Curso introductorio a Estadística

Diego Velázquez

Instituto Nacional de Medicina Genómica

22 de abril de 2023

Presentación

Mi nombre es Diego Velázquez, estudié la licenciatura en Matemáticas Aplicadas (ITAM) y la licenciatura de Ciencias de la Computación (Fac Ciencias, UNAM). Soy apasionado del modelaje matemático, análisis de datos, inteligencia artificial, algoritmos genéticos, métodos de simulación MCMC y de la Estadística Bayesiana.

Datos de contacto:

- Correo institucional: **velazquez@ciencias.unam.mx**
- Correo personal: **d.velazquez.mc@gmail.com**
- Github: **DiegoArturoVelazquezTrejo**

Notas sobre el curso

- Teórico-práctico en lenguaje de programación R,
- Repositorio del curso: https://github.com/DiegoArturoVelazquezTrejo/Estadistica-1-INMEGEN-2023,
- Días jueves 10:00 a.m. a 11:00 a.m.,
- Requisitos: crearse una cuenta de GitHub `https://github.com/` y/o RStudioCloud `https://posit.cloud/`,
- Herramientas en R: plotly, ggplot2, y más ...
- Última sesión: lectura de artículo científico con Dr. Pablo Oliva.

Introducción: Github

GitHub es una plataforma para gestionar y compartir código de software. Ayuda a los desarrolladores a colaborar, rastrear cambios y trabajar en equipo en proyectos de programación. Usaremos Github como repositorio para toda la parte práctica del curso. Por otro lado, nos aprovecharemos de codespaces para la ejecución del código.

Figure: Logo de GitHub

Introducción: RStudio Cloud

Posit Cloud de R es una plataforma para el manejo y análisis de datos en la nube. Ayuda a los profesionales a almacenar, procesar y trabajar en equipo con datos, permitiendo análisis y colaboración eficiente. Trabajar en la nube nos ahorra el tiempo y problemas de instalación en nuestros equipos de R y del entorno necesario para trabajar.

Figure: Logo de GitHub

Introducción: ¿Qué es la estadística?

Supóngase que se desea diseñar:

- Una encuesta para recabar los primeros informes en un día de elecciones con el fin de pronosticar los resultados.
- Muestreos de consumidores para predecir preferencias hacia ciertos productos.
- Experimentos para determinar el efecto de diversos medicamentos y condiciones ambientales controladas en seres humanos para inferir el tratamiento adecuado.

Introducción: Definiciones

Estadística: "Rama de las matemáticas que estudia la recolección, análisis, interpretación y presentación de masas de información numérica" (**Webster 's New Collegiate Dictionary**).

Estadística: "La tecnología del método científico que se ocupa del diseño de experimentos e investigaciones, y de la inferencia estadística" (**Mood, Graybill y Boes 1974**).

Meta de la estadística: "La meta de la estadística es hacer una inferencia acerca de una población, con base en información contenida en una muestra de esa población y dar una medida de bondad asociada para la inferencia" (**Wackerly, Mendenhall, Scheaffer 2009**).

Introducción: Algunas nociones básicas

- La colección de toda la posible información que caracteriza a un fenómeno se denomina **población**, y el subconjunto seleccionado de ella es una **muestra**,
- La información útil para inferir alguna característica de una población,
- **Por ejemplo**: las mediciones en los pacientes para un experimento médico representan una muestra de una población conceptual formada por todos los pacientes que de modo semejante sufren hoy, así como los que estarán afectados en el futuro próximo.

Introducción: Teoría y realidad

- Las teorías son conjeturas propuestas para explicar fenómenos del mundo real; *aproximaciones* o modelos de la realidad.
- Cuando escogemos un modelo matemático para un proceso físico, esperamos que el modelo refleje fielmente, en términos matemáticos, los *atributos del proceso* físico.
- El proceso de hallar una buena ecuación no es necesariamente sencillo y por lo general requiere de varias *suposiciones o simplificaciones*.

Conceptos Básicos

- *Variable aleatoria*: es una cantidad numérica que toma diferentes valores en función de los resultados de un experimento o proceso aleatorio *(matemáticamente es una función de valor real para la cual el dominio es un espacio muestral)*.
- *Variable dependiente*: variable que se mide u observa para evaluar cómo cambia en respuesta a la manipulación o cambio de otra variable.
- *Variable independiente*: variable que se manipula o cambia deliberadamente para observar su efecto sobre otra variable (la dependiente).
- *Distribución*: describe cómo se agrupan o extienden los valores numéricos de una variable en un conjunto de observaciones.

