99981. Al final del reporte se menciona que los parámetros se estimaron con el método de Máxima Verosimilitud (Maximum Likelihood –*ML estimate*–).

Ninguno de los ajustes probados es aceptable

En este caso, pudiera ser que los datos sean insuficientes o que no tengan muy buena calidad; pero si no se tiene más información a la mano, y no se ha encontrado ajuste aceptable para ninguna de las distribuciones que se supusieron factibles para los datos, o para todas las opciones del paquete estadístico que se esté utilizando, aún es posible generar la *distribución empírica* a partir de los datos de que se dispone.

La distribución empírica $F_E(x)$ tiene la misma definición que una FDA teórica de cualquier distribución: $F_E(x) = P[X \leq x]$, solo que esta probabilidad está referida a la muestra de datos, y no a una fórmula analítica como ocurre en las distribuciones discutidas en el capítulo 2. (Statistics How To, 2018).

La función de distribución empírica (FDE) de una muestra con *n* datos se genera, primero poniendo orden creciente los datos, y luego se asignando la aportación porcentual de cada dato al total; así, el primer dato x_1 tiene la probabilidad acumulada de $\frac{1}{n}$, el segundo dato tiene la probabilidad acumulada $\frac{2}{n}$, etc. Cada dato agrega un porcentaje acumulado igual a $\frac{1}{n}$. En caso de que un dato se repita, digamos 3 veces, su aportación agregaría $\frac{3}{n}$ al acumulado, y así sucesivamente.

La tabla 5.1 muestra datos de toneladas/carro en movimientos ferroviarios de aceite de soya.

Tabla 5.1. Embarques ferroviarios de aceite de soya. (elaboración propia)

	Embarques ferroviarios de Aceite de Soya. Tons/carro						
69.540	86.284	84.929	85.812	89.811	73.853	92.996	89.529
70.166	67.606	85.789	89.552	84.630	78.399	87.329	84.031
70.382	78.378	74.160	78.521	84.657	87.151	88.577	85.226
70.382	84.079	74.165	72.463	84.934	88.609	76.299	83.979
83.810	84.380	74.321	72.500	85.785	76.327	83.993	84.725
83.813	85.422	74.440	84.796	76.795	85.513	84.204	85.889
83.836	89.207	78.319	84.981	76.906	74.185	84.625	83.873
83.958	92.217	84.183	85.268	89.278	75.885	84.881	89.691
83.979	70.678	85.074	85.558	83.813	83.918	84.953	86.214
84.092	84.122	85.005	85.876	85.250	84.349	85.238	86.247

El histograma correspondiente se muestra en la figura 5.1.

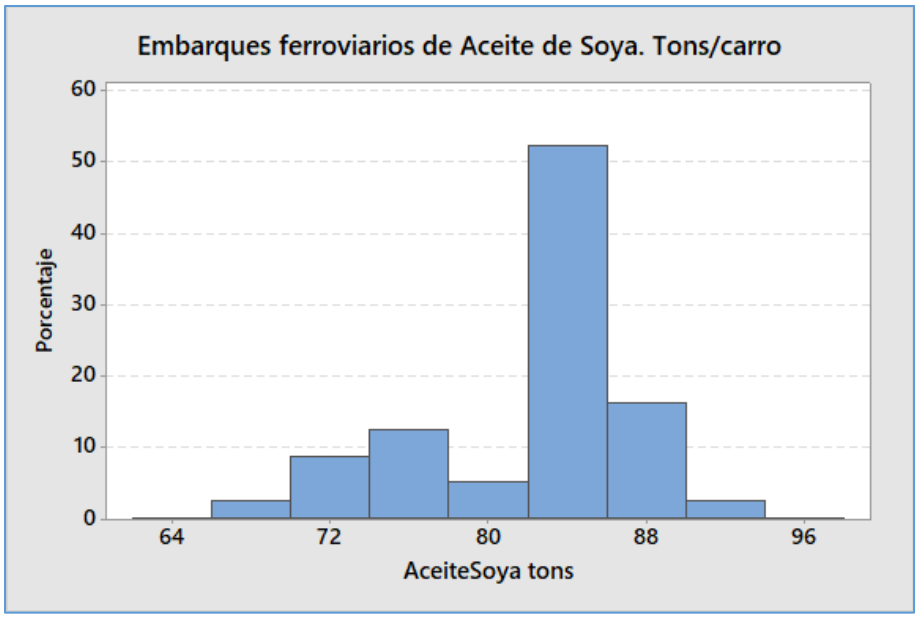


Figura 5.1. Histograma de embarques ferroviarios de aceite de soya. (elaboración propia)

