Simulación Aplicada a la Logística

Lectura 2

Distribuciones de Probabilidad

Profesor:
Ricardo Caballero, M.Sc.
ricardo.caballero@utp.ac.pa

Introducción: ¿Qué es estadística?

- **La estadística** es la ciencia de recopilar, organizar, analizar, interpretar y presentar datos.
- **Una estadística** es una medida única, expresada como un número, que se utiliza para resumir un conjunto de datos de muestra; por ejemplo, la altura promedio de los estudiantes en una universidad.

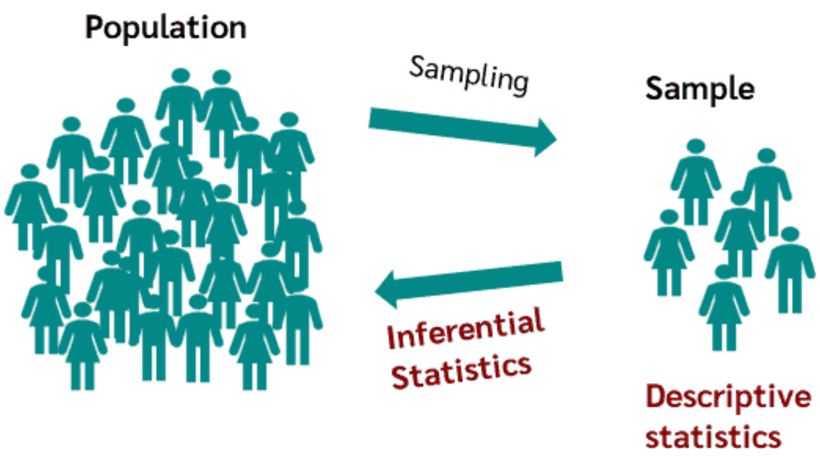
Hay dos tipos principales de estadísticas:

Estadística descriptiva

se refiere a la recopilación, presentación y resumen de datos (ya sea mediante cuadros y gráficos o mediante un resumen numérico).

Estadística inferencial

se refiere a generalizar de una muestra a una población, estimar parámetros poblacionales desconocidos, sacar conclusiones y tomar decisiones.



Introducción: Razones para estudiar estadística

- El conocimiento estadístico brinda a una empresa una ventaja competitiva frente a organizaciones que no pueden comprender los datos de su mercado interno o externo.

- El dominio de las estadísticas básicas le da a un gerente individual una ventaja competitiva a medida que avanza en el proceso de promoción o cuando cambia a un nuevo empleador.

- **Comunicación:** Comprender el lenguaje de las estadísticas facilita la comunicación y mejora la resolución de problemas.

- **Gestión de la información:** Las estadísticas ayudan a resumir cantidades grandes y pequeñas de información (datos) y a revelar relaciones subyacentes.

- **La mejora de procesos:** Las grandes empresas tienen sistemas formales para la mejora continua de la calidad. Las estadísticas ayudan a las empresas a supervisar a sus proveedores, monitorear sus operaciones internas e identificar problemas. La mejora de la calidad va mucho más allá de las estadísticas, pero se espera que cada graduado universitario conozca suficientes estadísticas para comprender su papel en la mejora de la calidad.

- **Revisión de cuentas:** La empresa ha sabido que algunas facturas se están pagando incorrectamente, pero no sabe hasta qué punto está extendido el problema. Se puede utilizar una muestra de facturas para estimar la proporción de facturas pagadas incorrectamente.

Introducción: Razones para estudiar estadística y sus usos en distintas áreas

- **Marketing:** Muchas empresas utilizan la gestión de relaciones con el cliente (CRM) para analizar los datos de los clientes de múltiples fuentes. Con herramientas estadísticas y analíticas como la correlación y la extracción de datos, identifican necesidades específicas de diferentes grupos de clientes y esto les ayuda a comercializar sus productos y servicios de forma más eficaz.

- **Cuidado de la salud:** Evalúe a 100 pacientes entrantes utilizando un cuestionario de evaluación física y mental de 42 ítems.

- **Mejora de calidad: i**niciar un programa de triple inspección, estableciendo sanciones para los trabajadores que produzcan productos de mala calidad.

- **Adquisición/compras:** Un productor de alimentos compra envases de plástico para envasar su producto. La inspección del envío más reciente de 500 contenedores encontró que 3 de ellos estaban defectuosos. La tasa histórica de defectos del proveedor es 0,005. ¿Ha aumentado realmente la tasa de defectos o se trata simplemente de un lote “malo”?

- **Medicina:** Determinar si un nuevo fármaco es realmente mejor que el placebo o si la diferencia se debe al azar.

- **Gestión de operaciones:** Administre el inventario pronosticando la demanda de los consumidores.

- **La garantía del producto:** Determine el costo promedio en dólares de los reclamos de garantía de un motor híbrido nuevo.

Introducción: Errores estadísticos más comunes

- La estadística es una parte esencial del **pensamiento crítico** porque nos permite probar una idea con evidencia empírica.

- Los **datos empíricos** representan datos recopilados a través de observaciones y experimentos.

- Se utilizan herramientas estadísticas para comparar ideas previas con datos empíricos, pero **ocurren errores**.

Error 1 — **Conclusiones de pequeñas muestras.** Tenga cuidado al hacer generalizaciones a partir de muestras pequeñas (por ejemplo, un grupo de 10 pacientes que mostraron mejoría).

Error 2 — **Conclusiones de muestras no aleatorias.** Tenga cuidado al hacer generalizaciones a partir de muestras pequeñas y de estudios retrospectivos de grupos especiales (p. ej., estudiar pacientes con ataque cardíaco sin definir un grupo de control equivalente).

Error 3 — **Conclusiones de eventos raros.** Tenga cuidado al sacar inferencias sólidas de eventos que no son sorprendentes cuando se analiza a toda la población (por ejemplo, ganar la lotería).

Error 4 — **Métodos de encuesta deficientes.** Tenga cuidado al utilizar métodos de muestreo deficientes o preguntas redactadas de manera vaga (por ejemplo, encuestas o cuestionarios anónimos).

Introducción: Errores estadísticos más comunes

Error 5 **Suponer un vínculo causal.** Tenga cuidado al sacar conclusiones cuando no existe un vínculo de causa y efecto (por ejemplo, los equipos que juegan en estadios con nombre (Citi Field para los Mets de Nueva York) tienden a perder más juegos de los que ganan). En realidad, son los jugadores y los entrenadores quienes determinan si un equipo gana.

Error 6 **Generalización a individuos.** Evite interpretar demasiado las generalizaciones estadísticas (por ejemplo, los hombres son más altos que las mujeres). Sí, pero sólo en un sentido estadístico. Los hombres son en promedio más altos, pero muchas mujeres son más altas que muchos hombres

Error 7 **Sesgo inconsciente.** Tenga cuidado de no permitir, inconsciente o sutilmente, que el sesgo influya en el manejo de los datos (por ejemplo, enfermedades cardíacas en hombres frente a mujeres). Los síntomas en los hombres son más evidentes que en las mujeres.

Error 8 **Significancia versus importancia.** Los efectos estadísticamente significativos pueden carecer de importancia práctica (por ejemplo, los reclutas militares austriacos nacidos en primavera son en promedio 0,6 cm más altos que los nacidos en otoño). ¿Alguien notaría esta diferencia?

Recolección de datos: datos y variables

Variable cualitativa

- Variable que describe o jerarquiza un elemento de la población
- Cuando el objeto se observa y registra como una característica no numérica

Ordinal	Nominal
Variable que toma distintos valores ordenados siguiendo una escala establecida.	Variable donde los valores no pueden ser sometidos a un criterio de orden.

- Nota en el examen: A, B, C, D, F
- Nivel de educación superior: licenciatura, maestría, doctorado, ..
- Clase social: alta, media, baja

- Afiliación religiosa: católico, evangélico, musulman,...
- Estado civil: soltero, casado, divorciado, viudo.
- Nacionalidad: panameño, colombiano, venezolano,…

Variable cuantitativa

- Es aquella variable que se expresa en cantidades con un valor numérico.
- Variable que cuantifica un elemento de una población

Discreta	Continua
Variable que puede tomar valores numéricos enteros	Variable que puede adquirir cualquier valor dentro de un rango específico de valores.

- Cantidad de hijos que tiene una familia: 0, 1, 2, 3, 4, …
- Edad: 19 años, 20 años, 21 años, 22 años,…
- Número de estudiantes en el salón: 15, 20, 30,…

- Peso: 70 kg, 70,5 kg, 71 kg, 71,5 kg,...
- Estatura: 1,68 m, 1,69 m, 1,70 m,…
- Volumen de botellas de agua: 0,5 litros, 1 litro, 1,5 litros,…

Recolección de datos: niveles de medición

Nivel de medición	Características	Ejemplo
Nominal	Solo categorías	Color de ojo (azul, marrón, verde, etc.)
Ordinal	El rango tiene significado. No hay un significado claro para distanciar	Rara vez, nunca
Intervalo	La distancia tiene significado	Temperatura (57º Celsius)
Razón	Existe un cero significativo	Cuentas por pagar ($21,7 millones)

Nominal

- Los datos nominales simplemente identifican una categoría.
- Los datos nominales son datos cualitativos, de atributos, categóricos o de clasificación, y pueden codificarse numéricamente (por ejemplo, 1 = Apple, 2 = Samsung, 3 = Dell, 4 = HP).
- Las únicas operaciones matemáticas son contar (por ejemplo, frecuencias) y estadísticas simples.

Ordinal

- Los códigos de datos ordinales se pueden clasificar (p. ej., 1 = Con frecuencia, 2 = A veces, 3 = Rara vez, 4 = Nunca).
- La distancia entre códigos de datos no es significativa (por ejemplo, la distancia entre 1 y 2, o entre 2 y 3, o entre 3 y 4 carece de significado).
- Existen muchas pruebas estadísticas útiles para datos ordinales, que son especialmente útiles en ciencias sociales, marketing y investigación de recursos humanos.

Intervalo

- Los datos no solo se pueden clasificar, sino que también pueden tener intervalos significativos entre los puntos de la escala (por ejemplo, la diferencia entre 60°F y 70°F es la misma que la diferencia entre 20°F y 30°F).
- Dado que los intervalos entre números representan distancias, se pueden realizar operaciones matemáticas (por ejemplo, promedio).
- El punto cero de las escalas de intervalo es arbitrario, por lo que las proporciones no son significativas (por ejemplo, 60°F no es dos veces más cálido que 30°F).

Razón

- Debido a este punto cero, las proporciones de los valores de los datos son significativas (por ejemplo, una ganancia de 20 millones de dólares es el doble que 10 millones de dólares).
- El cero no tiene por qué ser observable en los datos; es un punto de referencia absoluto.

Recolección de datos: escalas de Likert

Escalas Likert

- Un caso especial de datos de intervalo utilizados frecuentemente en la investigación por encuestas.
- La escala Likert se refiere al número de puntos de la escala (normalmente 5 o 7).
- La escala de Likert generalmente consta de una serie de afirmaciones o preguntas sobre un tema específico, y los encuestados deben indicar su grado de acuerdo o desacuerdo con cada afirmación. Por lo general, las opciones de respuesta se presentan en una escala ordinal que varía de 1 a 5 o 1 a 7, donde 1 representa un fuerte desacuerdo o una respuesta negativa, y 5 o 7 representan un fuerte acuerdo o una respuesta positiva. Los valores intermedios indican grados variables de acuerdo o desacuerdo.

"College-bound high school students should be required to study a foreign language." (check one)

❑ Strongly Agree ❑ Somewhat Agree ❑ Neither Agree Nor Disagree ❑ Somewhat Disagree ❑ Strongly Disagree

"How would you rate your Internet service provider?" (check one)

❑ Terrible ❑ Poor ❑ Adequate ❑ Good ❑ Excellent

Website User Survey

1. The website has a user friendly interface.
strongly agree — agree — neutral — disagree — strongly disagree

2. The website is easy to navigate.
strongly agree — agree — neutral — disagree — strongly disagree

3. The website's pages generally have good images.
strongly agree — agree — neutral — disagree — strongly disagree

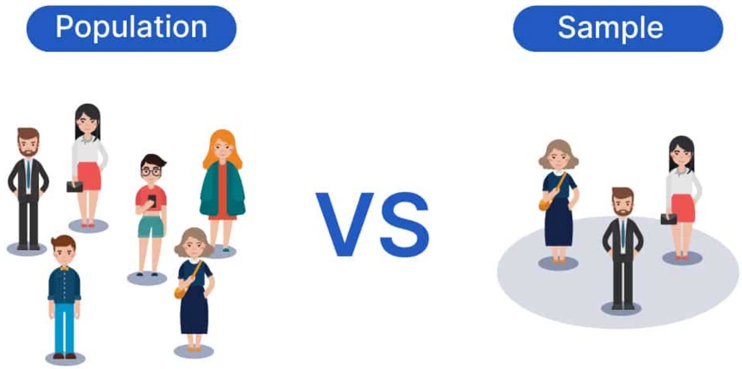
4. The website allows users to upload pictures easily.
strongly agree — agree — neutral — disagree — strongly disagree

5. The website has a pleasing color scheme.
strongly agree — agree — neutral — disagree — strongly disagree

Recolección de datos: muestreo

Población

involucra todos los artículos que nos interesan en el estudio. Puede ser finito (por ejemplo, todos los pasajeros de un avión) o infinito (por ejemplo, todas las Coca-Colas producidas en un proceso de embotellado en curso).



Muestra

es un subconjunto de la población e implica observar solo algunos de los elementos seleccionados de la población.

Métodos de muestreo aleatorio

Muestreo aleatoria simple	Utilice números aleatorios para seleccionar elementos de una lista (por ejemplo, titulares de tarjetas Visa).
Muestreo sistemática	Seleccione cada elemento k de una lista o secuencia (por ejemplo, clientes de un restaurante).
Muestreo estratificada	Seleccione aleatoriamente dentro de estratos definidos (por ejemplo, por edad, ocupación, género).
Muestreo de conglomerados	Seleccione regiones geográficas aleatorias (por ejemplo, códigos postales) que representen a la población.

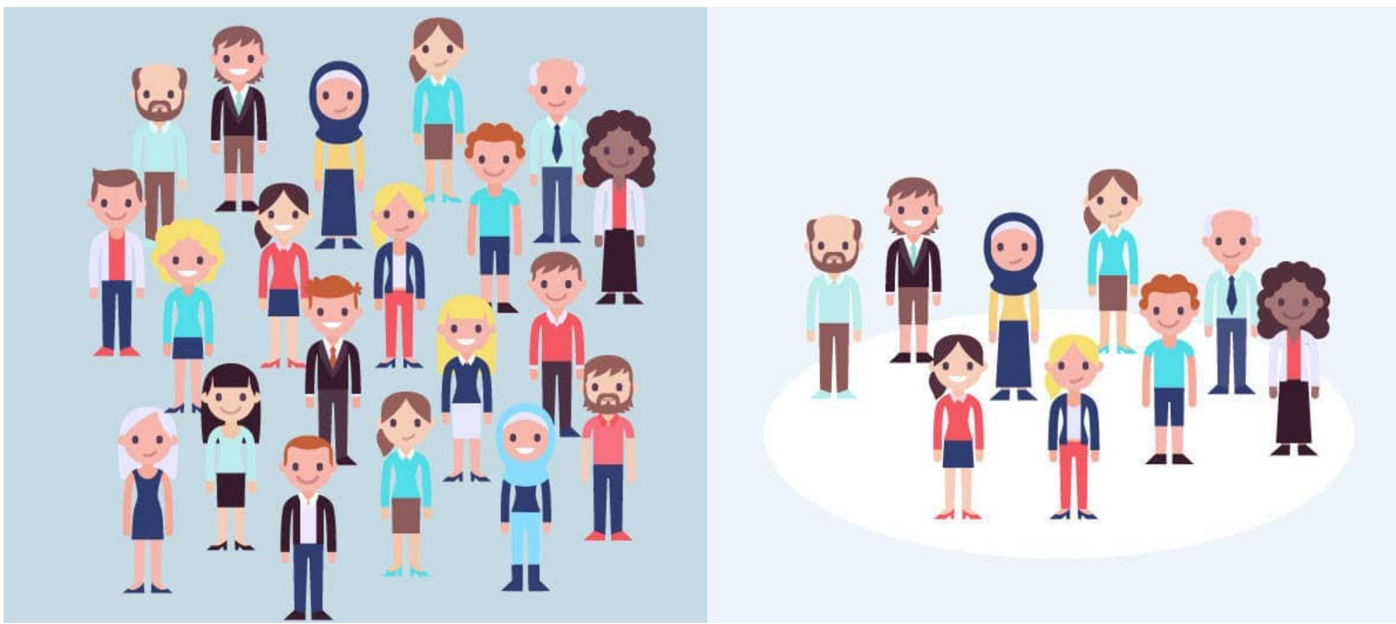
Recolección de datos: muestreo

Muestreo aleatorio simple

Generalmente denotamos el tamaño de la población por N y el tamaño de la muestra por n.