Introducción: Tipos de variables

- *Variables categóricas nominales*: representa estados/categorías (categorías mutuamente excluyentes, no hay orden inherente o numérico).
- *Variables categóricas ordinales*: representa estados/categorías y que poseen un orden inherente (relación jerárquica). Ejemplos: niveles de satisfacción, clasificación educativa, escala de dolor.
- *Variables discretas*: variable que puede tomar sólo un número finito o contablemente infinito de valores.
- *Variables continuas*: cualquier variable que puede tomar un valor dentro de un rango continuo infinito.

Introducción: transformación de variables categóricas en R

En R se tiene que transformar las variables categóricas a tipo factor para que el lenguaje identifique la naturaleza categórica de la variable. Para variables categóricas nominales basta con: Repositorio del curso:

https://github.com/DiegoArturoVelazquezTrejo/Estadistica-1-INMEGEN-2023

Introducción: frecuencias

1. **Frecuencias absolutas**: número de veces que un valor específico o un evento ocurre en un conjunto de datos.
2. **Frecuencias relativas**: proporción o fracción de veces que un valor específico o un evento ocurre en relación con el total de observaciones en un conjunto de datos.

Histograma de frecuencia relativa: representa la distribución de frecuencias relativas. Consiste en subdividir el eje de medición en intervalos de igual ancho. Se construye un rectángulo sobre cada intervalo, de modo que la altura del rectángulo sea proporcional a la fracción del número total de mediciones que caen en cada intervalo.

¿Cómo construirlo en R? https://github.com/DiegoArturoVelazquezTrejo/Estadistica-1-INMEGEN-2023

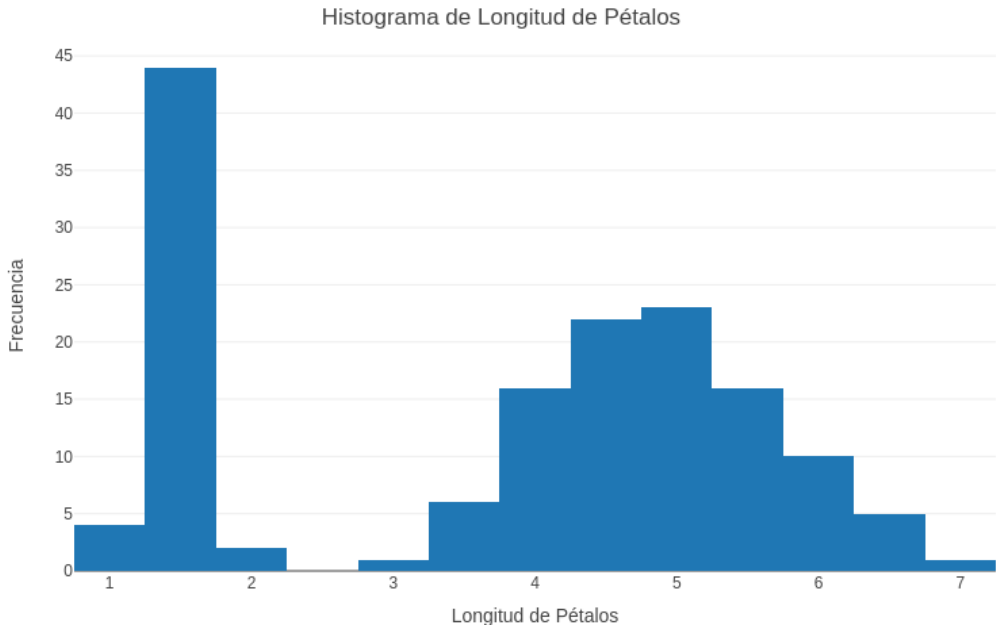

Figure: Ejemplo de un histograma

Introducción: distribución de una variable

¿Por qué es necesario conocer la distribución de una variable?

- Comprensión de los datos,
- Inferencia estadística (aplicar pruebas de hipótesis, intervalos de confianza, métodos paramétricos y no paramétricos),
- Selección de modelos (la elección del modelo adecuado a menudo depende de la distribución de la variable),
- Identificación de outliers.