El histograma sugiere una distribución sesgada a la derecha, aunque la barra entre las 80 y las 88 toneladas parece muy alta. Al probar los datos con todas las opciones de ajuste que ofrece Minitab, se obtuvo el siguiente resultado.

Distribution ID Plot for AceiteSoya tons

Goodness of Fit Test

Distribution	AD	P	LRT P
Normal	4.862	<0.005	
Box-Cox Transformation	3.361	<0.005	
Lognormal	5.316	<0.005	
3-Parameter Lognormal	4.897	*	0.023
Exponential	32.074	<0.003	
2-Parameter Exponential	14.307	<0.010	0.000
Weibull	3.284	<0.010	
3-Parameter Weibull	2.972	<0.005	0.372
Smallest Extreme Value	2.971	<0.010	
Largest Extreme Value	6.299	<0.010	
Gamma	5.198	<0.005	
3-Parameter Gamma	5.100	*	0.333
Logistic	4.563	<0.005	
Loglogistic	4.929	<0.005	
3-Parameter Loglogistic	4.563	*	0.038

Las 15 opciones de ajuste de Minitab dan valores-p mucho menores a 5%, por lo que ninguna de ellas es aceptable. El ordenamiento de datos y su porcentaje acumulado, queda como sigue (Minitab).

Tally for Discrete Variables: AceiteSoya tons

AceiteSoya tons	Count	Percent	CumPct
67.606	1	1.25	1.25
69.540	1	1.25	2.50
70.166	1	1.25	3.75
70.382	2	2.50	6.25
70.678	1	1.25	7.50
72.463	1	1.25	8.75
72.500	1	1.25	10.00
73.853	1	1.25	11.25
74.160	1	1.25	12.50
74.165	1	1.25	13.75
74.185	1	1.25	15.00
74.321	1	1.25	16.25
74.440	1	1.25	17.50
75.885	1	1.25	18.75
76.299	1	1.25	20.00
76.327	1	1.25	21.25
76.795	1	1.25	22.50
76.906	1	1.25	23.75
78.319	1	1.25	25.00
78.378	1	1.25	26.25
78.399	1	1.25	27.50
78.521	1	1.25	28.75
83.810	1	1.25	30.00
83.813	2	2.50	32.50
83.836	1	1.25	33.75
83.873	1	1.25	35.00
83.918	1	1.25	36.25
83.958	1	1.25	37.50
83.979	1	1.25	38.75
83.979	1	1.25	40.00
83.993	1	1.25	41.25
84.031	1	1.25	42.50
84.079	1	1.25	43.75
84.092	1	1.25	45.00
84.122	1	1.25	46.25
84.183	1	1.25	47.50
84.204	1	1.25	48.75
84.349	1	1.25	50.00
84.380	1	1.25	51.25
84.625	1	1.25	52.50
84.630	1	1.25	53.75
84.657	1	1.25	55.00
84.725	1	1.25	56.25
84.796	1	1.25	57.50
84.881	1	1.25	58.75
84.929	1	1.25	60.00
84.934	1	1.25	61.25
84.953	1	1.25	62.50