En una muestra aleatoria simple, cada elemento de la población de N elementos tiene la misma probabilidad de ser elegido en la muestra de n elementos.

Un experimento físico para lograr esto sería escribir cada uno de los N valores de datos en una ficha de póquer y luego sacar n fichas de un tazón después de agitarlo bien.

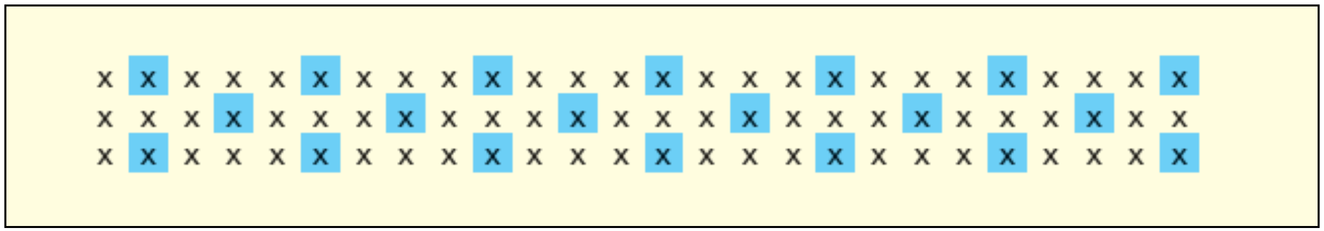


Recolección de datos: muestreo

Muestreo sistemático

Muestreo eligiendo cada k-ésimo elemento de una lista, comenzando desde una entrada elegida al azar en la lista.

Por ejemplo, comenzando en el elemento 2 (ver más abajo), tomamos muestras cada 4 elementos para obtener una muestra de n = 20 elementos de una lista de N = 78 elementos. Tenga en cuenta que N/n = 78/20 = 4 (periodicidad).



Recolección de datos: muestreo

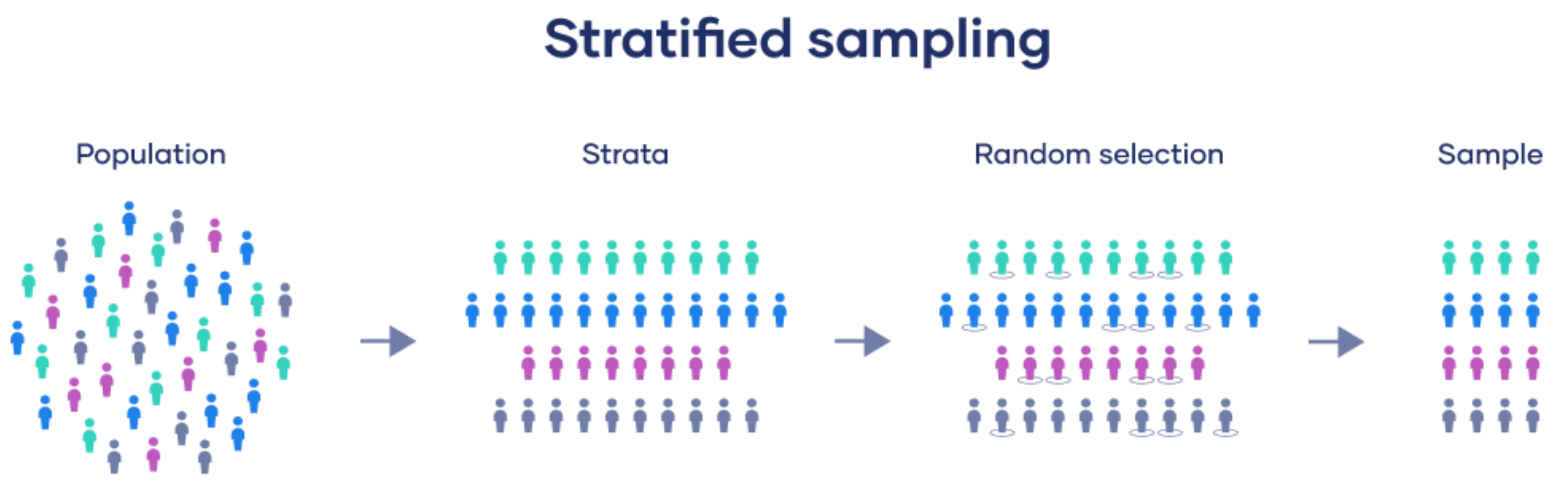
Muestreo estratificado

Utiliza información previa sobre la población.

Aplicable cuando la población puede dividirse en subgrupos relativamente homogéneos de tamaño conocido (estratos).

Dentro de cada estrato se toma una muestra aleatoria simple del tamaño deseado.

Alternativamente, se podría tomar una muestra aleatoria de toda la población y luego combinar las estimaciones de los estratos individuales utilizando ponderaciones apropiadas.



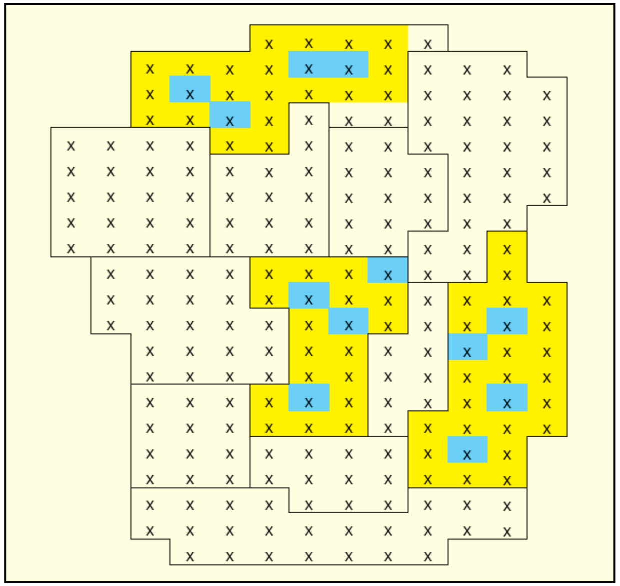
Recolección de datos: **muestreo**

Muestreo por conglomerados (cluster sample)

Los estratos consisten en regiones geográficas.

Muestreo por **conglomerados en una etapa:** la muestra consta de todos los elementos en cada una de las k subregiones (conglomerados) elegidas al azar.

Muestreo de **conglomerados en dos etapas**: primero se eligen k subregiones (conglomerados) y luego se elige una muestra aleatoria de elementos dentro de cada conglomerado.



¿Cómo se puede caracterizar la incertidumbre?

Maneras de caracterizar una distribución o un conjunto de datos:

Característica	Definición
Tendencia central	Indicadores estadísticos de posición que dan una descripción compacta de como están centrado los datos y una visualización más clara del nivel que alcanza la variable *¿Dónde se concentran los valores de los datos? ¿Cuáles parecen ser valores de datos típicos o medios? ¿Existe una tendencia central?*
Variabilidad	Grado en que los datos numéricos tienden a esparcirse alrededor de un valor promedio, es la variabilidad de un conjunto de datos. *¿Cuánta dispersión hay en los datos? ¿Qué tan dispersos están los valores de los datos? ¿Hay valores inusuales?*
Forma	La forma de una distribución puede evaluarse observando el histograma o comparando la media y la mediana. *¿Los valores de los datos están distribuidos simétricamente? ¿Sesgados? ¿Con un pico pronunciado? ¿Planos? ¿Bimodal?*

Medidas de tendencia central: Media

- Es una medida estadística que representa el comportamiento de todos los datos.
- La medida estadística de tendencia central más familiar es la media. Es la suma de los valores de los datos dividida por el número de elementos de datos.

Media aritmética

Se encuentra al sumar todos los valores de la variable x y dividir la suma entre el numero total de datos.

Media poblacional

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Donde:

$\mu =$ media poblacional

x_i = valores de los datos

N = población

Media muestral

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Donde:

$\bar{X} =$ media muestral

x_i = valores de los datos

n = muestra

Valor esperado

Promedio ponderado de los valores de la variable aleatoria

$$E(x) = \sum_{i=1}^{n} x_i P(x_i)$$

Donde:

$x_i = valores\ posibles\ de\ la\ variable\ aleatoria$

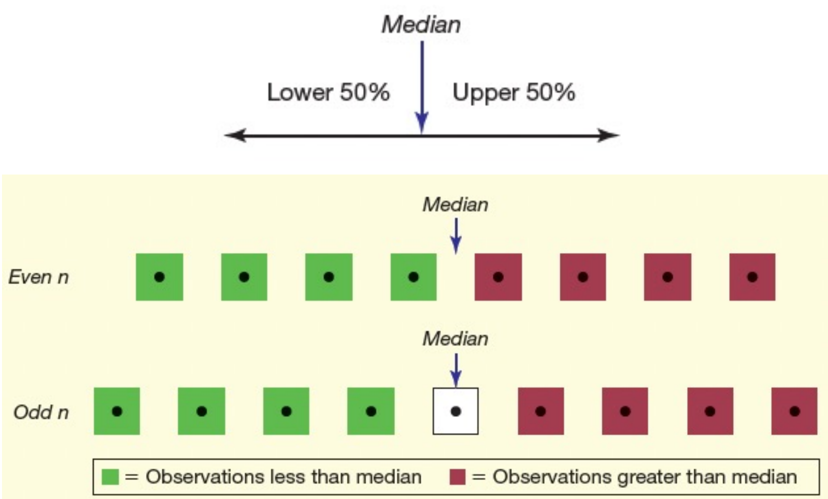
$P(x_i) = probabilidad\ de\ cada\ valor\ posible\ de\ la\ variable\ aleatoria$

$E(x) = valor\ esperado\ o\ media\ de\ la\ variable\ aleatoria$

Medidas de tendencia central: Mediana y moda

Mediana

- La mediana (M) es el percentil 50 o punto medio de los datos de la muestra ordenados.
- M separa las mitades superior e inferior de las observaciones ordenadas.
- Si n es impar, la mediana es la observación del medio en el conjunto de datos ordenado.
- Si n es par, la mediana es el promedio de las dos observaciones del medio en el conjunto de datos ordenado.



Moda

- El valor de los datos que ocurre con más frecuencia.
- Puede tener múltiples modas o ninguna moda.
- La moda es más útil para datos discretos o categóricos con solo unos pocos valores de datos distintos. Para datos continuos o datos con un amplio rango, la moda rara vez es útil.

Ejemplo: Caracterizando incertidumbre

Ventas por semana por tienda			
Semana	Tienda A	Tienda B	Tienda C
1	1	4	7
2	5	2	8
3	3	0	7
4	2	1	6
5	3	2	7
6	3	5	7
7	3	1	6
8	2	2	6
9	5	2	7
10	2	2	6
11	3	0	8
12	2	4	7
13	3	3	7
14	4	2	8
15	2	3	7
16	1	3	6
17	3	3	6
18	4	3	5
19	4	1	6
20	3	4	5
21	2	2	8
22	4	0	7
23	4	0	6
24	3	2	7
25	4	1	7
26	1	2	5
27	2	3	6
28	3	1	6
29	4	1	8
30	4	2	5
31	1	0	7
32	2	1	7
33	3	0	7
34	4	4	7
35	5	2	7
36	5	3	7
37	1	0	5
38	5	1	8
39	5	2	5
40	1	2	6
41	2	1	7
42	3	1	6
43	3	2	8
44	3	4	8
45	4	1	7
46	1	1	7
47	2	1	6
48	3	1	5
49	4	1	6
50	2	0	7

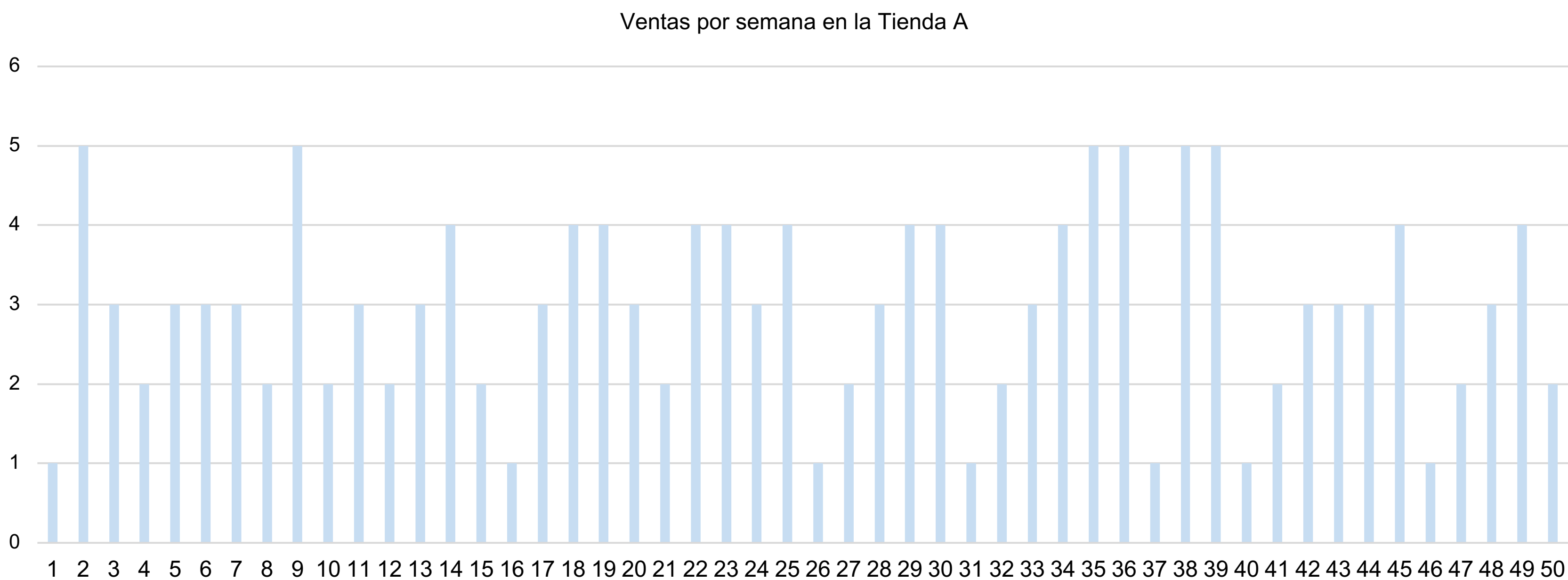
Supongamos que una empresa vende pastas de dientes

La empresa cuenta con tres sucursales y nos gustaría comparar el comportamiento de la demanda en cada una de ellas.