Caracterización de un conjunto de mediciones

Buscamos algunos números que tienen interpretaciones significativas y que se pueden usar para describir la distribución de frecuencias de cualquier conjunto de mediciones.

Definición: Media muestral

La media de una muestra de n respuestas medidas $y_1, y_2, \ldots, y_n$ está dada por:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Por lo general no podemos medir el valor de la media poblacional, μ; más bien, μ es una constante desconocida que podemos estimar usando información muestral. En R se calcula usando la función *mean()*.

Definición: Varianza muestral

La **varianza de una muestra** de mediciones $y_1, y_2, \ldots, y_n$ es la suma del cuadrado de las diferencias entre las mediciones y su media, dividida entre $n-1$. Simbólicamente, la varianza muestral es:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

¿Por qué se divide entre $n-1$ en lugar de n? Se puede demostrar que el estimador de la varianza muestral con $n-1$ es insesgado y da una mejor estimación de la verdadera varianza poblacional, σ^2. En R se puede estimar usando el comando `var(datos)`.

Definición: Desviación estándar

La desviación estándar de una muestra de mediciones es la raíz cuadrada positiva de la varianza, esto es,

$$s = \sqrt{s^2}$$

Por lo que la correspondiente desviación estándar poblacional está denotada por $\sigma = \sqrt{\sigma^2}$. La desviación estándar se puede usar para dar una imagen más o menos precisa de la variación de datos para un conjunto de mediciones. De ahí sale la siguiente regla empírica:

Regla empírica

Para una distribución de mediciones que sea aproximadamente normal (forma de campana/ distribución gaussiana), se deduce que el intervalo con puntos extremos:

$\mu \pm \sigma$ contiene aproximadamente el 68% de las mediciones,

$\mu \pm 2\sigma$ contiene aproximadamente el 95% de las mediciones,

$\mu \pm 3\sigma$ contiene aproximadamente todas las mediciones.

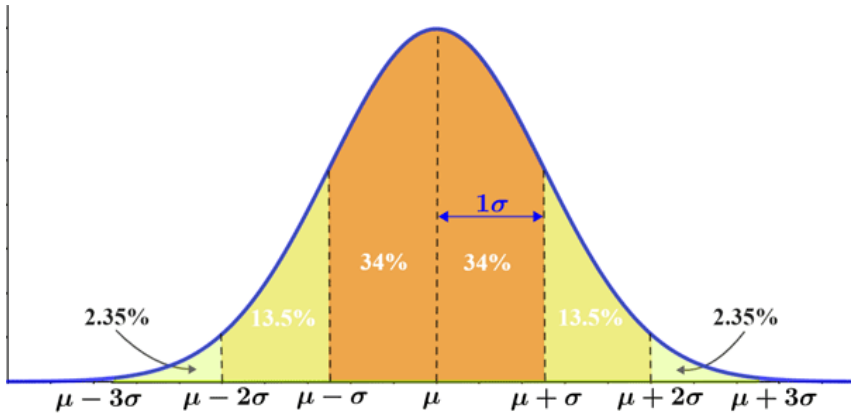

Figure: Regla empírica para distribuciones normales

Introducción: Recursos gráficos para la presentación de datos

Existen varios recursos básicos para la presentación de datos, entre ellos:

- **Gráficos de barra**: frecuencia o proporción de diferentes categorías mediante barras rectangulares. Útiles para comparar valores entre diferentes grupos.
- **Histogramas**: para estudiar la distribución de datos continuos.
- **Gráficas de líneas**: conectan puntos de datos con líneas, mostrando la evolución o tendencia de una variable a lo largo del tiempo.

- **Gráficos circulares**: muestran proporciones de diferentes partes de un todo como sectores de un círculo. A menudo se utilizan para resaltar la composición porcentual.
- **Gráficos de dispersión**: representan pares de datos en un plano, permitiendo observar la relación entre dos variables y detectar posibles patrones o correlaciones.
- **Box-plots**: muestran la distribución de los datos que incluye cuartiles, mediana, valores atípicos y rango intercuartílico.
- **Gráficos de burbujas**: variante de los gráficos de dispersión, donde se agrega un tercer valor para determinar el tamaño de las burbujas, visualizando información adicional.