84.981	1	1.25	63.75
85.005	1	1.25	65.00
85.074	1	1.25	66.25
85.226	1	1.25	67.50
85.238	1	1.25	68.75
85.250	1	1.25	70.00
85.268	1	1.25	71.25
85.422	1	1.25	72.50
85.513	1	1.25	73.75
85.558	1	1.25	75.00
85.785	1	1.25	76.25
85.789	1	1.25	77.50
85.812	1	1.25	78.75
85.876	1	1.25	80.00
85.889	1	1.25	81.25
86.214	1	1.25	82.50
86.247	1	1.25	83.75
86.284	1	1.25	85.00
87.151	1	1.25	86.25
87.329	1	1.25	87.50
88.577	1	1.25	88.75
88.609	1	1.25	90.00
89.207	1	1.25	91.25
89.278	1	1.25	92.50
89.529	1	1.25	93.75
89.552	1	1.25	95.00
89.691	1	1.25	96.25
89.811	1	1.25	97.50
92.217	1	1.25	98.75
92.996	1	1.25	100.00
N=	80		

La figura 5.2 muestra la función empírica de distribución acumulada resultante.

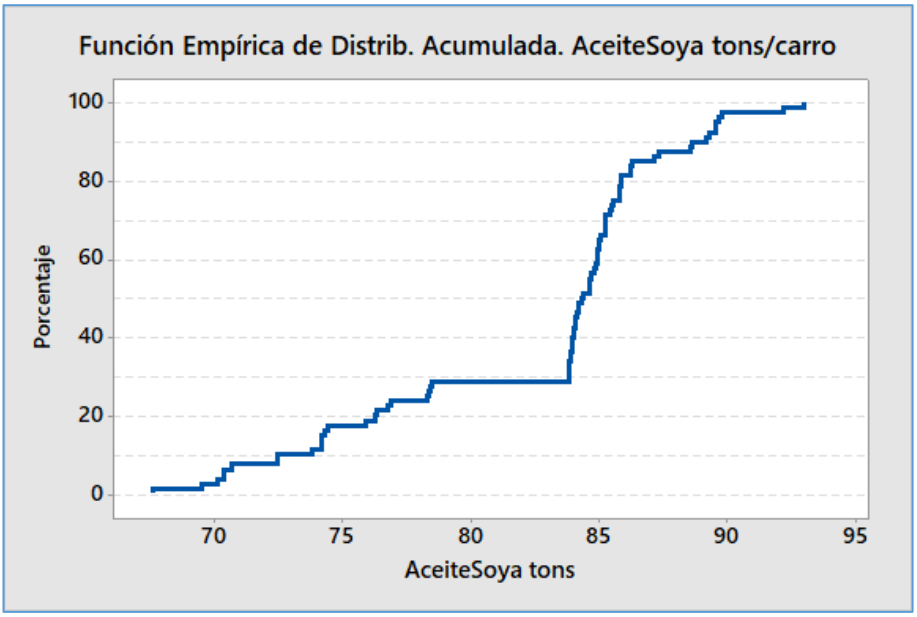


Figura 5.2. Función de Distribución Acumulada Empírica. Aceite de soya t/carro. (elaboración propia)

Esta distribución empírica permite estimar como primera aproximación, probabilidades de eventos de interés para el modelador directamente de la tabla de frecuencias o equivalentemente de la gráfica de la Función Empírica.

Por ejemplo, si se quiere saber cuál es la probabilidad de tener un embarque de aceite de soya que lleve entre 85 y 90 toneladas por carro, se calcula directamente:

$$P[85 \leq Peso \leq 90] = FDE(90) - FDE(85) \cong$$

$$FDE(89.811) - FDE(85.005) \cong 97.50\% - 65.00\% \approx 32.50'\%$$

En el desarrollo de este trabajo, se han mostrado las distribuciones de probabilidad más comunes que aparecen en problemas de transporte e ingeniería, y se ha incluido una breve lista de aplicaciones de estas distribuciones que la literatura ha reportado como exitosas, permitiendo así aprovechar la experiencia de otros investigadores que ya han avanzado en distintos campos de aplicación.

Para tener el marco de referencia completo, también se han resumido en los capítulos previos, el uso de las funciones estadísticas de Excel; los métodos de estimación de parámetros más comunes; y el desarrollo de las tres técnicas de bondad de ajuste usadas en la práctica: Ji-cuadrada, Kolmogorov-Smirnov y Anderson-Darling.

Con estas referencias se puede proceder a hacer los cálculos necesarios para obtener los ajustes de datos que se deseen, y también se puede aprovechar dichas referencias para aprovechar algún software estadístico del que se disponga, y poder interpretar adecuadamente los resultados que se obtengan.