Se ha tomado una muestra de 50 semanas de ventas en cada sucursal

Analice el comportamiento de las ventas y brinde sus conclusiones

Ejemplo: Caracterizando incertidumbre – media aritmética



Semana	Unidades vendidas
1	1
2	5
3	3
4	2
5	3
6	3
7	3
8	2
9	5
10	2
11	3
12	2
13	3
14	4
15	2
16	1
17	3
18	4
19	4
20	3
21	2
22	4
23	4
24	3
25	4
26	1
27	2
28	3
29	4
30	4
31	1
32	2
33	3
34	4
35	5
36	5
37	1
38	5
39	5
40	1
41	2
42	3
43	3
44	3
45	4
46	1
47	2
48	3
49	4
50	2

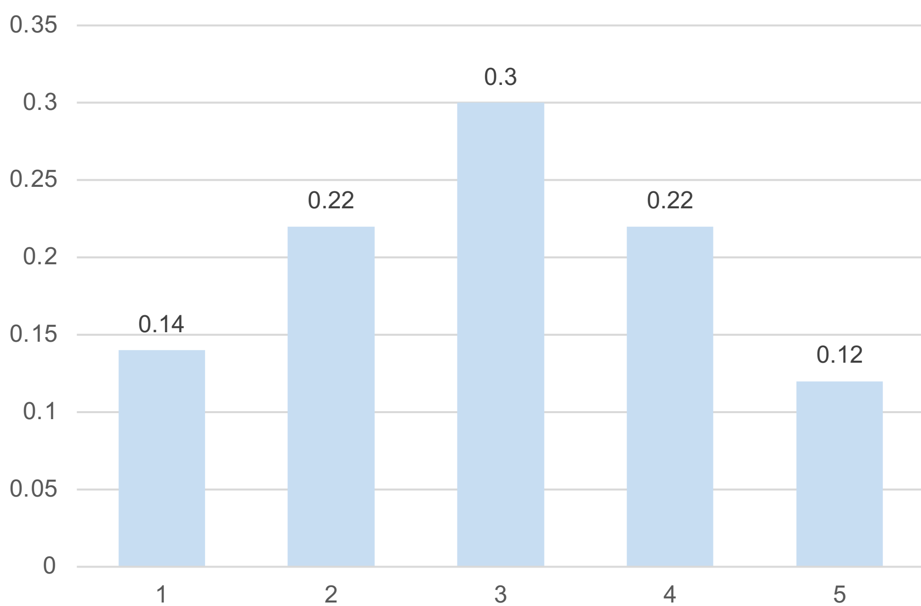
$$n = 50 \; semanas$$

$$\sum_{1}^{50} x_1 + x_2 + .. + x_{50} = 148 \; unidades$$

$$\mu = \frac{148 \; unidades}{50 \; semanas} = 2.96 \; unidades/semana$$

Ejemplo: Caracterizando incertidumbre – valor esperado

Unidades vendidas	Semanas	Probabilidad
1	7	0.14
2	11	0.22
3	15	0.3
4	11	0.22
5	6	0.12
	50	



$$E(x) = \sum_{i=1}^{n} x_i P(x_i)$$

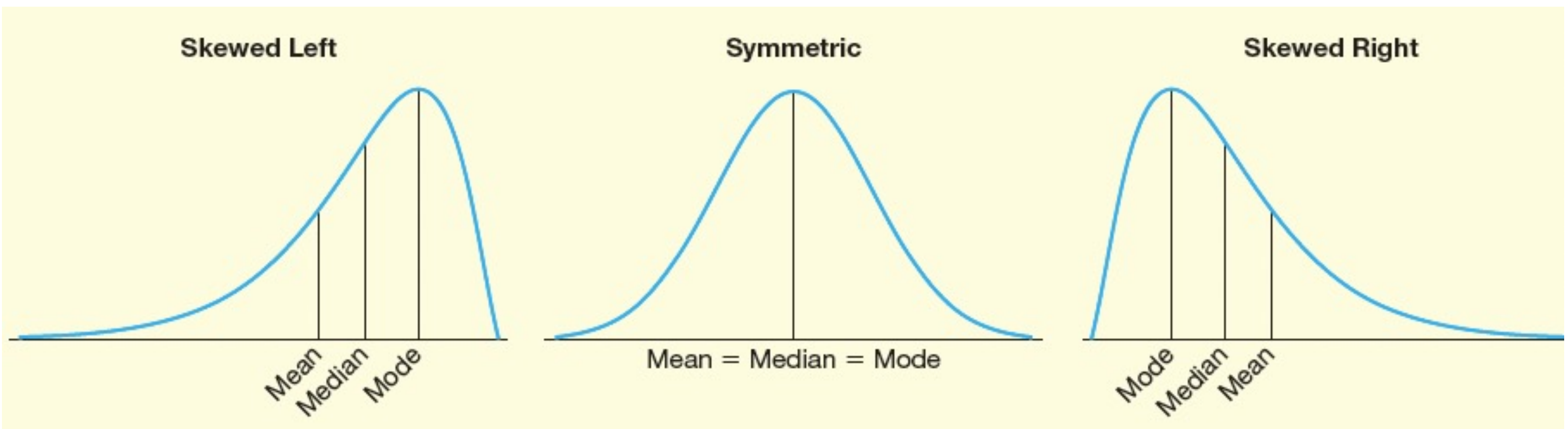
$$E(x) = \sum_{i=1}^{6} x_1 P(x_1) + x_2 P(x_2) + \ldots$$

$$= (1)(0.14) + 2(0.22) + 3(0.3) + 4(0.22) + 5(0.12)$$

$$E(x) = 2.96\, unidades/semana$$

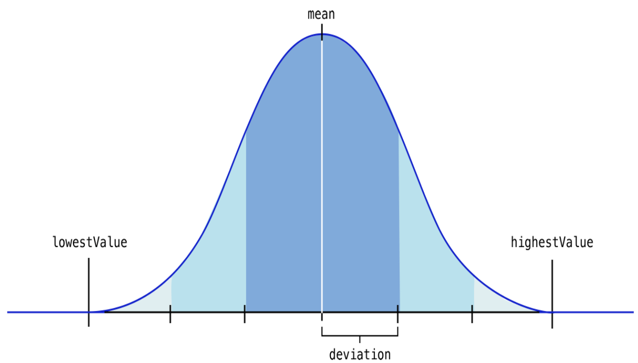
La forma de una distribución puede evaluarse observando el histograma o comparando la media y la mediana

Forma de la distribución	Apariencia del histograma	Estadísticos
Sesgado hacia la izquierda (sesgo negativo)	La cola larga del histograma apunta a la izquierda (unos pocos valores bajos pero la mayoría de los datos a la derecha)	Media < Mediana
Simétrico	Las colas del histograma están equilibradas (valores bajos/altos compensados)	Media ≈ Mediana
Sesgado hacia la derecha (sesgo positivo)	La cola larga del histograma apunta a la derecha (la mayoría de los datos a la izquierda pero unos pocos valores altos)	Media > Mediana



Medidas de variabilidad o dispersión

- La variación es la "dispersión" de los puntos de datos alrededor del centro de la distribución en una muestra.
- Grado en que los datos numéricos tienden a esparcirse alrededor de un valor promedio, es la variabilidad de un conjunto de datos.
- Si los datos se encuentran muy dispersos, la posición central es menos representativa de los datos, como un todo, que cuando estos se agrupan más cerca alrededor de la media.



1. Rango

Se define como la distancia entre el valor máximo y el valor mínimo de un conjunto total de datos.

$$Rango = Valor\ máximo\ - Valor\ mínimo$$

Características:

- Fácil de encontrar
- Puede haber valores muy alto y muy bajos que no necesariamente representan la media
- Solo consideran dos valores del total de datos
- Menos útil

2. Varianza

Es el promedio de las observaciones respecto a su media elevada al cuadrado

Significa que:

- Se encuentra la cantidad por la cual cada observación se desvía de la media.
- Se elevan al cuadrado tales desviaciones.
- Se halla la media de tales desviaciones elevadas al cuadrado

Para datos **no agrupados**

Varianza poblacional	Varianza muestral
$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$	$S^2 = \frac{\sum(x_i - \bar{X})^2}{n-1}$

Para **datos agrupados**

Varianza poblacional	Varianza muestral
$\sigma^2 = \frac{\sum f_i(x_i - \mu)^2}{N}$	$S^2 = \frac{\sum f_i(x_i - \bar{X})^2}{n-1}$

Medidas de variabilidad o dispersión

3. Desviación estándar

Es útil para describir como se apartan de la media la distribución de los elementos de forma individual.
La desviación estándar es un número único que nos ayuda a entender cómo los valores individuales en un conjunto de datos varían con respecto a la media.

Características:

- Uno de los parámetros estadísticos más importantes
- Más útil
- Calcula la distancia promedio en que los valores se alejan de la media
- Nunca negativa
- Solo 0 cuando todos sus datos son iguales.
- Es la raíz cuadrada de la varianza

Para datos **no agrupados**

Desviación estándar poblacional

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Desviación estándar muestral

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n - 1}}$$

Para **datos agrupados**

Desviación estándar poblacional

$$\sigma = \sqrt{\frac{\sum f_i(x_i - \mu)^2}{N}}$$

Desviación estándar muestral

$$S = \sqrt{\frac{\sum f_i(x_i - \bar{X})^2}{n - 1}}$$

Medidas de variabilidad o dispersión

4. Coeficiente de variación

- Medida relativa de dispersión, que puede compararse para diferentes distribuciones y que expresa la desviación estándar como porcentaje de la media
- Para comparar la dispersión en conjuntos de datos con unidades de medida disímiles (por ejemplo, kilogramos y onzas) o medias disímiles (por ejemplo, precios de viviendas en dos ciudades diferentes), definimos el coeficiente de variación (CV), que es una medida de dispersión libre de unidades:

$$C.V. = \frac{S}{\overline{X}} * 100$$

- Útil para la comparación de dos conjuntos de datos.
- El CV es la desviación estándar expresada como un porcentaje de la media.
- En algunos conjuntos de datos, la desviación estándar puede de hecho superar la media, así que el CV puede exceder el 100%.
- Esto puede suceder en conjuntos de datos sesgados, especialmente si hay valores atípicos.
- El CV es útil para comparar variables medidas en diferentes unidades.

Medidas de variabilidad o dispersión

5. Desviación absoluta de las medias (MAD)

- Una medida adicional de dispersión es la **desviación media absoluta (MAD).**
- Esta estadística revela la distancia promedio desde el centro. Se deben utilizar valores absolutos; de lo contrario, las desviaciones alrededor de la media sumarían cero.
- La MAD es atractiva debido a su interpretación simple y concreta. La MAD nos indica cuál es la distancia promedio desde un punto de datos individual.

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Medidas de posición

- Los **percentiles** son datos que han sido **divididos en 100 grupos**.

Por ejemplo, si obtienes un puntaje en el percentil 83 en una prueba estandarizada, eso significa que el 83% de las personas que hicieron la prueba obtuvieron un puntaje por debajo del tuyo.

- Los **deciles** son datos que han sido **divididos en 10 grupos.**
- Los **cuartiles** son datos que han sido **divididos en 4 grupos.**

Cuartiles son puntos de escala que dividen los datos ordenados en cuatro grupos de tamaño aproximadamente igual.

<table>
<tr><td></td><td>Q_1</td><td></td><td>Q_2</td><td></td><td>Q_3</td><td></td></tr>
<tr><td>⇦Lower 25%⇨</td><td>|</td><td>⇦Second 25%⇨</td><td>|</td><td>⇦Third 25%⇨</td><td>|</td><td>⇦Upper 25%⇨</td></tr>
</table>

Los tres valores que separan los cuatro grupos se llaman Q1, Q2 y Q3, respectivamente.

Medidas de posición

El **segundo cuartil Q_2** es la **mediana**, una medida de tendencia central.

	Q_2	
⇦ Lower 50% ⇨	|	⇦ Upper 50% ⇨

Q_1 y Q_3 miden la dispersión ya que el **rango intercuartil $Q_3 - Q_1$** mide el grado de dispersión en el 50 por ciento medio de los valores de datos.

	Q_1		Q_3	
⇦Lower 25%⇨	|	⇦ Middle 50% ⇨	|	⇦Upper 25%⇨

El primer cuartil Q_1 es la mediana de los valores de datos que están por debajo de Q_2, y el tercer cuartil Q_3 es la mediana de los valores de datos que están por encima de Q_2.

	Q_1		Q_2		Q_3	
⇦Lower 25%⇨	|	⇦Second 25%⇨	|	⇦Third 25%⇨	|	⇦Upper 25%⇨

Para la primera mitad de los datos, el 50% está por encima, el 50% por debajo de Q_1.

Para la segunda mitad de los datos, el 50% está por encima, el 50% por debajo de Q_3.

Ejemplo: Caracterizando incertidumbre - Dispersión

Unidades vendidas (xi)	Semanas	Probabilidad (pi)	xipi	xi-μ	(xi−μ)^2	pi (xi−μ)^2
1	7	0.14	0.14	-1.96	3.84	0.5378
2	11	0.22	0.44	-0.96	0.92	0.2028
3	15	0.3	0.9	0.04	0.00	0.0005
4	11	0.22	0.88	1.04	1.08	0.2380
5	6	0.12	0.6	2.04	4.16	0.4994
	50		2.96			

Varianza

$$\sigma^2 = \sum_{i=1}^{n} P(x_i)(x_i - \mu)^2$$

$$\sigma^2 = 1.4784$$

Desviación estándar

$$\sigma = \sqrt{1.4784}$$

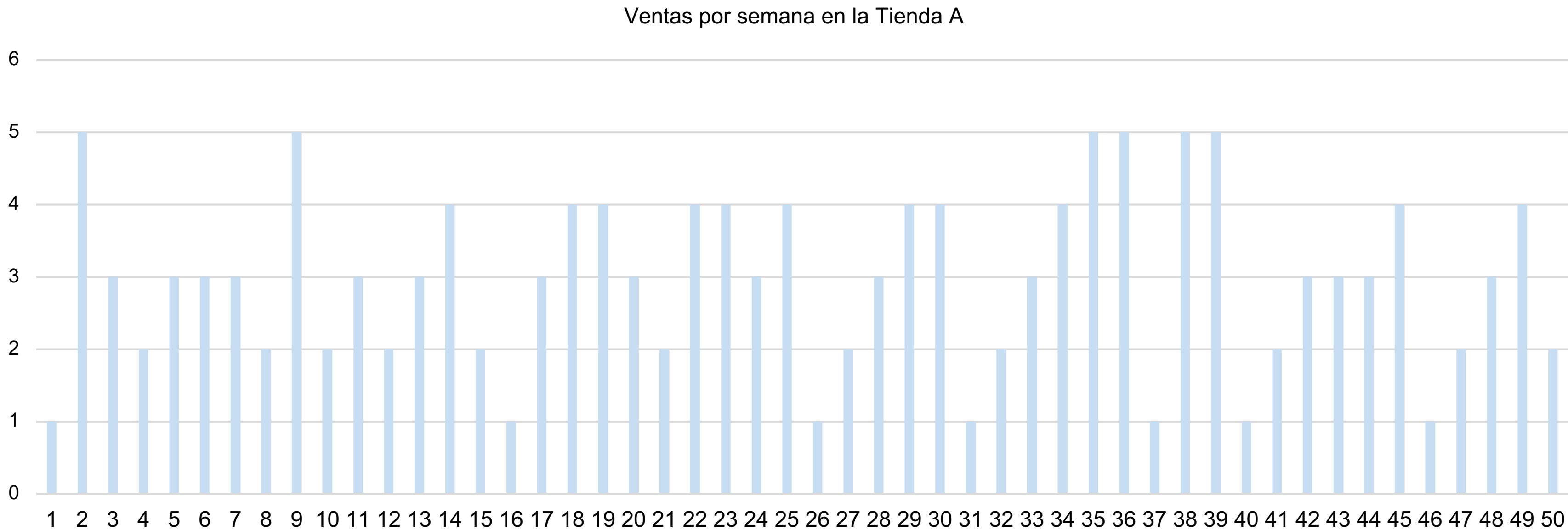
$$\sigma = 1.2159$$

Coeficiente de variación

$$C.V. = \frac{\sigma}{\mu} * 100$$

$$C.V. = \frac{1.2159}{2.96} * 100 = 0.411$$

Ejemplo: Caracterizando incertidumbre - Resumen



Valor mínimo = 1

Percentil 25 = 2

Media = 2.96

Mediana = Percentil 50 = 3

Moda = 3

Percentil 75 = 4

Valor máximo = 5

Rango = 4

Varianza = 1.48

Desviación estándar = 1.215

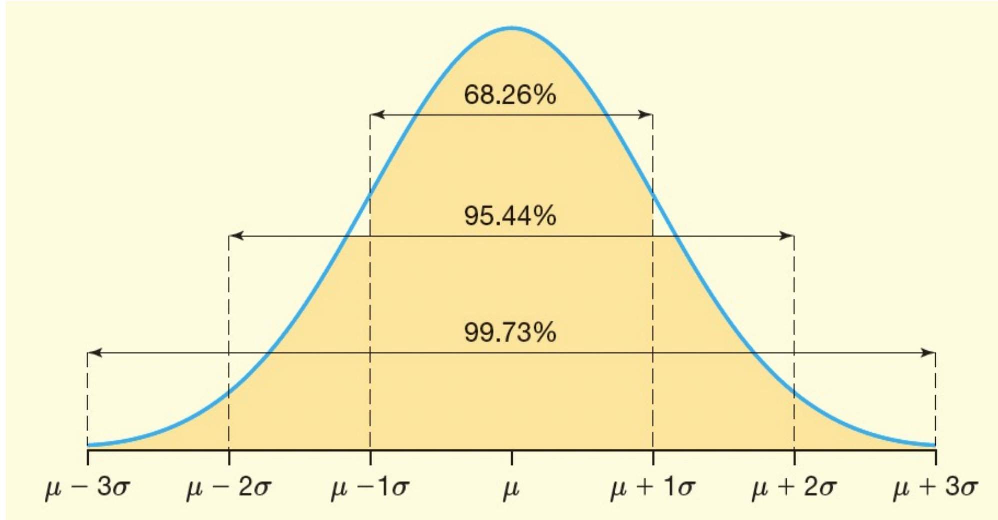
Coeficiente de variación = 0.411

Regla empírica

Distribución normal es una distribución de datos continuos (no discretos) que produce una curva simétrica en forma de campana