¿Cómo construirlo en R? https://github.com/DiegoArturoVelazquezTrejo/Estadistica-1-INMEGEN-2023

Recapitulación: Variables Aleatoria

DiegoArturoVelazquezTrejo github

Variable Aleatoria Discreta: Una variable aleatoria discreta es aquella que puede tomar un conjunto finito o numerable de valores distintos. Cada valor tiene una probabilidad asociada.
Ejemplo: El resultado del lanzamiento de un dado.

Variable Aleatoria Continua: Una variable aleatoria continua es aquella que puede tomar cualquier valor en un intervalo continuo. Se describe mediante una función de densidad de probabilidad.
Ejemplo: La altura de las personas.

Definición: Esperanza Teórica de una Variable Aleatoria

Esperanza Teórica de una Variable Aleatoria: La esperanza teórica (o valor esperado) de una variable aleatoria es un concepto que representa el valor medio ponderado de los posibles resultados de una variable aleatoria.

Para una variable discreta X con función de probabilidad $P(X = x_i)$ y valores x_i, la esperanza teórica $\mathbb{E}(X)$ se calcula como:

$$\mathbb{E}(X) = \sum_i x_i \cdot P(X = x_i)$$

La Esperanza Teórica

La Esperanza Teórica en concepto entendible:
La esperanza teórica de una variable aleatoria es como el valor promedio esperado de los posibles resultados. Imagina que lanzas un dado muchas veces y anotas los resultados. La esperanza es como el valor que esperarías obtener en promedio después de muchas tiradas. Para variables que pueden tomar diferentes valores con diferentes probabilidades (como números en un dado), la esperanza considera esos valores y sus probabilidades para calcular ese valor promedio.

Definición de Proporción

La proporción es una medida estadística que se refiere a la relación entre una cantidad particular y el total al que pertenece. Se utiliza para describir la fracción o porcentaje de un cierto evento o característica en una población o muestra.

$$\text{Proporción} = \frac{\text{Número de casos que cumplen la condición}}{\text{Total de casos observados}}$$

Las proporciones son especialmente útiles cuando se trabaja con variables categóricas o eventos binarios, donde se desea expresar la frecuencia relativa de un resultado particular en relación con el total de observaciones.

Definición: Variable de Bernoulli

Variable de Bernoulli con Ejemplos: Una variable de Bernoulli es un tipo de variable aleatoria discreta que toma solo dos valores posibles: 0 o 1. Estos valores representan dos resultados mutuamente excluyentes de un experimento aleatorio.

Ejemplo en Biología: Consideremos un experimento donde registramos si una especie de ave sí anida (1) o no (0) en un determinado sitio.

Ejemplo en Medicina: En un estudio clínico, registramos si un paciente **responde positivamente** a un tratamiento (1) o **no responde**(0). Lo mismo para el caso si contrae una enfermedad o no.

Variable de Bernoulli: Características y Función de Probabilidad

La variable de Bernoulli está caracterizada por un único parámetro p, que es la probabilidad de que ocurra el evento asociado al valor 1.

Función de probabilidad:

Dos posibles valores

$$P(X = x) = \begin{cases} p, & \text{si } x = 1 \\ 1 - p, & \text{si } x = 0 \end{cases}$$

$p + (1 - p) = 1$

Estimador para el Valor p

En muchas situaciones prácticas, no conocemos el valor real de la probabilidad p asociada a una variable de Bernoulli. Sin embargo, podemos usar los datos de una muestra para estimar este valor.

Un estimador común para p es la proporción de éxitos en la muestra, denotado como $\hat{p}$. Si tenemos una muestra de tamaño n y observamos x éxitos, entonces el estimador $\hat{p}$ se calcula como:

$$\hat{p} = \frac{x}{n}$$

$\hat{p} = \frac{80}{100} = 0.8$

$1 - \hat{p} = 1 - 0.8 = 0.2$

Trat. 100

Sí 80 pos

20 neg

Este estimador nos da una idea de cómo se comporta la probabilidad p en la población basándonos en los resultados observados en la muestra.