.

Bibliografía

Frías Bustamante, M. P. (2018). *Tablas de inferencia*. Estadística II. Ingeniería en Organización Industrial. Universidad de Jaén. España. En: http://www4.ujaen.es/~mpfrias/TablasInferencia.pdf

Gerlough, D.L. and Huber, M. J. (1975).*Traffic Flow Theory. A Monograph*. TRB Special Report 165. Transportation Research Board. Washington D.C.

Gutiérrez-González, E., et al. (2013). "Aplicación de un modelo de inventario con revisión periódica para la fabricación de transformadores de distribución". *Ingeniería Investigación y Tecnología*. Vol. XIV (número 4), octubre-diciembre 2013. Universidad Nacional Autónoma de México. Facultad de Ingeniería. México.

Krishnamoorty, K. (2006). *Handbook of Statistical Distributions with Applications.* Chapman & Hall/CRC. USA.

Mood, A.M.; Graybill, F.A. and Boes, D.C. (1963). *Introduction to the Theory of Statistics*. 3rd edition. Mcgraww-Hill Inc. Tokyo.

Real Statistics. (2018). *Anderson-Darling Test Table*. Real Statistics Using Excel. En: http://www.real-statistics.com/statistics-tables/anderson-darling-test-table/

Schwar, J.F. y Puy H.J. (1975). *Métodos estadísticos en Ingeniería de Tránsito"*. Co-editores. Asociación Mexicana de Caminos, A.C y Representaciones y Servicios de Ingeniería, S.A. México.

Statistics How To. (2018). *Empirical Distribution Function Definition.* Disponible en: https://www.statisticshowto.datasciencecentral.com/empirical-distribution-function/

Stephens, M. A. (1979). *The Anderson-Darling Statistic.* Technical Report No. 39. Department of Statistics. Stanford University. USA.

Upton, G. & Cook, I. (2002). *A Dictionary of Statistics.* Oxford University Press.UK.

Anexo. Referencias de las aplicaciones

Al-Gahmdi, A.S. (2000). “Probability approach for ranking high-accident locations”.*Urban Transport VI. C.A.* Brebbia & L.J. Sucharov (Editors). En: https://www.witpress.com/Secure/elibrary/papers/UT00/UT00050FU.pdf

Andreassen, D.C. (1986). “Interesection accident frequencies”. *Traffic Engineering + Control*. October 1986, pp. 514-517.

Bagnold, R.A. and Barndorff-Nielsen, O. (1980). “The pattern of natural size distributions”. *Sedimentology*.
En: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-3091.1980.tb01170.x

Baldeh, M., et al. (2016). “Poisson Process and Its Application to the Storm Water Overflows”. *Computational Water, Energy and Environmental Engineering*. *2016,5,47-53.* En: http://www.scirp.org/journal/cweee

Boll, J. et al, (1997). “Frequency distribution of water and solute transport properties derived from pan sampler data”. Water Resources Research, Vol. 33, No. 12, pages 2655–2664. En: http://soilandwater.bee.cornell.edu/publications/BollWRR97.pdf

Bowling, S.R. et al. (2009). “A logistic approximation to the cumulative normal distribution”. *Journal of Industrial Engineering and Management*. – 2(1): 114-127 – ISSN: 2013-0953. En: http://www.jiem.org/index.php/jiem/article/viewFile/60/27

Brown, W.K. and Wohletz, K.H. (1995). “Derivation of the Weibull distribution based on physical principles and its connection to the Rossin-Rammler and lognormal distributions”. *Journal of Applied Physics.* Vol.78, No. 4, 2758-2763., 15 August 1995. En: https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-94-3297

Büchel, B. and Corman, F. (2018). *Modelling probability distributions of public transport travel time components.* Conference Paper. ETH Zurich Research Collection. En: https://doi.org/10.3929/ethz-b-000263929.