Regla empírica:

- El 68% de los datos están entro de más o menos una desviación estándar de la media, (entre - 1σ y 1σ)
- El 95% de los datos están entro de más o menos dos desviaciones estándar de la media (entre - 2σ y 2σ)
- El 99.7% de los datos están entro de más o menos tres desviaciones estándar de la media (entre - 3σ y 3σ)



Variable estandarizada Z

- Es la variable que mide la desviación respecto a la media en términos de unidades de desviaciones estándar
- Permite comparar la desviación estándar entre dos conjunto de datos
- Da como resultado cuantas desviaciones estándar lejos de la media se encuentra el valor analizado.

$$z_i = \frac{x_i - \mu}{\sigma}$$ Para una población

$$z_i = \frac{x_i - \overline{x}}{s}$$ Para una muestra

Outliers

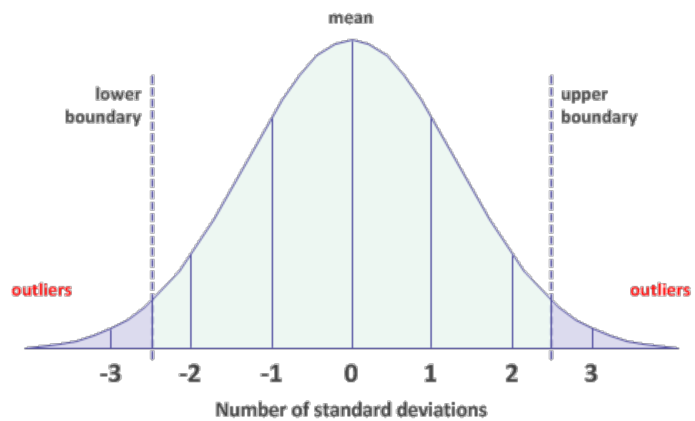
Outliers o valores atípicos

- Estos son puntos de datos que se diferencian drásticamente de otros valores en un conjunto de datos.
- Los valores atípicos pueden ser resultado de variabilidad en la medición o indicar un error experimental; en algunos casos, también pueden ser indicativos de una distribución con colas pesadas o ser simplemente variaciones extremas en los datos.
- Identificar valores atípicos es importante para el análisis estadístico, ya que pueden afectar significativamente los resultados y conclusiones.

Basado en su puntuación z estandarizada, se clasifica un valor de datos como:

Inusual si $|z_i| > 2$ (más allá de $\mu \pm 2\sigma$)

Outlier si $|z_i| > 3$ (más allá de $\mu \pm 3\sigma$)

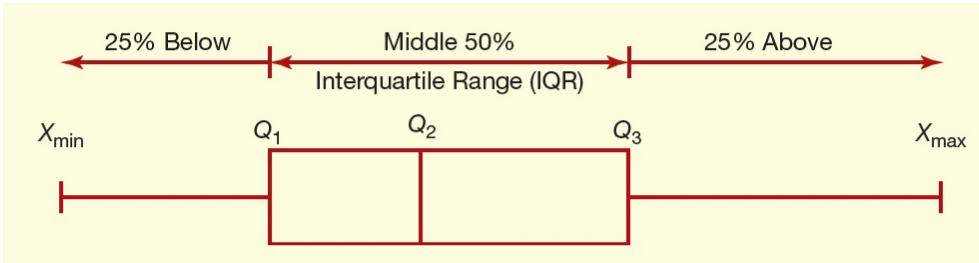


Box plots

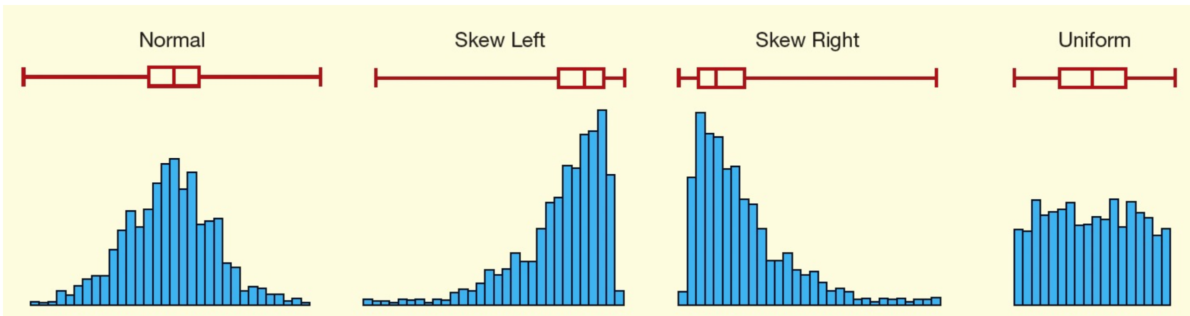
Diagrama de cajas

- Una herramienta útil del análisis exploratorio de datos (EDA) es el **diagrama de caja** (también llamado diagrama de caja y bigotes) basado en el resumen de cinco números:

$$x_{min}, Q_1, Q_2, Q_3, x_{max}$$



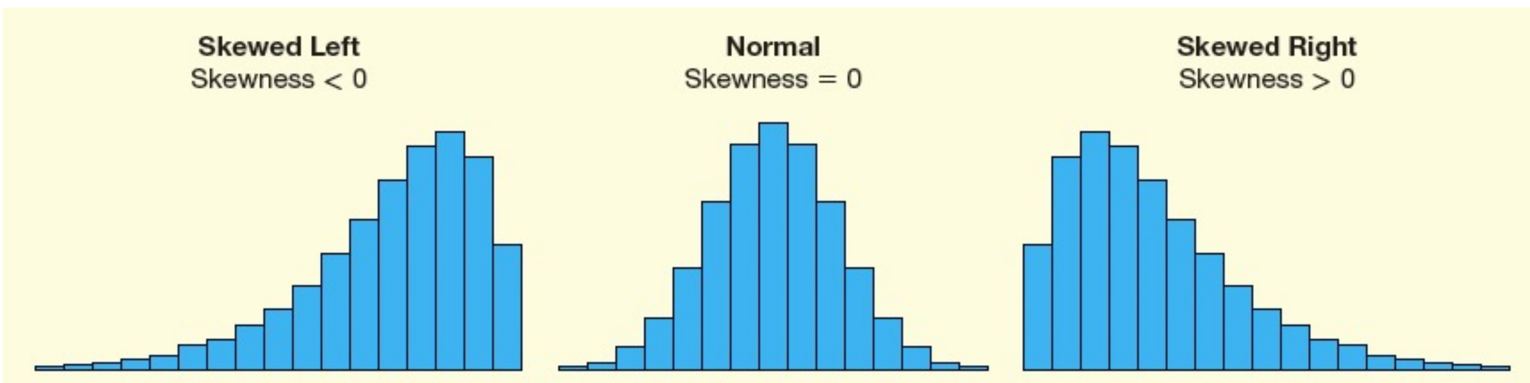
- Un diagrama de caja muestra el centro (posición de la mediana Q_2). Un diagrama de caja muestra la variabilidad (ancho de la "caja" definido por Q_1 y Q_3 y el rango entre x_{min} y x_{max}).
- Un diagrama de caja muestra la forma (asimetría si los "bigotes" son de longitud desigual y/o si la mediana no está en el centro de la caja).



Asimetría y curtosis

- En datos simétricos, la media y la mediana son aproximadamente iguales.
- Cuando los datos están sesgados hacia la derecha (o tienen sesgo positivo), la media supera a la mediana.
- Cuando los datos están sesgados hacia la izquierda (o tienen sesgo negativo), la media está por debajo de la mediana.

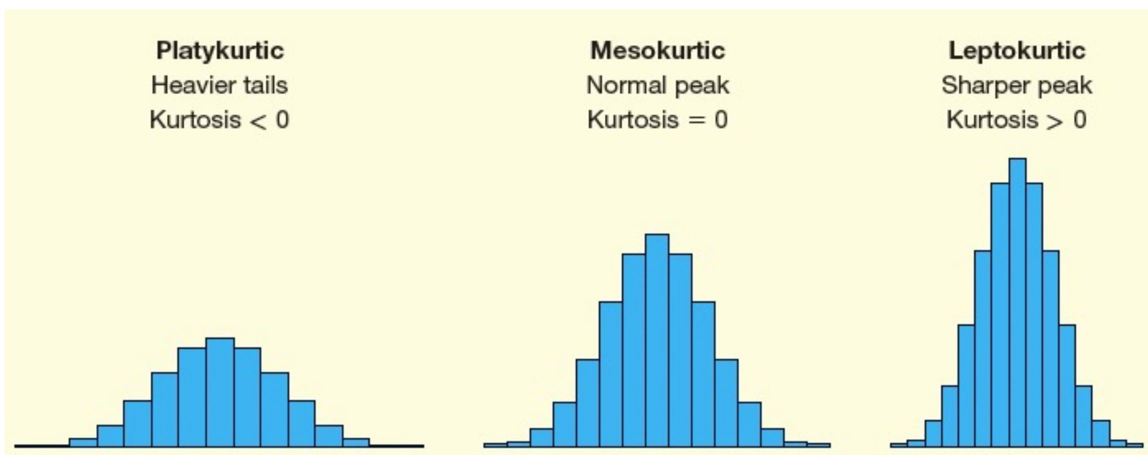
$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$



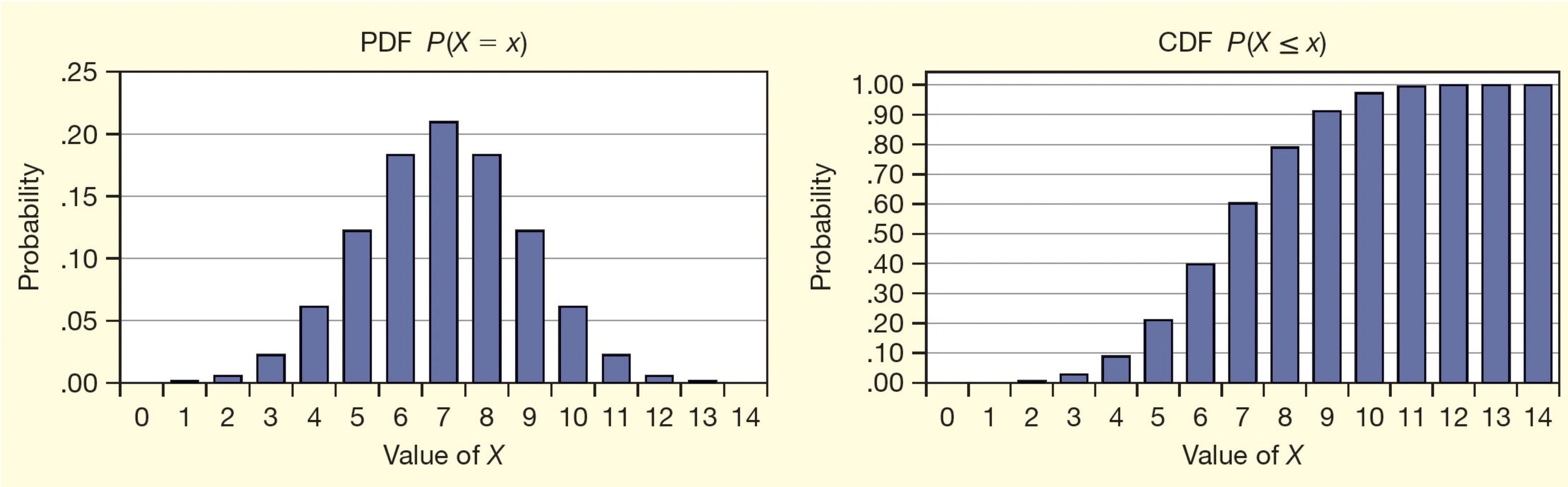
Asimetría y curtosis

- La curtosis se refiere a la longitud relativa de las colas y el grado de concentración en el centro.
- Una población con distribución normal se llama mesocúrtica y sirve como referencia.
- Una población que es más plana que una población normal (es decir, tiene colas más pesadas) se llama platicúrtica, mientras que una que es más puntiaguda que una población normal (es decir, tiene colas más delgadas) se llama leptocúrtica.
- La curtosis no es lo mismo que la variabilidad, aunque ambos conceptos suelen confundirse.
- Un histograma es una guía poco fiable para la curtosis porque su escala y las proporciones de los ejes pueden variar, por lo que se necesita una estadística numérica:

$$\text{Kurtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$



Distribuciones de probabilidad



Función de probabilidad (PDF)

Función de distribución acumulada (CDF)

Una función de distribución de probabilidad (PDF) es una función matemática que muestra la probabilidad de cada valor X.

Una función de distribución acumulada (CDF) es una función matemática que muestra la suma acumulada de probabilidades, sumando desde el valor X más pequeño al más grande, acercándose gradualmente a la unidad.

- Las variables aleatorias y sus distribuciones se describen mediante sus parámetros.
- Las ecuaciones para la PDF, la CDF y las características de la distribución (como la media y la desviación estándar) dependerán de los parámetros del proceso.

Distribución de probabilidad discreta

- Describe el comportamiento de una variable aleatoria.
- Los valores de la distribución son valores enteros

- Los valores de una distribución discreta pueden pueden provenir de fuentes:
  - **Empíricas** → basado en data actual o real (e.g. medidas, reportes de ventas, etc)
  - **Teórica** →basado en formas matemáticas

- Una distribución de probabilidad discreta asigna una probabilidad a cada valor de una variable aleatoria discreta X.

- Para que sea una distribución de probabilidad válida, se deben satisfacer las siguientes condiciones.

- Si X tiene n valores distintos $x_1, x_2, \ldots, x_n$, entonces

$$0 \leq P(x_i) \leq 1$$

$$\sum_{i=1}^{n} P(x_i) = 1$$

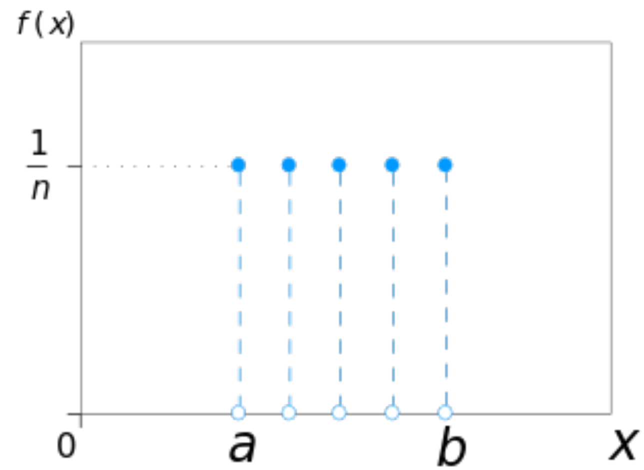
Qué tipo de fuente es mejor?

- *Depende de lo que se desee alcanzar*
- *Las distribuciones empíricas siguen un historial pasado*
- *Las distribuciones teóricas permiten un modelado más robusto*
- *Usualmente, se busca por la distribución teórica que encaje con los datos.*

Distribución Uniforme Discreta

La distribución uniforme discreta describe una variable aleatoria con un número finito de valores enteros de *a* - *b* (los únicos dos parámetros).

Cada valor de la variable aleatoria tiene la misma probabilidad de ocurrir.