Ejemplo Médico: Resultados de una Prueba Médica

```r
# Cantidad de pacientes
n <- 200

# Simular n resultados de una prueba m dica
# (positivo = 1, negativo = 0)
resultados <- sample(c(0, 1), n,
replace = TRUE, prob = c(0.8, 0.2))

# Calcular la proporcion de resultados positivos
proporcion_positivos <- sum(resultados) / n
```

resultados = [0,1,0, 1,1, 0,1]

Sum(1 1 1) = 0+1+0+1+1+1+0+1 = 4

4/20

Variable Binomial: Introducción

$$f(x) = \binom{n}{x} p^{x} (1-p)^{n-x}$$

coef de combinación — éxito — fracaso

Variable Binomial

La variable binomial es un tipo de variable aleatoria que modela el número de éxitos en una serie de ensayos independientes e idénticos, donde cada ensayo tiene dos posibles resultados: éxito o fracaso.

En otras palabras, la variable binomial es utilizada cuando estamos interesados en contar cuántas veces ocurre un evento de interés (éxito) en un número fijo de ensayos, donde la probabilidad de éxito en cada ensayo es constante.

$\hat{p} = 0.6$

$n = 100$

¿Cuál es la prob. de que hayan exactam. 27 personas que respondieron positiv.?

Variable Binomial: Ejemplo en Medicina

Ejemplo Aplicado a la Medicina

Veamos un ejemplo en el ámbito médico: supongamos que estamos realizando un estudio clínico para evaluar la eficacia de un nuevo medicamento para tratar una enfermedad. Queremos determinar cuántos pacientes de un grupo de 50 experimentarán mejoría significativa con el medicamento.

Cada paciente puede experimentar mejoría (éxito) o no (fracaso) de manera independiente. Aquí, la variable binomial puede utilizarse para modelar el número de pacientes que experimentan mejoría en el grupo después de un cierto número de días de tratamiento.

Variable Binomial: Distribución, Media y Esperanza

Distribución Binomial

Una variable binomial X con parámetros n (número de ensayos) y p (probabilidad de éxito en un ensayo) sigue una distribución binomial. Su función de probabilidad es:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

donde $\binom{n}{k}$ es el coeficiente binomial.

Media y Esperanza

La media (μ) y la esperanza ($E[X]$) de una variable binomial son iguales y se calculan como: $n = 100, \quad p = 0.6$

$$\mu = E[X] = np \qquad \mu = 100 \times 0.6 = 60$$

Estos valores representan el número esperado de éxitos en n ensayos, dado p.