Cabrera, J.B.D. et al. (2004). *On the Statistical Distribution of Processing Times in Network Intrusion Detection*. Proceedings of the 43rd IEEE Conference on Decision and Control, Bahamas, December 2004. En: http://wenke.gtisc.gatech.edu/papers/cdc04.pdf

Cooray, K. (2005). *Analyzing lifetime data with long-tailed skewed distribution: the logistic-sinh family*. Statistical Modelling Society SMS. En:

http://journals.sagepub.com/doi/abs/10.1191/1471082X05st099oa?journalCode=smja.

Dauxois, J.Y., Jomhoori, S. ad Yousefzadeh, F. (2014). ' Testing an "Exponential Delay Time model" against a "Random Sign Censoring model" in Reliability '. *Journal de la Société Française de Statistique.* Vol. 155 No. 3. En: https://dialnet.unirioja.es/servlet/articulo?codigo=5055099

Deni, S.D. and Jemain, A.A. (2009). "Mixed log series geometric distribution for sequences of dry days". Atmospheric Research. Vo. 92, Issue 2, April 2009, Pages 236-243.

Dianty, M.A; M Yahaya, A.S. and Ahmad, F. (2014)." Probability Distribution of Engineering Properties of Soil at Telecomunication Sites in Indonesia". *International Journal of Scientific Research in Knowledge,* 2(3), pp. 143-150. En: https://www.researchgate.net/publication/274973546_Probability_Distribution_of_Engineering_Properties_of_Soil_at_Telecomunication_Sites_in_Indonesia

Drâgulescu, A. and Yakovenko, V.M. (2001). "Evidence for the exponential distribution of income in the USA". *The European Physical Journal B* 20, 585-589. En: http://terpconnect.umd.edu/~yakovenk/papers/EPJB-20-585-2001.pdf

Fahidy, T.Z. (2012). *Application of the Negative Binomial/Pascal Distribution in Probability Theory to Electrochemical Processes*. Recent Trend in Electrochemical Science and Technology. En: https://www.intechopen.com/books/recent-trend-in-electrochemical-science-and-technology/application-of-the-negative-binomial-pascal-distribution-in-probability-theory-to-electrochemical-pr

Fieller, N.R. J; Gilbertson, D.D. and Olbricht, W. (1984). "A new method for environmental analysis of particle size distribution data from shoreline sediments". *Nature*. 311, 648-651. En: https://www.nature.com/articles/311648a0

Friedman, L.C., Bradford, W.L. & Peart, D.B. (1983). *Application of binomial distributions to quality assurance of quantitative chemical analyses*. Taylor & Francis Oonline. https://www.tandfonline.com/doi/abs/10.1080/10934528309375123.

Gerlough, D.L. (1955). *Use of Poisson Distribution in Highway Traffic*. Eno Foundation dfor Highway Traffic Control, Inc. Columbia University Press. USA

Harris, M. (2006). *Analysis and modelling of train delay data*. Dissertation submitted for the MSc in Mathematics with Modern Applications Department of Mathematics, University of York, Complexity Research Group, BT, Martlesham, UK. En: http://keithbriggs.info/documents/Mark_Harris_MSc_Dissertation_colour.pdf

Hashim, I.H. (2011). "Analysis of speed characteristics for rural two-lane roads: A field study from Minoufiya Governorate, Egypt". Ain Shams Engineering Journal (2011) 2, 43–52. En: https://ac.els-cdn.com/S2090447911000165/1-s2.0-

S2090447911000165-main.pdf?_tid=e507d5b2-26b2-4027-84de-9e3dd8b41919&acdnat=1538691277_e911e30d20e38d96782b1e150fe664b9

Husak, G.J., Michaelsen, J. and Funk, C. (2006). "Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications". *International Journal of Climatology.* Published online in Wiley InterScience. (www.interscience.wiley.com) DOI: 10.1002/joc.1441. En: . http://chg.ucsb.edu/publications/pdfs/2006_Husaketal_GammaDistribution.pdf

Jánosikova, L. and Slavik, M. (2014). "Modelling passenger's arrivals at public transport stops". *European Transport\ Trasporti Europei*. Issue 56, paper No. 7. En: http://www.istiee.org/te/papers/N56/P07_56_12_2014.pdf