PMF

Notación

$$U(a, b)$$

$$n = b - a + 1$$

Donde:
a = valor $mínimo$
b = valor $máximo$

Media y mediana

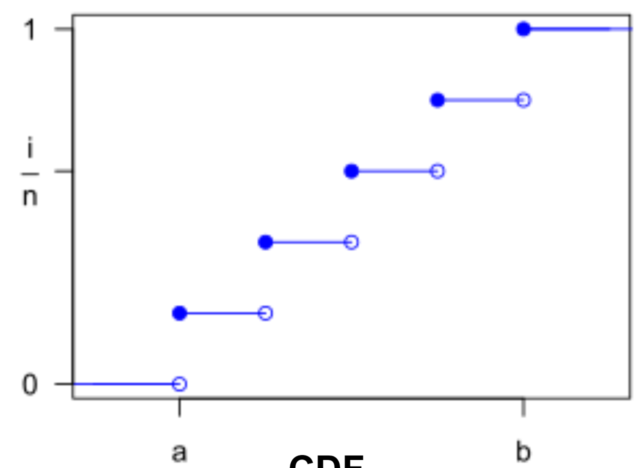
$$\mu = \frac{a+b}{2}$$

Varianza

$$\sigma^2 = \frac{(b-a+1)^2 - 1}{12}$$

Función de densidad (PMF)

$$P[X = x] = f(x|a, b) = \begin{cases} \frac{1}{n}, & a \leq x \leq b \\ 0, & en\ otro\ caso \end{cases}$$



CDF

Distribución Uniforme Discreta

Parameters	a = lower limit b = upper limit
PDF	$P(X = x) = \dfrac{1}{b - a + 1}$
CDF	$P(X \leq x) = \dfrac{x - a + 1}{b - a + 1}$
Domain	$x = a, a + 1, a + 2, \ldots, b$
Mean	$\dfrac{a + b}{2}$
Standard deviation	$\sqrt{\dfrac{[(b - a) + 1]^2 - 1}{12}}$
Random data generation in Excel	=RANDBETWEEN(a, b)
Comments	Useful as a benchmark, to generate random integers for sampling, or in simulation models.

Ejemplo: Distribución Uniforme Discreta

Suponga que se lanza un dado. Cuál es la probabilidad que salga 1, 2, 3, 4, 5 o 6?

xi	fi	pi
1	1	0.1667
2	1	0.1667
3	1	0.1667
4	1	0.1667
5	1	0.1667
6	1	0.1667
	6	

Media

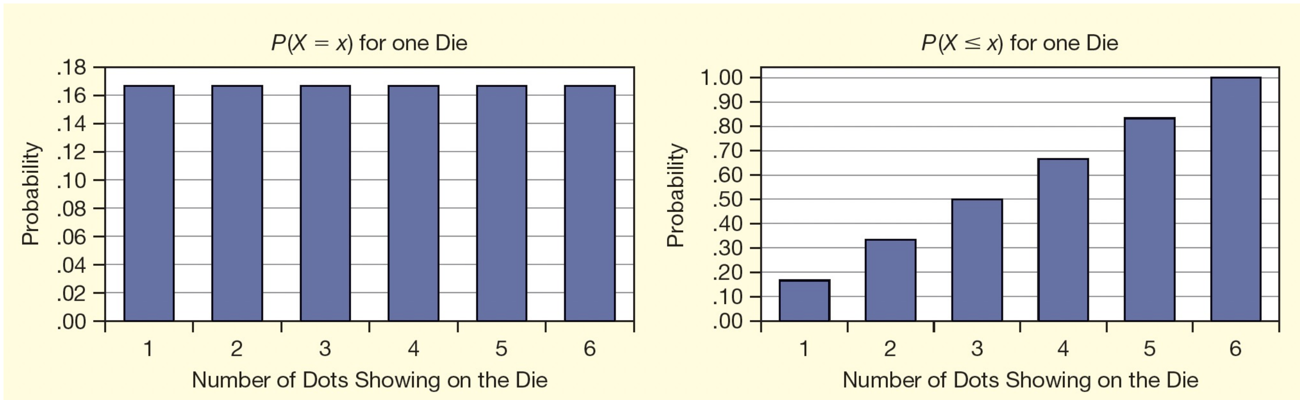
$$\mu = \frac{a+b}{2} = \frac{1+6}{2} = 3.5$$

Varianza

$$\sigma^2 = \frac{(b-a+1)^2-1}{12} = \frac{(6-1+1)^2-1}{12} = 2.916$$

Desviación estándar

$$\sigma = \sqrt{\sigma^2} = 1.708$$



Distribución de Poisson

- Describe el número de ocurrencias dentro de una unidad de tiempo (por ejemplo, minuto, hora) La traducción en el contexto de estadística es la siguiente:
- La distribución de Poisson describe el número de ocurrencias dentro de una unidad de tiempo (por ejemplo, minuto, hora)
- Los eventos ocurren de manera aleatoria e independiente a lo largo de un continuo de tiempo o espacio.
- Llamaremos al continuo 'tiempo' ya que la aplicación más común de Poisson es modelar las llegadas por unidad de tiempo."

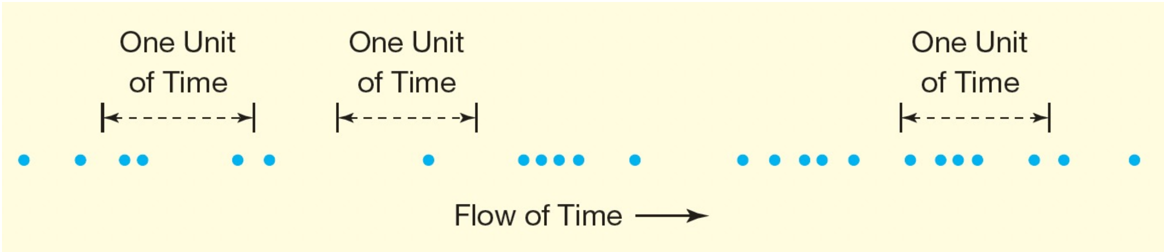
$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \text{for } X = 1, 2, 3, 4, \ldots$$

Donde:
λ = es el número promedio de resultados por unidad de tiempo, distancia,
e = 2.71828…
x! = factorial de x

Media λ

Distribución de Poisson: **características**

- Todas las características del modelo de Poisson están determinadas por su media λ.
- La constante e es aproximadamente 2,71828
- La media de la distribución de Poisson es λ y su desviación estándar es la raíz cuadrada de la media.
- La simplicidad de las fórmulas de Poisson las convierte en un modelo atractivo
- El número de eventos que pueden ocurrir en una unidad de tiempo dada no está acotado.
- Las probabilidades de Poisson disminuyen gradualmente hacia cero a medida que aumenta X, por lo que el rango efectivo suele ser pequeño.
- Un evento de interés ocurre aleatoriamente en el tiempo o en el espacio.
- La tasa de llegada promedio (λ) permanece constante.
- Las llegadas son independientes entre sí.
- La variable aleatoria (X) es el número de eventos dentro de un intervalo de tiempo observado.



Cada punto (•) representa una ocurrencia del evento de interés.

Distribución de Poisson: características

- A menudo llamamos a la distribución de Poisson el **modelo de llegadas** (clientes, defectos, accidentes).
- Las llegadas pueden considerarse razonablemente como eventos de Poisson si cada evento es independiente (es decir, la ocurrencia de cada evento no tiene efecto sobre la probabilidad de que ocurran otros eventos). Algunas situaciones carecen de esta característica.
- Por ejemplo:
  - *X = número de clientes que llegan a un cajero automático de un banco en un minuto dado.*
  - *X = número de infecciones por virus en un servidor de archivos en un centro de datos durante un período de 24 horas.*
  - *X = número de llegadas de pacientes con asma en una hora dada en una clínica sin cita previa.*
  - *X = número de paradas de motor en aviones Airbus 330 por cada 100,000 horas de vuelo.*
  - *X = número de imperfecciones por hoja de papel bond blanco.*

Distribución de Poisson: características

Parameter	λ = mean arrivals per unit of time or space
PDF	$P(X = x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$
Excel* PDF	=POISSON.DIST(x, λ, 0)
Excel* CDF	=POISSON.DIST(x, λ, 1)
Domain	$x = 0, 1, 2, \ldots$ (no obvious upper limit)
Mean	λ
Standard deviation	$\sqrt{\lambda}$
Comments	Always right-skewed, but less so for larger λ.

Ejemplo: Distribución de Poisson

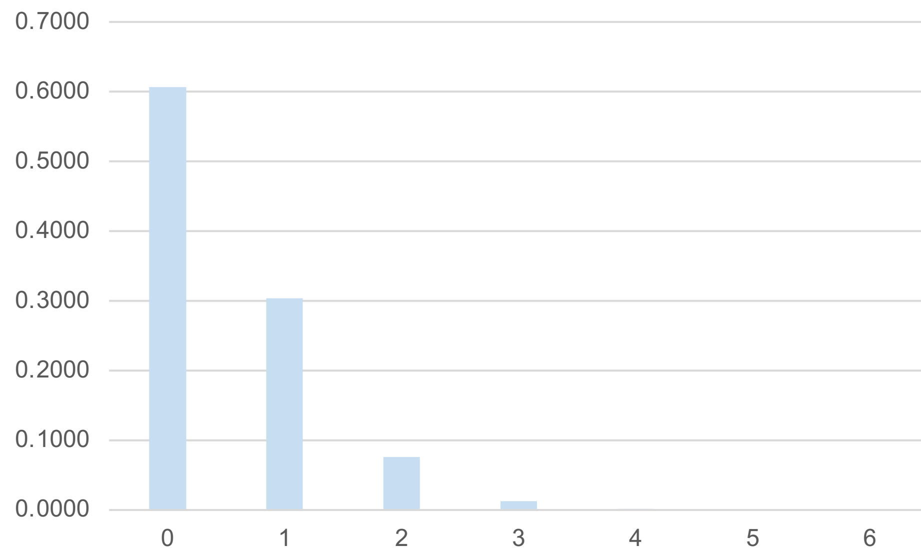
Suponga que $\lambda = 0.5$. Cómo se comportará la distribución?
Luego pruebe con $\lambda = 2\ , 5, 7\ y\ 10$

Ejemplo: Distribución de Poisson

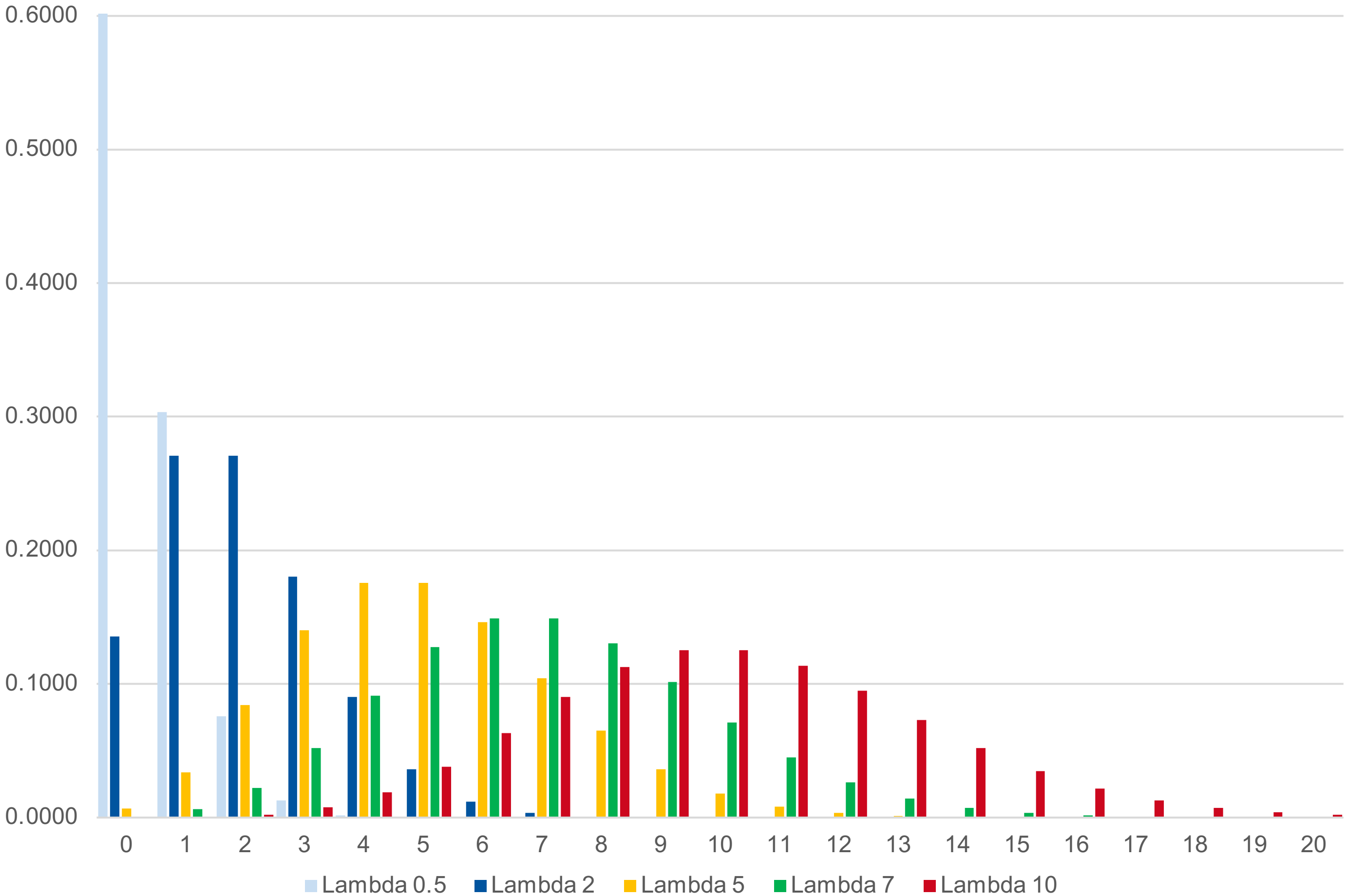
Suponga que $\lambda = 0.5$. Cómo se comportará la distribución?

$\lambda = 0.5$

xi	Pi
0	0.6065
1	0.3033
2	0.0758
3	0.0126
4	0.0016
5	0.0002
6	0.0000



Ejemplo: Distribución de Poisson



	Lambda				
xi	0.5	2	5	7	10
0	0.6065	0.1353	0.0067	0.0009	0.0000
1	0.3033	0.2707	0.0337	0.0064	0.0005
2	0.0758	0.2707	0.0842	0.0223	0.0023
3	0.0126	0.1804	0.1404	0.0521	0.0076
4	0.0016	0.0902	0.1755	0.0912	0.0189
5	0.0002	0.0361	0.1755	0.1277	0.0378
6	0.0000	0.0120	0.1462	0.1490	0.0631
7	0.0000	0.0034	0.1044	0.1490	0.0901
8	0.0000	0.0009	0.0653	0.1304	0.1126
9	0.0000	0.0002	0.0363	0.1014	0.1251
10	0.0000	0.0000	0.0181	0.0710	0.1251
11	0.0000	0.0000	0.0082	0.0452	0.1137
12	0.0000	0.0000	0.0034	0.0263	0.0948
13	0.0000	0.0000	0.0013	0.0142	0.0729
14	0.0000	0.0000	0.0005	0.0071	0.0521
15	0.0000	0.0000	0.0002	0.0033	0.0347
16	0.0000	0.0000	0.0000	0.0014	0.0217
17	0.0000	0.0000	0.0000	0.0006	0.0128
18	0.0000	0.0000	0.0000	0.0002	0.0071
19	0.0000	0.0000	0.0000	0.0001	0.0037
20	0.0000	0.0000	0.0000	0.0000	0.0019

Ejemplo: Distribución de Poisson

El jueves por la mañana entre las 9 a.m. y las 10 a.m., los clientes llegan y entran en la cola en la Cooperativa de Crédito a una tasa promedio de 1.7 clientes por minuto. Utilizando las fórmulas de Poisson con, encuentra la función de densidad de probabilidad (PDF), la media y la desviación estándar.

$$\text{PDF: } P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{(1.7)^x e^{-1.7}}{x!}$$

$$\text{Mean: } \lambda = 1.7$$

$$\text{Standard deviation: } \sigma = \sqrt{\lambda} = \sqrt{1.7} = 1.304$$

¿Cuál es la probabilidad de que lleguen dos o menos clientes en un minuto dado?

$$P(X \leq 2) = P(0) + P(1) + P(2) = .1827 + .3106 + .2640 = .7573.$$

¿Cuál es la probabilidad de que lleguen al menos tres clientes (el evento complementario)?

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - .7573 = .2427$$

Ejemplo: Distribución de Poisson

Un taller de reparación de motores pequeños llegan trabajos de reparación a razón de 10 por día.

a. ¿Cuál es el número promedio de trabajos que se reciben a diario en el taller?
b. ¿Cuál es la probabilidad de que no lleguen trabajos durante cualquier hora, suponiendo que el taller está abierto 8 horas al día?

