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

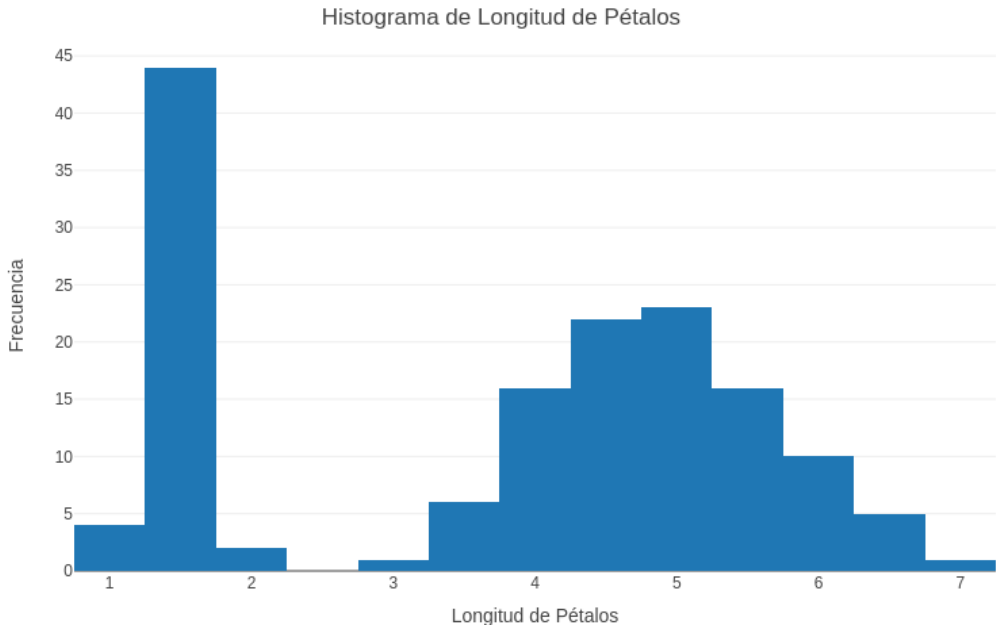

Figure: Ejemplo de un histograma

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.

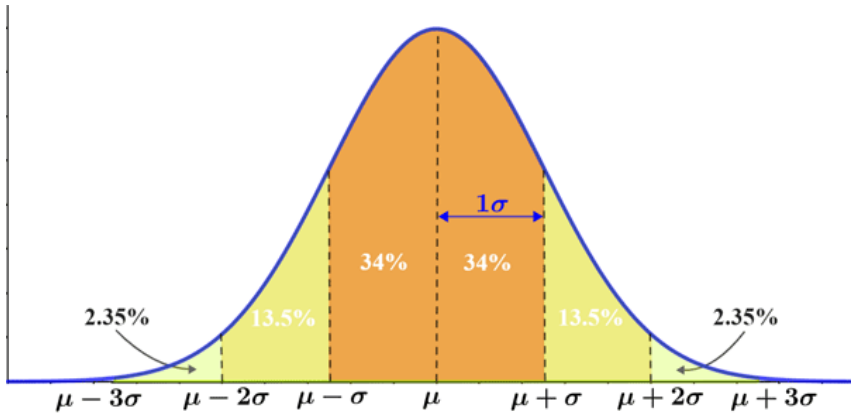

Figure: Regla empírica para distribuciones normales

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