Jiang, R. and Murthy, D.N.P. (2011). "A study of Weibull shape parameter: Properties and significance". *Reliability Engineering & System Safety.*Volume 96, Issue 12, December 2011, Pages 1619-1626. En: https://www.sciencedirect.com/science/article/pii/S095183201100175X?via%3Dihub

Jiao, T; Wen, X and Wang, X. (2013). "Study on Probability Distribution Models of the Subgrade Continuous Compaction Indicator". *Applied Mechanics and Materials* ISSN: 1662-7482, Vols. 438-439. En: https://www.scientific.net/AMM.438-439.1060

Knecht, W.R. (2015). *Predicting Accident Rates From General Aviation Pilot Total Flight Hours.* Report DOT/FAA/AM-15/3. Office of Aerospace Medicine. Federal Aviation Administration. En: http://libraryonline.erau.edu/online-full-text/faa-aviation-medicine-reports/AM15-03.pdf.

Kozik, M. (2014). *Application of binomial distribution to interpret ^{31}P NMR for aqueous solution of alpha-dodecatungstophosphoric acid, $H_3[PW_{12}O_{40}]$*. VIPEr. En: https://www.ionicviper.org/class-activity/application-binomial-distribution-interpret-31p-nmr-aqueous-solution-alpha.

Lee, K.M. et al. (2011). *Application of binomial and multinomial probability statistics to the sampling design process of a global grain tracing and recall system*. Food Control 22, pp 1085-1094.

Letkowski, J. (2012). *Applications of the Poisson Probability*. Academic and Business Research Institute. Disponible en: http://www.aabri.com/SA12Manuscripts/SA12083.pdf

Limpert, E. Stahel, W.A. and Abbt, M. (2001). "Log-normal Distributions acdoss the Sciences: Keys and Clues". *BioScience* May 2001/Vol. 51 No. 5, pp. 341- 352.

Mehri, H., Djemel, T. and Kammoun, H. (2006). *Solving of waiting lines models in the airport using queuing theory model and linear programming. The practice case: A.I.M.H.B.*HAL archives-ouvertes.fr. En: https://hal.archives-ouvertes.fr/hal-00263072/document

Önoz, B. and Bayazit, M. (1995). "Best.fit distribution of largest available flood samples". *Journal of Hydrology*. 167 (1-4): 195-208. En: https://www.researchgate.net/publication/222213660_Best-fit_distribution_of_largest_available_flood_samples

Poch, M. and Mannering, F. (1961). "Negative Binomial Analysis of Intersection-Accident Frequencies". *Journal of Transportation Engineering.* March/April 1996, pp. 105-113. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.178.4343&rep=rep1&type=pdf

Rajkai, K; Kabos, S; Van Genuchten, M.T. and Jansson, P.E. (1996). "Estimation of water-retention characteristics from the bulk density of Swedish soils". *Soil Science* 161(12):832-845. En: https://www.researchgate.net/publication/234203956_Estimation_of_water-retention_characteristics_from_the_bulk_density_of_Swedish_soils

Razali, A.M; Abidin, R.Z.; Zaharim, A. and Sopian, K. (2008). *Fitting of Statistical Distributions to Wind Speed Data*. 4th IASME/WSEAS International Conference on ENERGY, ENVIRONMENT, ECOSYSTEMS and SUSTAINABLE DEVELOPMENT (EEESD'08).Algarve, Portugal, June 11-13, 2008. En: http://www.wseas.us/e-library/conferences/2008/algarve/EEESD/023-588-353.pdf

Rehman, K; Burton, P.W. and Weatherill, G.A. (2018). "Application of Gumbel I and Monte Carlo methods to assess seismic hazard in and around Pakistan". *Journal of Seismology* .Volume 22, Issue 3, pp 575-588. En: https://link.springer.com/article/10.1007/s10950-017-9723-8

Ruimin, L., Chai, H. and Tang, J. (2013). "Empirical Study of Travel Time Estimation and Reliability". *Mathematical Problems in Engineering*. Volume 2013. Aritcle ID 504579. En: https://www.hindawi.com/journals/mpe/2013/504579/

Sexauer, J.M., McBee, K.D. & Bloch, K. A. (2011). *Applications of probability model to analyze the effects of electric vehicle chargers on distribution transformers*. 2011 IEEE Electrical Power and Energy Conference. En: https://ieeexplore.ieee.org/document/6070213.