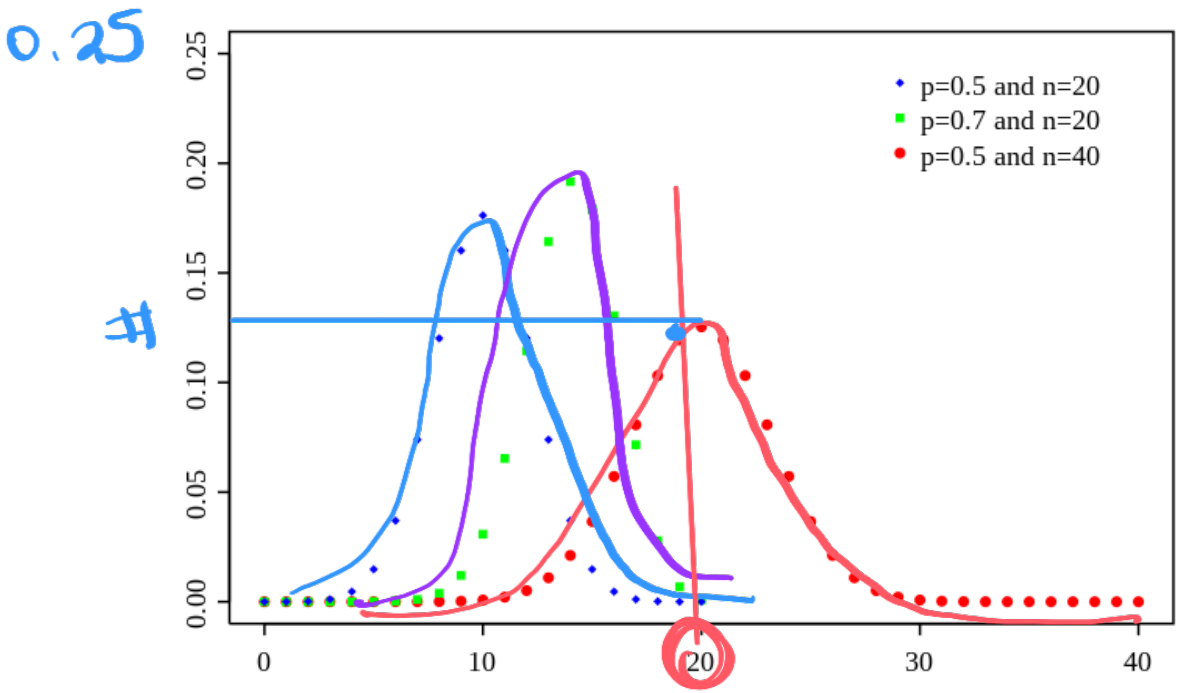

Figure: Distribución Binomial

Ejemplo: Distribución Binomial y Histograma (Parte 1)

Ejemplo: Distribución Binomial y Histograma
Supongamos que estamos realizando un estudio clínico en el que se administra un tratamiento a un grupo de pacientes. La probabilidad de que un paciente mejore con el tratamiento es $p = 0.3$, y el estudio se realiza con $n = 20$ pacientes.

Utilizando una distribución binomial, podemos simular el número de pacientes que mejoran en el grupo.

Ejemplo: Distribución Binomial y Histograma (Parte 2)

Ejemplo práctico

```
# Cargar la librer a plotly
library(plotly)

# Par metros de la distribuci n binomial
n <- 20  # N mero de ensayos
p <- 0.3 # Probabilidad de  xito  en un ensayo

# Generar datos a partir de una distribuci n binomial
datos_binom <- rbinom(1000, n, p)

# Crear un histograma interactivo con Plotly
histograma <- plot_ly(x = datos_binom, type = "histogram",
                      histnorm = "probability",
                      marker = list(color = "rgba(100, 149, 237, 0.7)"),
                      opacity = 0.7) %>%
    layout(title = "Distribuci n Binomial",
           xaxis = list(title = "N mero de  xitos "),
           yaxis = list(title = "Probabilidad"))

histograma
```

Distribución Poisson

La distribución Poisson es utilizada para modelar el número de eventos que ocurren en un intervalo de tiempo o espacio dado, cuando estos eventos son raros y aleatorios. Estos eventos deben ser mutuamente excluyentes.
Está caracterizada por un único parámetro λ, que representa el número medio de eventos en el intervalo.

$$P(X = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}$$

donde:

- X es la variable aleatoria que representa el número de eventos.
- k es el valor específico de eventos.
- e es la base del logaritmo natural (aproximadamente 2.71828).
- λ es la tasa de ocurrencia esperada de eventos.
- $k!$ es el factorial de k.