Ejemplo: Distribución de Poisson

Un taller de reparación de motores pequeños llegan trabajos de reparación a razón de 10 por día.

a. ¿Cuál es el número promedio de trabajos que se reciben a diario en el taller?
b. ¿Cuál es la probabilidad de que no lleguen trabajos durante cualquier hora, suponiendo que el taller está abierto 8 horas al día?

El número promedio de trabajos recibidos por día es igual a λ = 10 trabajos por día.

Para calcular la probabilidad de que lleguen trabajos por hora, temenos que calcular la tasa de llegada por hora; es decir $\lambda = \frac{10}{8} = 1.25$ trabajos de reparación por hora. Por lo tanto

$$x = 0$$
$$\lambda = 1.25$$

$$p(x;\ \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$p(no\ hay\ llegadas\ por\ hora) = \frac{1.25^0 e^{-1.25}}{0!}$$

$$p(no\ hay\ llegadas\ por\ hora) = 0.2865$$

Ejemplo: Distribución de Poisson

Suponga que está atendiendo un centro de atención al cliente. Las llamadas de los clientes llegan a una tasa de 2.2 llamadas por minuto.

a. ¿Cuál es la probabilidad de que no haya llamadas en el próximo minuto?
b. ¿Cuál es la probabilidad de que dos o menos llamadas entren en el próximo minuto?
c. ¿Cuál es la probabilidad de al menos una llamada entre en el próximo minuto?

Ejemplo: Distribución de Poisson

a. ¿Cuál es la probabilidad de que no haya llamadas en el próximo minuto?

$$P[X=0]=\frac{\lambda^x e^{-\lambda}}{x!}=\frac{2.2^0 e^{-2.2}}{0!}=0.11 \rightarrow 11\%$$

b. ¿Cuál es la probabilidad de que dos o menos llamadas entren en el próximo minuto?

$$P[X=0]=\frac{\lambda^x e^{-\lambda}}{x!}=\frac{2.2^0 e^{-2.2}}{0!}=0.11 \rightarrow 11\%$$

$$P[X=1]=\frac{\lambda^x e^{-\lambda}}{x!}=\frac{2.2^1 e^{-2.2}}{1!}=0.24 \rightarrow 24\%$$

$$P[X \leq 2]=62\%$$

$$P[X=2]=\frac{\lambda^x e^{-\lambda}}{x!}=\frac{2.2^2 e^{-2.2}}{2!}=0.27 \rightarrow 27\%$$

c. ¿Cuál es la probabilidad de al menos una llamada entre en el próximo minuto?

$$P[X>0]=1-P[X=0]=1-0.11=0.89 \rightarrow 89\%$$

Ejemplo: Distribución de Poisson

En una clínica ambulatoria de salud mental, las cancelaciones de citas ocurren a una tasa promedio de 1.5 por día en un miércoles típico.

(a) Justifica el uso del modelo de Poisson.
(b) ¿Cuál es la probabilidad de que no ocurran cancelaciones en un miércoles en particular?
(c) ¿Una?
(d) ¿Más de dos?
(e) ¿Cinco o más?"

Ejemplo: Distribución de Poisson

(a) Justifica el uso del modelo de Poisson.

Las cancelaciones son independientes y ocurren de una en una.

También sabemos que las cancelaciones son discretas y que se nos ha dado una tasa media por unidad de tiempo.

(b) ¿Cuál es la probabilidad de que no ocurran cancelaciones en un miércoles en particular?

$$P(X=0)=\frac{(1.5)^0(e)^{-1.5}}{0!}=.2231$$

(c) ¿Una?

$$P(X=1)=\frac{(1.5)^1(e)^{-1.5}}{1!}=.3347$$

(d) ¿Más de dos?

$$P(X>2)=1-P(X\leq 2)=.1912$$

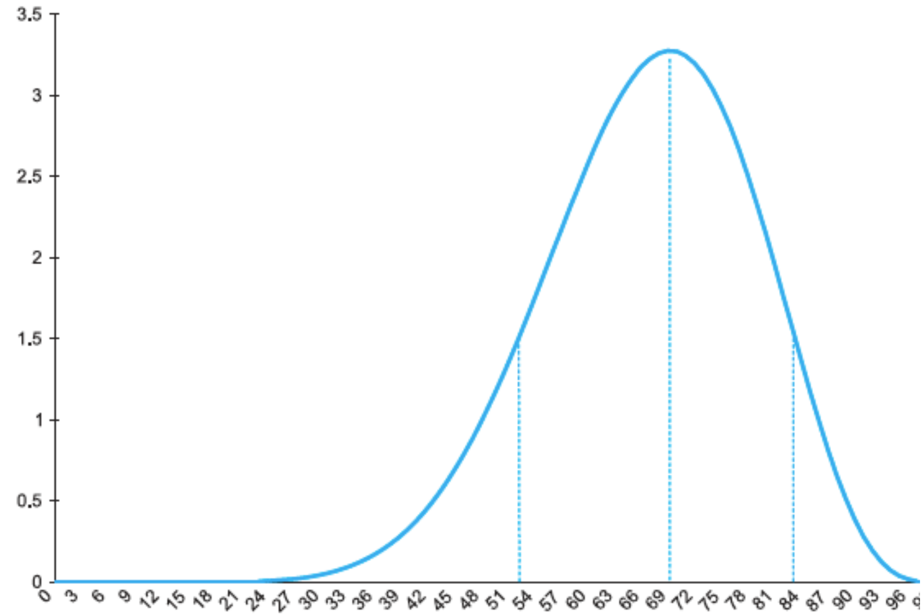
(e) ¿Cinco o más?"

$$P(X\geq 5)=1-P(X\leq 4)=.0186$$

Distribución de probabilidad continua

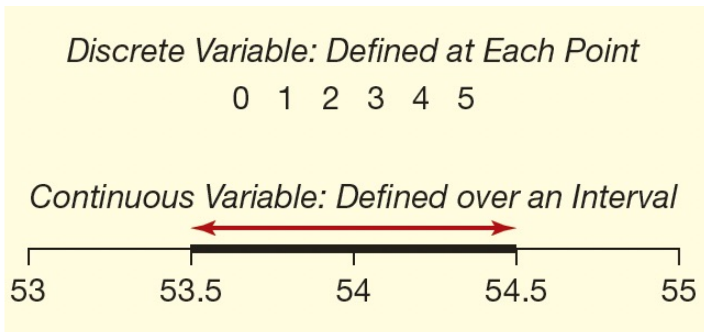
Una distribución de probabilidad en la que la variable aleatoria X puede tomar cualquier valor (es continua). Debido a que existen infinitos valores que X podría asumir, la probabilidad de que X adopte cualquier valor específico es cero.

Una **función de densidad**, generalmente denotada por f (x), especifica la distribución de probabilidad de una variable aleatoria continua X. Cuanto mayor es f (x), más probable es x. Además, el área total entre la gráfica de f (x) y el eje horizontal, que representa la probabilidad total, es igual a 1. Finalmente, f (x) no es negativo para todos los valores posibles de X.



Distribución de probabilidad continua

- La probabilidad para una variable discreta se define en un punto como $P(X = 3)$ o como una suma sobre una serie de puntos como $P(X \leq 2) = P(0) + P(1) + P(2)$.
- Cuando X es una variable continua (por ejemplo, tiempo de espera), no tiene sentido hablar de probabilidad "en" un valor particular de X (por ejemplo, X = 54 segundos) porque los valores de X no son un conjunto de puntos discretos.
- Las probabilidades para variables continuas se definen como áreas bajo una curva llamada función de densidad de probabilidad (PDF, por sus siglas en inglés).
- Las probabilidades para una variable aleatoria continua también se definen en intervalos como $P(53.5 \leq X \leq 54.5)$ o $P(X < 54)$ o $P(X \geq 53)$.

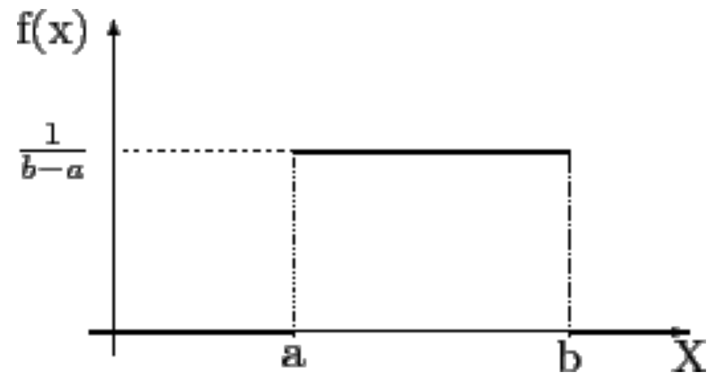


Distribución Uniforme Continua

Esta distribución se caracteriza por una función de densidad que es "plana", por lo cual la probabilidad es uniforme en un intervalo cerrado, [a, b].

La **función de densidad (pmf)** de la variable aleatoria continua X en el intervalo [a, b] es:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & en\ otro\ caso \end{cases}$$



La función de densidad forma un rectángulo con base $b - a$ y altura $\frac{1}{b-a}$

Media y mediana

$$\mu = \frac{a+b}{2}$$

Varianza

$$\sigma^2 = \frac{(b-a)^2}{12}$$

Función de distribución (cdf)

$$F(t \mid a,b) = \begin{cases} 0 & if\ t < a \\ \dfrac{t-a}{b-a} & if\ a \leq t \leq b \\ 1 & if\ t > b \end{cases}$$

Distribución Uniforme Continua

Parameters	a = lower limit b = upper limit
PDF	$f(x) = \frac{1}{b - a}$
CDF	$P(X \leq x) = \frac{x - a}{b - a}$
Domain	$a \leq x \leq b$
Mean	$\frac{a + b}{2}$
Standard deviation	$\sqrt{\frac{(b - a)^2}{12}}$
Shape	Symmetric with no mode.
Random data in Excel	=a+(b–a)*RAND()
Comments	Used as a conservative what-if benchmark and in simulation.

Ejemplo: Distribución uniforme contínua

Una empresa tiene un departamento de entrega a domicilio. Distribuyen el producto desde una locación en el centro de la ciudad hacia residencias y oficinas en la ciudad. Las entregas son realizadas en scooters y entregada directamente a cada cliente. Se registra que la distancia entre cada cliente es 2.75 km a 6.50 km. Asuma una distribución uniforme contínua

a. ¿Cuál es la distancia promedio y el coeficiente de varianza?
b. ¿Cuál es la probabilidad de que la distancia sea mayor que 5 km?
c. ¿Cuál es la probabilidad de que la distancia es +/- 1σ de μ?

Ejemplo: Distribución uniforme contínua

a. ¿Cuál es la distancia promedio, la desviación estándar y coeficiente de variación?

$$\mu = \frac{a+b}{2} = \frac{2.75+6.50}{2} = 4.625\ km$$

$$\sigma^2 = \frac{(b-a)^2}{12} = \frac{(6.50-2.75)^2}{12} = 1.1719 \rightarrow \sigma = \sqrt{1.1719} = 1.0825$$

$$C.V. = \frac{\sigma}{\mu} = \frac{1.0825}{4.625} = 0.23$$

b. ¿Cuál es la probabilidad de que la distancia sea mayor que 5 km?

$$F[X \le t] = P[X \le t] = \frac{t-a}{b-a}$$

$$P[X \ge 5] = 1 - P[X \le t]$$

$$= 1 - \frac{t-a}{b-a} = 1 - \frac{5-2.75}{6.5-2.75} = 0.4 \rightarrow 40\%$$

Ejemplo: Distribución uniforme contínua

c. ¿Cuál es la probabilidad de que la distancia es +/- 1σ de μ?

$\sigma = 1.0825$

$\mu = 4.625$

$$\mu + \sigma = 4.625 + 1.0825 = 5.7075$$
$$\mu - \sigma = 4.625 - 1.0825 = 3.5425$$

$$P[X \leq 5.7075] = \frac{5.7075 - 2.75}{6.5 - 2.75} = 0.789$$

$$P[X \geq 3.5425] = \frac{3.5425 - 2.75}{6.5 - 2.75} = 0.211$$

$$P[3.5425 \geq X \leq 5.7075] = 0.789 - 0.211 = 58\%$$

Ejemplo: Distribución uniforme contínua

Un cirujano oral inyecta un analgésico antes de extraer un diente. Dadas las características variadas de los pacientes, el dentista considera el tiempo de efectividad de la anestesia como entre 15 minutos y 30 minutos.

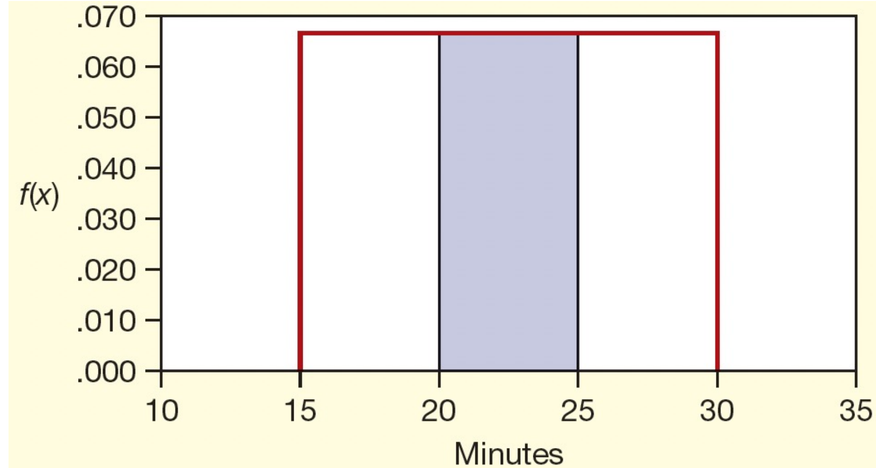
Calcule la media y la desviación estandar
Calcule la probabilidad de que el anestésico tenga una duración de entre 20 y 25 minutos.

Ejemplo: Distribución uniforme contínua

Calcule la media y la desviación estandar

X sigue una distribución uniforme U(15, 30)

a = 15, b = 30.



$$\mu = \frac{a+b}{2} = \frac{15+30}{2} = 22.5\ minutos$$

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(30-15)^2}{12}} = 4.33\ minutos$$

Calcule la probabilidad de que el anestésico tenga una duración de entre 20 y 25 minutos.

$$P(20 < X < 25) = (25 - 20)/(30 - 15) = 5/15 = 0.3333 = 33.33\%$$

Distribución Exponencial

- Esta distribución con frecuencia describe el tiempo requerido para atender a un cliente.
- La distribución exponencial es una distribución continua.
- Si los eventos por unidad de tiempo siguen una distribución de Poisson, el tiempo hasta el siguiente evento sigue una distribución exponencial.
- El tiempo hasta el próximo evento es una variable continua.