Simon, L.J. (1962). *An introduction to the negative binomial distribution and its applications*. Proceedings. Vol. XLIX, Part I. No. 91. Causal Actuarial Society. En : https://www.casact.org/pubs/proceed/proceed62/62001.pdf

Singh, A., Singh, A. K. and Iaci, R.J. (2002). *Estimation of the Exposure Point Concentration Term Using a Gamma Distribution*. EPA Technology Support Center Issue. En: https://www.itrcweb.org/ism-1/references/289cmb02.pdf

Taylor, C.J. (1961). "The Application of the Negative Binomial Distribution to Stock Control Problems". *Operational Research Quarterly*. Volume 12. Number 2. En: https://www.jstor.org/stable/i353509

Turowski, J.M. (2010). “Probability distributions of bed load transport rates: A new derivation and comparison with field data”. *Water Resources Research*, Vol. 46.

TutorVista.com. (2018). *Logarithmic Distribution*. Disponible en: https://math.tutorvista.com/statistics/logarathmic-distribution.html

Van beek, P. (1978). “An application of the logistic density on a stochastic continuous review stock control model”. *Unternehmensforschung Operations Research* 22(1). En: https://www.researchgate.net/publication/40167730_An_application_of_the_logistic_density_on_a_stochastic_continuous_review_stock_control_model

Wikipedia. (2018a) *Particle-size distribution/Probability distributions*. En: https://en.wikipedia.org/wiki/Particle-size_distribution#Probability_distributions.

Wikipedia (2018b). *Hyperbolic distribution*. En: https://en.wikipedia.org/wiki/Hyperbolic_distribution

Wikipedia (2018c). *Log-Laplace distribution*. En: https://en.wikipedia.org/wiki/Log-Laplace_distribution

Wikipedia (2018d). *Birnbaum–Saunders distribution* En: https://en.wikipedia.org/wiki/Birnbaum%E2%80%93Saunders_distribution

WSDOT. (1994). *Statistical Mehtods for WSDOT Pavement and Material Applications*. WA-RD 315.1. Interim Report February 1994. En: https://www.wsdot.wa.gov/research/reports/fullreports/315.1.pdf

Xu, Y.L. (1995). “Model- and full-scale comparison of fatigue-related characteristics of wind pressures on the Texas Tech Building”. *Journal of Wind Engineering and Industrial Aerodynamics*. 58(3):147-173. En: https://www.researchgate.net/publication/222726298_Model-_and_full-scale_comparison_of_fatigue-related_characteristics_of_wind_pressures_on_the_Texas_Tech_Building

Yuan, J; Goverde, R.M.P. and Hansen, I.A. (2006). *Evaluating stochastic train process time distribution models on the basis of empirical detection data*. Conference paper: COMPRAIL 2006. En: https://www.researchgate.net/publication/271450505_Evaluating_stochastic_train_process_time_distribution_models_on_the_basis_of_empirical_detection_data

Zamani, H. and Ismail, N. (2010). “Negative Binomial-Lindley Distribution and Its Application”. *Journal of Mathematics and Statistics* 6 (1): 4-9. Science Publications.

Zobeck, T.M; Gill, T.E and Popham, T.W. (1999). *A two-parameter Weibull function to describe airborne dust particle size distributions*. Wiley Online library. En: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291096-9837%28199909%2924:10%3C943::AID-ESP30%3E3.0.CO;2-9

Km 12+000 Carretera Estatal 431 “El Colorado-Galindo”
Parque Tecnológico San Fandila
Mpio. Pedro Escobedo, Querétaro, México
CP 76703
Tel +52 (442) 216 9777 ext. 2610
Fax +52 (442) 216 9671

publicaciones@imt.mx

http://www.imt.mx/

Esta publicación fue desarrollada en el marco de un sistema de gestión de calidad certificada bajo la norma ISO 9001:2015