Distribución Poisson

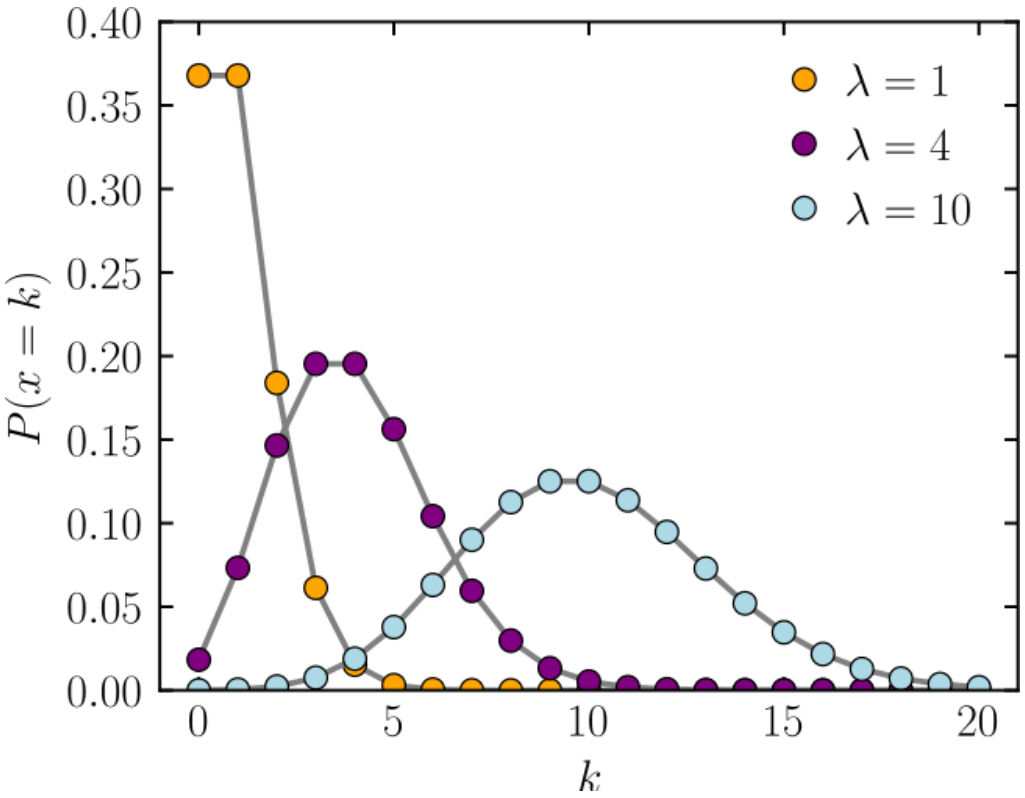

Figure: Distribución Poisson

Ejemplo de Distribución Poisson: Contexto Biológico

Ejemplo: Frecuencia de Mutaciones en un Gen
En un estudio genético, se analiza la frecuencia de mutaciones en un gen particular en una población de organismos. Se sabe que la mutación ocurre en promedio 2 veces por día en esta población.

Utilizando la distribución Poisson, podemos modelar la probabilidad de que ocurran diferentes cantidades de mutaciones en un día.

Supongamos que estamos interesados en la probabilidad de que haya exactamente 3 mutaciones en un día. Podemos aplicar la fórmula de la distribución Poisson:

$\lambda = 2$

$k = 3$

$$P(X = 3) = \frac{e^{-2} \cdot 2^3}{3!}$$

$$\frac{e^{-\lambda} \cdot \lambda^k}{k!}$$

Ejemplo de Distribución Poisson en R

```
# Generar datos simulados de una distribución Poisson
lambda <- 2
datos_poisson <- rpois(1000, lambda)

histograma <- plot_ly(x = datos_poisson, type = "histogram",
                      histnorm = "probability",
                      xbins = list(size = 0.5),
                      colors = "blue",
                      marker = list(line = list(color = "black",
             layout(title = "Histograma de Distribución Poisson",
                    xaxis = list(title = "Valores"),
                    yaxis = list(title = "Probabilidad"))
histograma

# Calcular la probabilidad de exactamente 3 mutaciones
prob_3_mutaciones <- dpois(3, lambda)
```

Ejemplos de Identificación de Tipo de Variable (Parte 1)

- Estás estudiando la cantidad de veces que una persona llama al médico en un [redacted]. ¿Qué tipo de variable se relaciona con esta situación? Poisson
- Realizas una encuesta en la que preguntas a las personas si han recibido la vacuna contra la gripe en el último año. Registras "Sí" o "No" para cada persona. ¿Qué tipo de variable representa esta encuesta? Bernoulli

Ejemplos de Identificación de Tipo de Variable (Parte 2)

- En un estudio de ensayos clínicos, estás contando la cantidad de pacientes que experimentan efectos secundarios graves después de recibir un nuevo medicamento. ¿Qué tipo de variable estás observando? Binomial
- Analizas la cantidad de bacterias por mililitro en una muestra de agua potable para determinar la concentración promedio. ¿Qué tipo de variable se relaciona con esta medición? Poisson

Estimadores de Máxima Verosimilitud para Media y Varianza

Tipo de Variable	Estimador MV para Media	Estimador MV para Varianza
Bernoulli	$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\hat{p}(1-\hat{p})$
Binomial	$\hat{p} = \frac{x}{n}$	$\frac{x}{n}(1-\frac{x}{n})$
Poisson	$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\hat{\lambda}$

Ejercicios Prácticos de Estimación de Medias con Datos

Ejercicio 1 - Variable Bernoulli: En un estudio clínico, se examina la efectividad de un nuevo tratamiento para una enfermedad. De un total de 50 pacientes tratados, 25 experimentan mejoras significativas. Calcula la estimación puntual y el intervalo de confianza del 95% para la proporción de pacientes que mejoraron.