Su función de probabilidad está dada por

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

$$P(X \leq x) = 1 - e^{-\lambda x} \qquad \text{(probability of waiting } x \text{ or less)}$$

$$P(X > x) = e^{-\lambda x} \qquad \text{(probability of waiting } \textit{more} \text{ than } x\text{)}$$

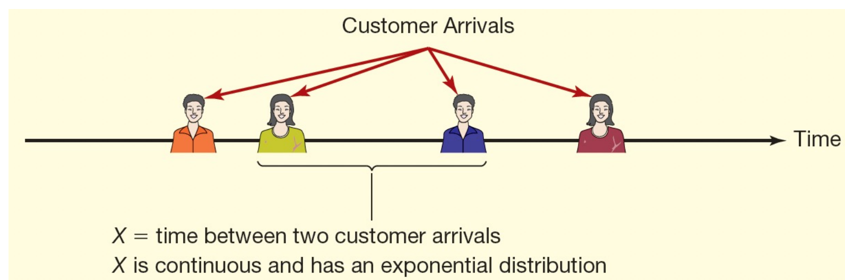
donde

$x = variable\ aleatoria\ (tiempo\ de\ servicio)$

$\lambda = tasa\ media\ de\ llegada\ por\ unidad\ de\ tiempo\ o\ espacio$

$e = 2.71828$

El promedio y desviación estándar se calcula como $= \frac{1}{\lambda}$,

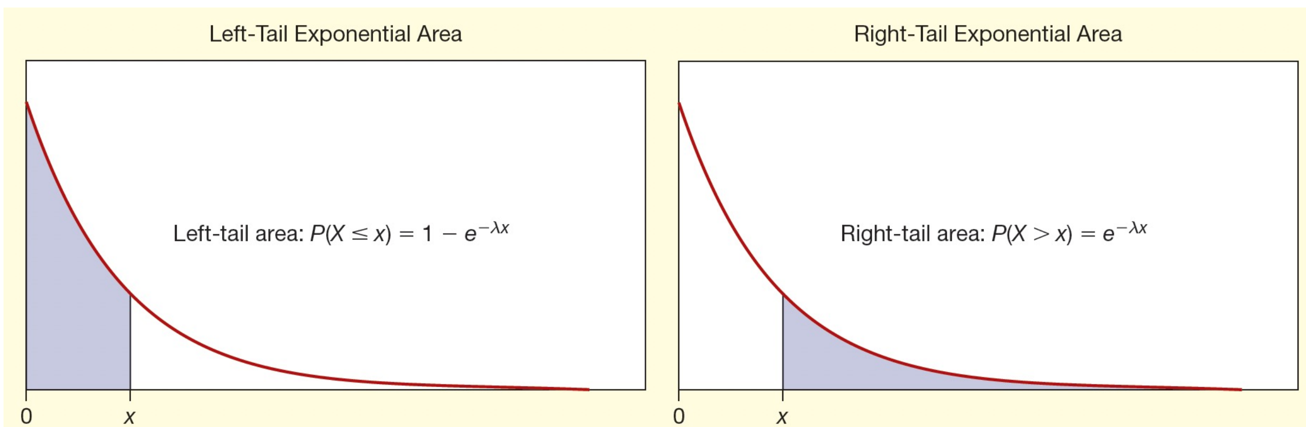


Distribución Exponencial

Parameter	λ = mean arrival rate per unit of time or space (same as Poisson mean)
PDF	$f(x) = \lambda e^{-\lambda x}$
CDF	$P(X \leq x) = 1 - e^{-\lambda x}$
Domain	$x \geq 0$
Mean	$1/\lambda$
Standard deviation	$1/\lambda$
Shape	Always right-skewed.
CDF in Excel	=EXPON.DIST(x,λ,1)
Random data in Excel	=–LN(RAND())/λ
Comments	Waiting time is exponential when arrivals follow a Poisson model. Often $1/\lambda$ is given (mean time between events) rather than λ. The value of e is approximately 2.71828.

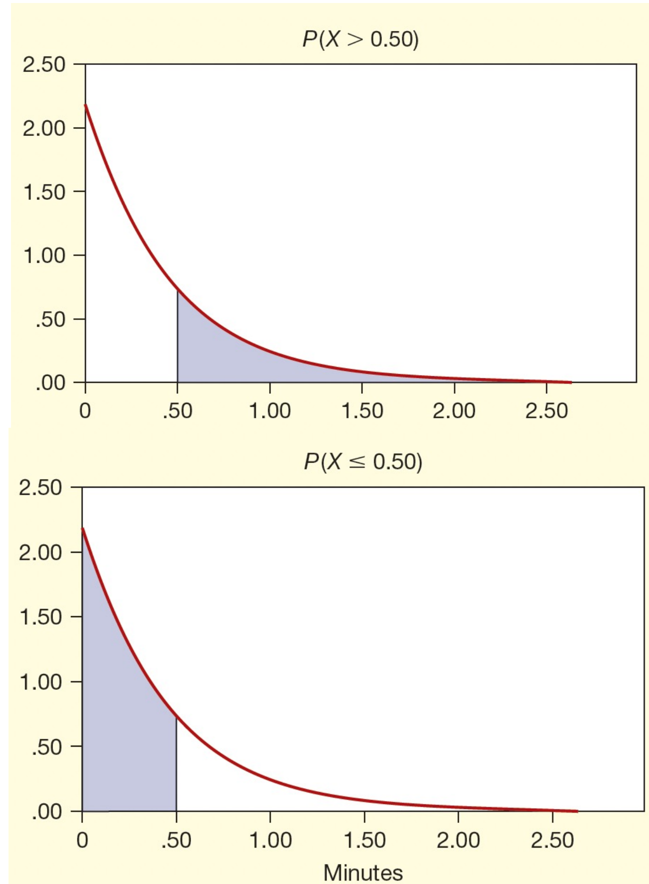
Distribución Exponencial: **características**

- La distribución exponencial tiene un parámetro, la tasa media de llegada λ.
- Cada valor de λ da una PDF diferente, como se ilustra en la figura
- Sin embargo, todas las curvas de la familia PDF exponencial tienen la misma forma.
- La intersección del eje vertical siempre es λ porque insertar x = 0 en la fórmula PDF da f(0) = λ.
- La función de densidad de probabilidad exponencial se aproxima a cero a medida que x aumenta y está sesgada hacia la derecha.
- Por lo general, no nos interesa la altura de la función f(x) sino las áreas bajo la curva.
- Afortunadamente, el CDF es simple. No se necesitan tablas, solo una calculadora que tenga la tecla de función e^x.
- La probabilidad de esperar más de x unidades de tiempo hasta la próxima llegada es e−λx, mientras que la probabilidad de esperar x unidades de tiempo o menos es $1 - e^{-\lambda x}$.



Ejemplo : Distribución Exponencial

Entre las 2 p.m. y las 4 p.m. del miércoles, las consultas de seguros de pacientes llegan a la aseguradora ASSA a una tasa media de 2.2 llamadas por minuto.
¿Cuál es la probabilidad de esperar más de 30 segundos (es decir, 0.50 minutos) para la siguiente llamada? La probabilidad de esperar menos que 30 segundos?



Establezca $\lambda = 2.2$ eventos/min y $x = 0.50$ min.

$$P(X > 0.50) = e^{(-\lambda x)} = e^{(-(2.2)(0.5))} = 0.3329$$

33.29% de probabilidad de esperar más de 30 segundos para la siguiente llamada.

La probabilidad de que se necesiten 30 segundos o menos (0.50 minutos) antes de que llegue la próxima llamada es

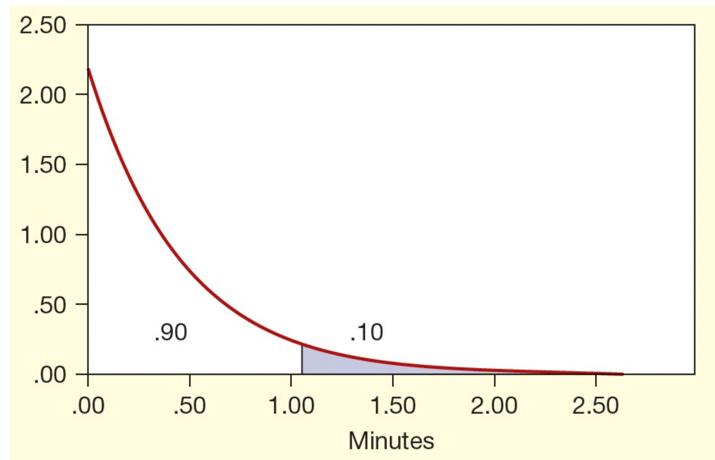
$$P(X \leq 0.50) = 1 - P(X > 0.50) = 0.6671$$

Ejemplo: Distribución Exponencial

Si la tasa media de llegadas es de 2.2 llamadas por minuto y queremos el percentil 90 para el tiempo de espera (el 10% superior del tiempo de espera).

Llame al tiempo desconocido x.
Dado que P(X ≤ x) = .90 implica que P(X > x) = .10,
establecemos el área de la cola derecha a .10, tomamos el logaritmo natural de ambos lados y resolvemos para x:

$$P(X \leq x) = 1 - e^{-\lambda x} = .90$$
$$\text{so} \quad e^{-\lambda x} = .10$$
$$-\lambda x = \ln(.10)$$
$$-(2.2)x = -2.302585$$
$$x = 2.302585/2.2$$
$$x = 1.0466 \text{ minutes}$$



Entonces el 90 por ciento de las llamadas llegarán dentro de 1.0466 minutos (o 62.8 segundos).

Tiempo medio entre eventos

- Los tiempos de espera exponenciales a menudo se describen en términos de **tiempo medio entre eventos (MTBE)** en lugar de en términos de llegadas de Poisson por unidad de tiempo.
- En otras palabras, se nos podría dar $1/\lambda$ en vez de λ.
- MTBE = $\frac{1}{\lambda}$ tiempo medio entre eventos (unidades de tiempo por evento).
- $\frac{1}{MTBE} = \lambda$ media de eventos por unidad de tiempo (eventos por unidad de tiempo).

EJEMPLO

Si el tiempo medio entre las llegadas de los pacientes a la sala de emergencias es de 20 minutos, entonces $\lambda = 1/20 = 0{,}05$ llegadas por minuto (o $\lambda = 3{,}0$ llegadas por hora).

Podríamos resolver un problema usando horas o minutos, siempre y cuando tengamos cuidado de asegurarnos de que x y λ se expresen en las mismas unidades cuando calculamos $e^{-\lambda x}$.

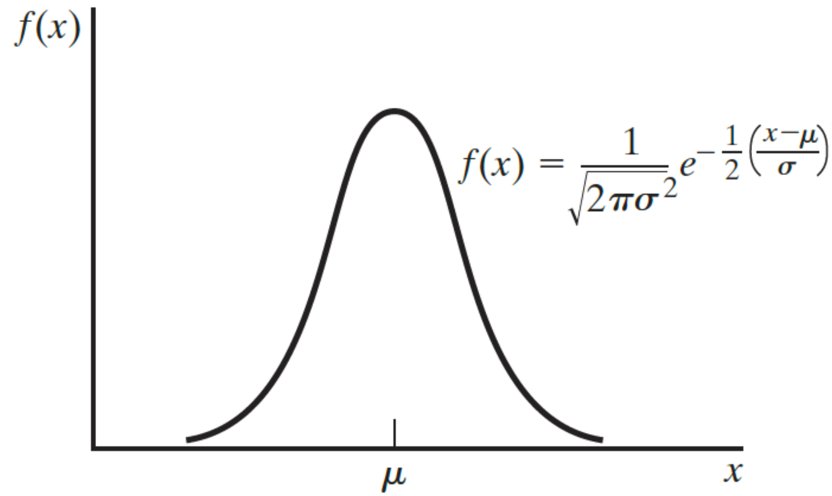
Por ejemplo, $P(X > 12 \text{ minutos}) = e^{-(0{,}05)(12)} = e^{-0{,}60}$ es lo mismo que $P(X > 0{,}20 \text{ hora}) = e^{-(3)(0{,}20)} = e^{-0{,}60}$.

Distribución Normal

Es una de las distribuciones de probabilidades continuas más populares y útiles.

La función de densidad de probabilidad está dada por la fórmula :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



- La distribución normal queda especificada por completo cuando se conocen los valores de la media, μ, y la desviación estándar, σ.
- La distribución normal es simétrica y su punto medio está en la media.
- Cambiando la media no cambia la forma de la distribución.
- Los valores en el eje X se miden en términos de cuántas desviaciones estándar están separados de la media.
- Cuando la desviación estándar se hace más grande, la curva se aplana.
- Cuando la desviación estándar se hace más pequeña, la curva es más pronunciada.

Distribución Normal: Propiedades

1. La moda, que es el punto sobre el eje horizontal donde la curva tiene su punto máximo, ocurre en $x = \mu$.
2. La curva es simétrica alrededor de un eje vertical a través de la media μ.
3. La curva tiene sus puntos de inflexión en $x = \mu \pm \sigma$, es cóncava hacia abajo si $\mu - \sigma < X < \mu + \sigma$, y es cóncava hacia arriba en otro caso.
4. La curva normal se aproxima al eje horizontal de manera asintótica, conforme nos alejamos de la media en cualquier dirección.
5. El área total bajo la curva y sobre el eje horizontal es igual a uno.

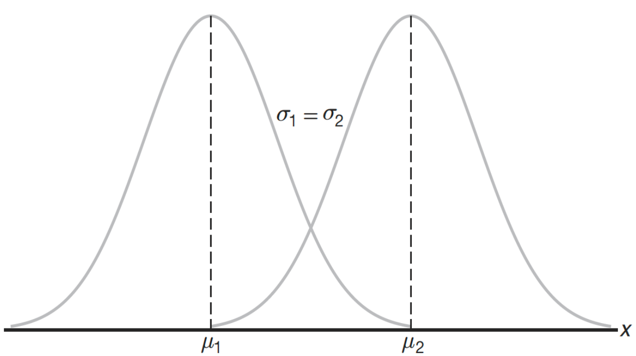


Figura 6.3: Curvas normales con $\mu_1 < \mu_2$ y $\sigma_1 = \sigma_2$.

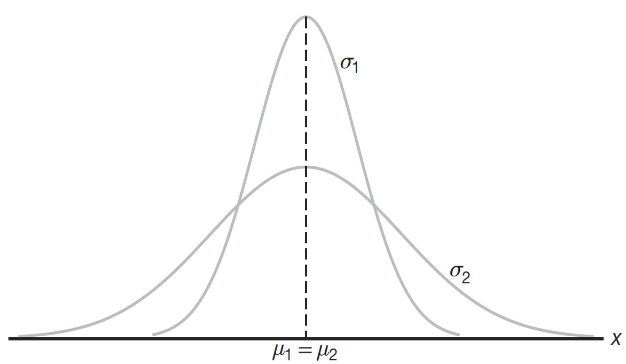


Figura 6.4: Curvas normales con $\mu_1 = \mu_2$ y $\sigma_1 < \sigma_2$.

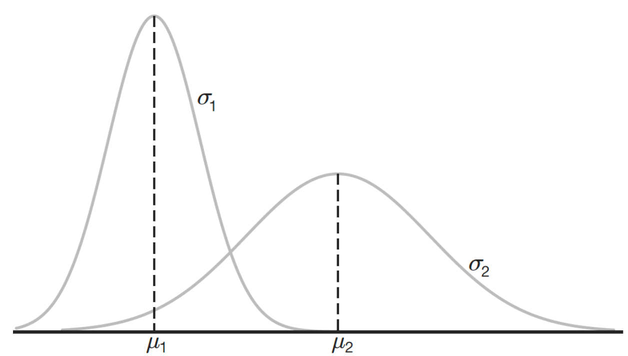


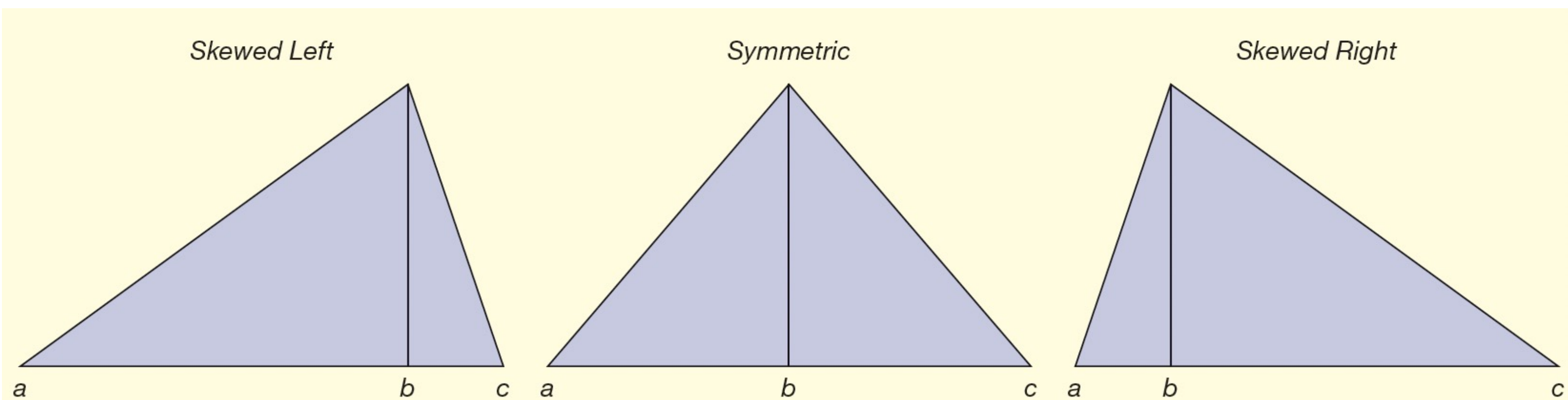
Figura 6.5: Curvas normales con $\mu_1 < \mu_2$ y $\sigma_1 < \sigma_2$.

Distribución Normal: Propiedades

Parameters	μ = population mean σ = population standard deviation
PDF	$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ where $z = \frac{x-\mu}{\sigma}$
Domain	$-\infty < z < +\infty$
Mean	0
Standard deviation	1
Shape	Symmetric, mesokurtic, and bell-shaped.
CDF in Excel*	=NORM.S.DIST(*z*,1)
Random data in Excel	=NORM.S.INV(RAND())
Comment	There is no simple formula for a normal CDF, so we need normal tables or Excel to find areas.