Ejercicio 2 - Variable Binomial: Se realiza un estudio para evaluar la tasa de éxito de un procedimiento quirúrgico. En una muestra de 100 procedimientos, se observan 20 casos exitosos. Estima la tasa de éxito en la población y proporciona un intervalo de confianza del 90%.

Ejercicios Prácticos de Estimación de Medias con Datos

Ejercicio 3 - Variable Poisson: En un hospital, se registran el número de pacientes que ingresan a urgencias cada hora durante un día. Los valores observados son: 2, 4, 1, 3, 2, 5, 2, 3, 4, 1, 3, 2, 1, 3, 2, 4. Calcula la estimación puntual y el intervalo de confianza del 95% para la tasa promedio de ingresos por hora.

Distribución Uniforme

La **distribución uniforme** es un modelo de probabilidad en el que todos los valores en un intervalo dado tienen la misma probabilidad de ocurrir. En otras palabras, es una distribución en la que cada valor dentro de un rango tiene la misma oportunidad de ser observado.

La función de densidad de probabilidad (pdf) de una variable aleatoria continua X que sigue una distribución uniforme en el intervalo $[a, b]$ se define como:

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

donde a es el límite inferior del intervalo y b es el límite superior.

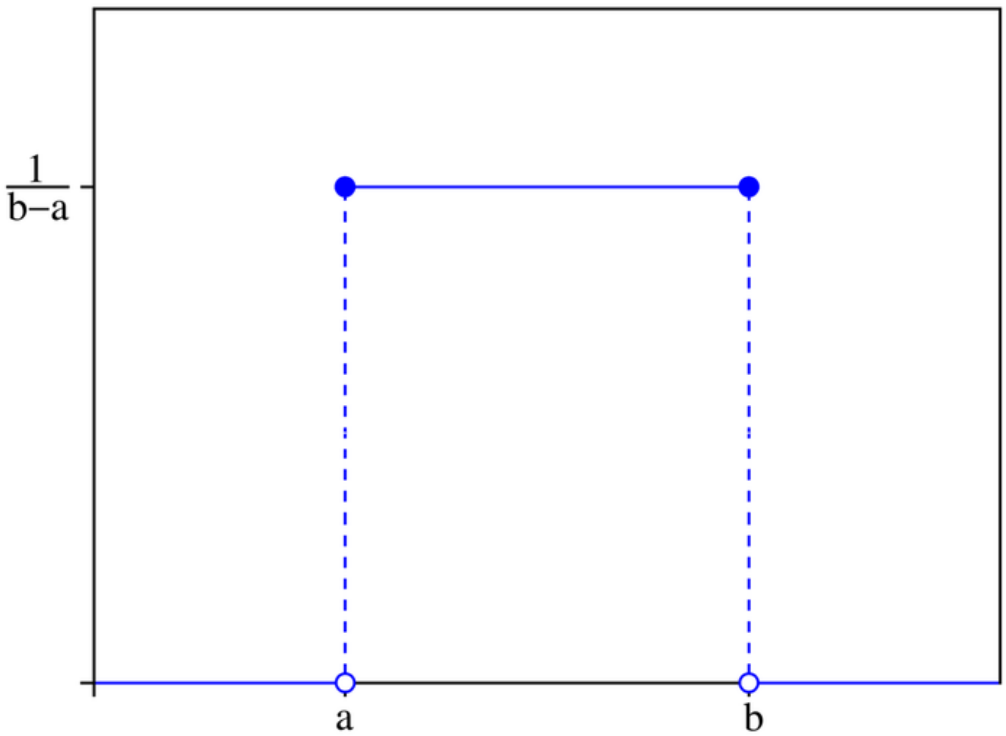

Figure: Distribución Uniforme

Ejemplo Médico de la Distribución Uniforme

Imagina que estás estudiando el tiempo que tardan los pacientes en completar un cuestionario de evaluación de dolor en una clínica de manejo del dolor. Supongamos que el cuestionario tiene un límite de tiempo de 5 minutos para ser completado.

En este escenario, el tiempo que los pacientes toman para completar el cuestionario puede modelarse utilizando una distribución uniforme en el intervalo [0, 5]. Cada valor dentro de este intervalo tiene la misma probabilidad de ser el tiempo de finalización.

Esto podría ser útil para analizar el flujo de pacientes en la clínica y asegurarse de que haya suficiente tiempo asignado para completar los cuestionarios de manera efectiva.