Distribución triangular

- La distribución triangular es una forma de pensar sobre la variación que se corresponde bastante bien con el análisis hipotético en los negocios.
- No sorprende que los analistas de negocios se sientan atraídos por el modelo triangular.
- Su rango finito y su forma simple son más comprensibles que una distribución normal..
- Es más versátil que una normal, porque se puede sesgar en cualquier dirección.
- El modelo triangular es especialmente útil para el análisis hipotético cuando el caso de negocios depende de predecir una variable estocástica (por ejemplo, el precio de una materia prima, una tasa de interés, un volumen de ventas).
- Si el analista puede anticipar el rango (a - c) y el valor más probable (b), será posible calcular las probabilidades de varios resultados.
- Muchas veces, tales distribuciones estarán sesgadas, por lo que una normal no sería de mucha ayuda.



Distribución triangular

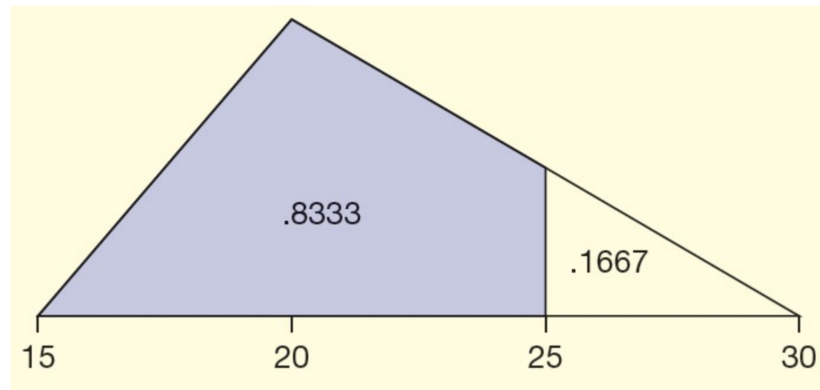
Parameters	a = lower limit b = mode c = upper limit
PDF	$f(x) = \dfrac{2(x-a)}{(b-a)(c-a)}$ for $a \le x \le b$ $f(x) = \dfrac{2(c-x)}{(c-a)(c-b)}$ for $b \le x \le c$
CDF	$P(X \le x) = \dfrac{(x-a)^2}{(b-a)(c-a)}$ for $a \le x \le b$ $P(X \le x) = 1 - \dfrac{(c-x)^2}{(c-a)(c-b)}$ for $b \le x \le c$
Domain	$a \le x \le c$
Mean	$\dfrac{a+b+c}{3}$
Standard deviation	$\sqrt{\dfrac{a^2+b^2+c^2-ab-ac-bc}{18}}$
Shape	Positively skewed if $b < (a+c)/2$. Negatively skewed if $b > (a+c)/2$.
Comments	Practical model, useful in business what-if analysis. A symmetric triangular is the sum of two identically distributed uniform variates.

Distribución triangular

Un cirujano oral inyecta un analgésico antes de extraer un diente. Dado que las características de los pacientes varían, el dentista considera que el tiempo de efectividad de la anestesia que toma entre 15 minutos y 30 minutos, siendo 20 minutos el tiempo más probable. Calcule la media, la desviación estandar y calcular la probabilidad de tardar menos de 25 minutos

La media es:

$$\mu = \frac{(a + b + c)}{3} = 21.7\ minutos$$



La desviación estándar es:

$$\sigma = \sqrt{\frac{(a^2 + b^2 + c^2) - ab - ac - bc}{18}} = 3.12 \text{ minutos}$$

Probabilidad de que el anestésico tarde menos de 25 minutos es:

$$P(X \leq 25) = 1 - \frac{(30-25)^2}{(30-15)(30-20)} = .8333$$

Pruebas para determinar si la data sigue una distribución específica:
Prueba Chi-cuadrado para bondad de ajuste

- La distribución de probabilidad de los datos históricos puede determinarse mediante las pruebas **Chi-cuadrada**, de **Kolmogorov-Smirnov** y de **Anderson-Darling**.
- La prueba de **bondad de ajuste (GOF)** ayuda decidir si la muestra se parece a un tipo particular de población.
- La prueba se basa en el nivel de ajuste que existe entre la frecuencia de ocurrencia de las observaciones en una muestra observada y las frecuencias esperadas que se obtienen a partir de la distribución hipotética.
- Las pruebas de bondad de ajuste son pruebas de hipótesis para verificar si los datos observados en una muestra aleatoria se ajustan con algún nivel de significancia a determinada distribución de probabilidad (uniforme, exponencial, normal, poisson, u otra cualquiera).
- La hipótesis nula H_o indica la distribución propuesta, mientras que la hipótesis alternativa H_1, nos indica que la variable en estudio tiene una distribución que no se ajusta a la distribución propuesta.

$$H_o: f(x) = f_o(x)$$

$$H_1: f(x) \neq f_o(x)$$

H_0: La población sigue una distribución____________.

H_1: La población no sigue una distribución__________.

El espacio en blanco puede contener el nombre de cualquier distribución teórica (por ejemplo, uniforme, Poisson, normal).

Pruebas para determinar si la data sigue una distribución específica:
Prueba Chi-cuadrado para bondad de ajuste

Una prueba de la bondad de ajuste entre las frecuencias observadas y esperadas se basa en la cantidad.

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

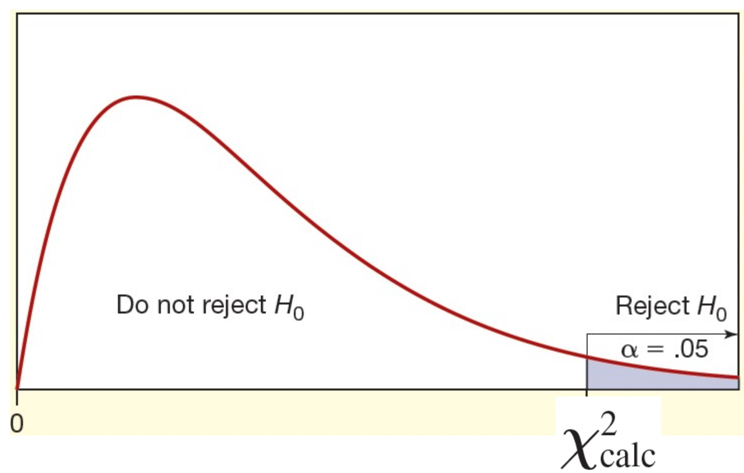
donde χ^2 es un valor de una variable aleatoria cuya distribución muestral se aproxima muy de cerca a la distribución chi cuadrada

Los símbolos

o_i = frecuencia observada
e_i = frecuencia esperada,
respectivamente, para la $i - ésima$ celda.

Si las frecuencias observadas se acercan a las frecuencias esperadas correspondientes, el valor χ 2 será pequeño, lo cual indica un buen ajuste.
Si las frecuencias observadas difieren de manera considerable de las frecuencias esperadas, el valor χ 2 será grande y el ajuste deficiente.

Un buen ajuste conduce a la aceptación de H_0, mientras que un ajuste deficiente conduce a su rechazo.
La región crítica caerá en la cola derecha de la distribución chi cuadrada.



Prueba Chi-cuadrado para bondad de ajuste: Distribución Poisson

- En un modelo de distribución de Poisson, X representa el número de eventos por unidad de tiempo o espacio.
- X es un entero discreto no negativo (X = 0, 1, 2, …)
- Las llegadas de eventos deben ser independientes entre sí.
- A veces se le llama modelo de eventos raros porque X normalmente tiene una media pequeña.
- La media λ es el único parámetro. Los pasos iniciales para la prueba son:
  - Paso 1: Calcule la frecuencia observada para cada uno de los valores de x.
  - Paso 2: Si se desconoce λ, estímelo a partir de la muestra.
  - Paso 3: Use el λ estimado para encontrar la probabilidad de Poisson P(X) para cada valor de X.
  - Paso 4: Multiplique P(X = x) por el tamaño de la muestra n para obtener las frecuencias esperadas.
  - Paso 5: Realice los cálculos de chi-cuadrado.
  - Paso 6: Toma la decisión.

Ejemplo: prueba Chi-cuadrado para Distribución Poisson

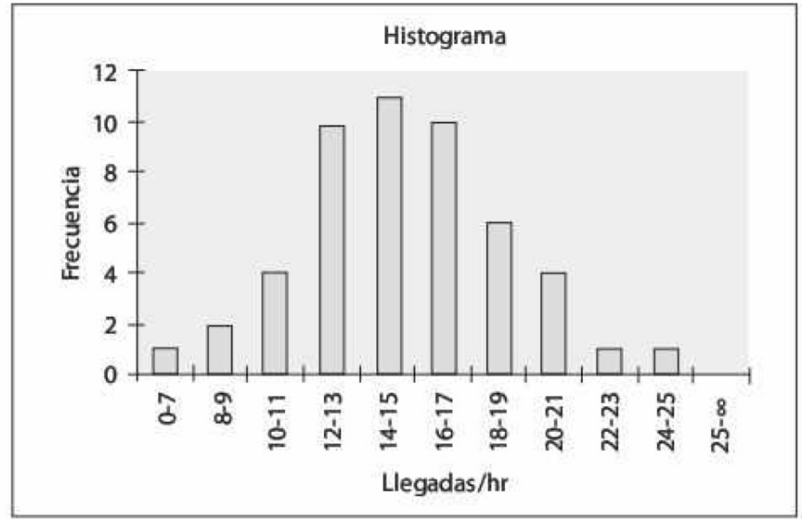
Dado los siguientes datos sobre el número de autos que entran a una gasolinera cada hora. Determine si la distribución es tipo Poisson utilizando la prueba Chi-cuadrado con un nivel de significancia del 5%

14	7	13	16	16	13	14	17	15	16
13	15	10	15	16	14	12	17	14	12
13	20	8	17	19	11	12	17	9	18
20	10	18	15	13	16	24	18	16	18
12	14	20	15	10	13	21	23	15	18

Ejemplo: prueba Chi-cuadrado para Distribución Poisson

El histograma de los n = 50 datos, que considera m = 11 intervalos, la media muestral de 15.04 y la varianza muestral de 13.14, permite establecer la siguiente hipótesis:

H_o: Poisson (λ = 15) automóviles/hora
H_1: Otra distribución



Comenzamos por calcular la probabilidad de cada intervalo a partir de la función de probabilidad de Poisson:

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0,1,2,\ldots$$

$$p(x) = \frac{15^x e^{-15}}{x!} \qquad x = 0,1,2,\ldots$$

Enseguida calculamos la frecuencia esperada en cada intervalo, multiplicando la probabilidad p(x) por el total de datos de la muestra:

$$E_i = np(x)$$
$$E_i = 50\,p(x)$$

Ejemplo: prueba Chi-cuadrado para Distribución Poisson

Y luego estimamos el estadístico de prueba:

$$\chi_o^2 = \sum_{i=1}^{m} \frac{(E_i - O_i)^2}{E_i} = \frac{(0.9001-1)^2}{0.9001} + \frac{(2.5926-2)^2}{2.5926} + \ldots + \frac{(0.3092-0)^2}{0.3092} = 1.7848$$

Intervalo	O_i	$p(x)$	$E_i = 50 * p(x)$	Error
0-7	1	0.0180	0.9001	0.0111
8-9	2	0.0519	2.5926	0.1354
10-11	4	0.1149	5.7449	0.5300
12-13	10	0.1785	8.9233	0.1299
14-15	11	0.2049	10.2436	0.0559
16-17	10	0.1808	9.0385	0.1023
18-19	6	0.1264	6.3180	0.0160
20-21	4	0.0717	3.5837	0.0483
22-23	1	0.0336	1.6821	0.2766
24-25	1	0.0133	0.6640	0.1700
25-8	0	0.0062	0.3092	0.3092
Total	50	1	50	1.78481

El valor del estadístico de prueba, $\chi_o^2 = 1.7848$, comparado con el valor de tablas crítico, $\chi^2_{0.05,11-0-1} = 18.307$, indica que no podemos rechazar la hipótesis de que la variable aleatoria se comporta de acuerdo con una distribución de Poisson, con una media de 15 automóviles/hora.

Algunas funciones de Excel para estadísticas

Función	Fórmula
Maximo	=MAX()
Mínimo	=MIN()
Mediana	=MEDIAN()
Moda	=MODE()
Media	=AVERAGE()
Desviación Media Absoluta (MAD)	=AVEDEV(Data)
Percentil	=PERCENTIL.INC()
Varianza poblacional	=VAR.P()
Varianza muestral	=VAR.S()
Valor estandarizado Z	=STANDARDIZE(XValue, Mean, StDev)
Desviación estándar poblacional	=STDEV.P()
Desviación estándar muestral	STDEV.S()
Distribución de Poisson	=POISSON.DIST(x, lambda, cumulativa)
Distribución exponencial	=EXPON.DIST(x, lambda, cumulativa)

Libros de referencia

- Render, B. (2016). Métodos cuantitativos para los Negocios. Editorial Pearson.
- Gohout, W. (2013). Operations Research. Oldenbourg Verlag München
- Taha, H. (2011). Investigación de Operaciones. Editorial Pearson.
- Hillier, F. & Lieberman, G. (2015). Investigación de Operaciones. McGraw-Hill
- Winston, W. (2004). Operations Research Applications and Algorithms. Thomson Brooks/Cole
- Anderson, D. & Sweeny, D. (2019). Métodos Cuantitativos para los Negocios. Cengage
- Eppen, D. (2000). Investigación de Operaciones en la Ciencia Administrativa. Pearson
- García et al. (2013). Simulación y Análisis de Sistemas con ProModel. Editorial Pearson.
- Srinivasan, G. (2010). Quantitative Models in Operations and Supply Chain Management. PHI Learning Private Limited
- Rardin, R. (2017). Optimization in Operations Research. Pearson
- Carter, M. et al. (2019). Operations Research A Practical Introduction. Taylor & Francis Group
- Aoroto Álvares, C., [et al] (2014) Operations research in business administration and managenent. Valencia: Universitat Politècnica de València
- Ravi Ravindran, A. (2008) Operations Research & Management Science Handbook. Taylor & Francis Group
- Rees, M. (2015). Business Risk and Simulation Modeling in Practice. John Wiley & Sons Ltd
- Sterman, J. (2000). Business Dynamics – Systems Thinking and Modeling for a Complex World. McGraw-Hill
- Winston, W. (2017) Microsoft Excel 2016 – Data Analysis and Busines Modeling. Microsoft press
- Schaffernicht, M. (2006). *Dinámica de Sistemas – Tomo 1: Fundamentos*.
- Alvarez, H. (2011). Introducción a la Simulación. Universidad Tecnológica de Panamá
- Cassandras & Lafortune. (2008). *Introduction to Discrete Event Systems*. Springer
- Bandyopadhyay, S. et al (2014). Discrete and Continuous Simulation: Theory and Practice. Taylor & Francis Group
- García et al. (2013). *Simulación y Análisis de Sistemas con ProModel*. Editorial Pearson.
- Schroeder et al. (2011). *Administración de Operaciones*. McGraw-Hill
- Render, B. & Heizer, J. (2014). *Principios de Administración de Operaciones*. Pearson
- Chase, R. & Jacobs, F. (2014). *Administración de Operaciones, Producción y Cadena de Suministro.* McGraw – Hill
- Rees, M. (2015). *Business Risk and Simulation Modeling in Practice*. John Wiley & Sons Ltd
- Sterman, J. (2000). *Business Dynamics – Systems Thinking and Modeling for a Complex World*. McGraw-Hill
- Winston, W. (2017) *Microsoft Excel 2016 – Data Analysis and Busines Modeling*. Microsoft press
- Slack, N., et al. (2016) . *Operations Management*. Pearson
- Stevenson, W. (2015). *Operations Management*. McGraw-Hill

Contacto

Ricardo Caballero, M.Sc.
Docente Tiempo Completo
Facultad de Ingeniería Industrial
Universidad Tecnológica de Panamá | Centro Regional de Chiriquí

E-Mail: ricardo.caballero@utp.ac.pa
Social: LinkedIn | ResearchGate
Website: https://www.academia.utp.ac.pa/ricardo-caballero

Project Manager
Giii Grupo de Investigación en Ingeniería Industrial

Website: www.giii.utp.ac.pa