Estadística II (CM0243) - Ingeniería de Producción - Notas de clase

Alejandro Calle Saldarriaga

19 de enero de 2022

Índice

Resumen

En este documento voy a ir subiendo las notas de clase (y los respectivos ejercicios recomendados) para complementar las clases que les voy a dar. Recomiendo fuertemente leerlo antes de cada clase, o en el peor de los casos, antes de cada parcial. También hay un par de recomendaciones de lecturas, por si no entienden los contenidos acá pueden ayudarse con los libros de los que saqué el material. Los talleres son los ejercicios propuestos acá. Es integral hacerlos: de eso depende su nota, además de que son la principal fuente para estudiar para los parciales. También acá voy poniendo las descripciones de las funciones que vamos a utilizar en R para hacer los análisis de datos correspondientes. Este documento va a ir evolucionando a medida que el curso avance. Cualquier error que vean, o duda puntual, escribirme a acalles@eafit.edu.co. El proceso para resolver dudas es el siguiente: primero, me escriben al correo contándome las dudas. Luego, como yo no tengo horario de oficina al ser de cátedra, cuadramos un espacio para resolver las dudas, ya sea por Microsoft teams o en algún lugar de la universidad. Todos los archivos de código en R, además de esté documento, estarán en la página web del curso: `https://acallesalda.github.io/teaching/2022-1-estadisticaII`. Al final de cada sección dejo una guía bilbiográfica, de la cual pueden guiarse para ver que otros recursos pueden mirar si acá no hay algo claro, o si quieren más ejemplos.

1. Introducción y pactos a la clase

La idea de la estadística es sencilla: tenemos datos. ¿Cómo sacamos información, conclusiones de esos datos, sabiendo que tenemos incertidumbre?

Exercise 1. *¿Qué fuentes de incertidumbre se les pueden ocurrir? Este es un ejercicio conceptual. No hay respuestas incorrectas. No se tienen que extender mucho, pero si les pido que tengan respuestas concretas.* □

La estadística es la base del método científico. Con ella podemos sacar conclusiones rigurosas a partir de las mediciones que le hacemos a ciertas variables de interés. Esto obviamente tiene un gran campo de aplicación a la empresa: ¿Como concluir que una máquina dada esta produciendo más que otra? ¿Como podemos predecir nuestro nivel de ventas en el próximo año? Estos son algunas de los miles de problemas que

se pueden atacar usando estadística. Ahora, más que nunca, se usa la estadística en contextos empresariales: estamos en la era de los datos (*big data*, como suele uno leer por ahí. Algunos llegan a decir que los datos son el nuevo petróleo.). ¿Cómo creen que INSTAGRAM conoce tanto de ustedes? Pues porque está sacando constantemente datos de sus hábitos de navegación, y sacando conclusiones a partir de estos (o sea, sacando información a partir de los datos usando estadística).

La teoría estadística fue creada más o menos en el siglo XVIII (y sigue siendo creada hoy en día), por algunos matemáticos brillantes, que se basaron en la teoría de la probabilidad. ¿Pero cómo recogían datos antes, sin computadores? Era muy difícil. Gracias a su ingenio, estos matemáticos lograron deducir lógicamente ciertas leyes que deben cumplir los datos y las propiedades que estos tienen, sin tener que hacer cálculos tediosos que a mano son prácticamente imposible. Gracias a estos matemáticos tenemos hoy estadística, y podemos usar los métodos que se han ido deduciendo a lo largo de los años para analizar nuestros datos. Hoy en día, con la cantidad tan absurda de datos que cada día va subiendo, hacer estadística con lápiz y papel es una cosa del pasado, relegado a los cursos de estadística o a la investigación. Tenemos computadores, que son capaces de hacer esos tediosos cálculos, mucho más rápido y más eficientemente que un ser humano. Por eso en un curso de estadística es esencial introducir software que nos permitan hacer estos cálculos. Acá usaremos R (pero hay muchos más: STATA, EVIEWS, PYTHON, C++, EXCEL, etc.), por razones que luego discutiremos. Manejar alguno de estos software (generalmente, manejar varios) y tener conocimiento estadístico es esencial para convertirse en lo que hoy se llama *Data Scientist* (o científico de datos), lo que el Harvard Bussines Review llama el trabajo más sexy del siglo XXI[1]. Aunque si me preguntan a mí, un científico de datos es un nombre de marketing para lo que antes se conocía como estadístico[2].

Las clases van a funcionar así. Yo les dictaré el curso primariamente usando teclado y marcador, con ciertas pausas para hacer cositas en R. Todos los códigos de clase se los mandó. En estas notas van a estar las cosas que les dicto con teclado/marcador, aunque un poco más profundamente, ya que el medio escrito se presta para más profundidad. Recomiendo fuertemente estudiar de la siguiente manera:

- Venir a clase y prestar la mayor atención posible. Acá voy a resolver algunos ejercicios (en tablero y marcador o en R), en lo posible tomar nota de eso.
- En este documento voy dejando ejercicios. Estos son talleres que yo voy a calificar. Los ejercicios van a estar marcados como prácticos o teóricos. Hay dos talleres, un trabajo práctico final (más tarde cuadramos entre todos que hacemos acá).
- Así se va a calificar el curso:
  - Taller 1: 15 % teórico, 10 % práctico. No desatiendan la parte teórica porque vale menos: esto será su principal herramienta para estudiar para el parcial. La parte práctica tiene que estar implementada en R. La parte práctica es una serie de ejercicios con unos datos que yo les comparta. La parte teórica me la pueden entregar en lápiz y papel (legible, o si quieren, en word o un pdf de LaTeX). Me lo tienen que mandar al correo antes de el Domingo 13 de Marzo, 23.59 PM. Me mandan la parte práctica escaneada y la parte teórica, todo en un archivo .zip. La parte práctica tiene dos componentes: Archivo de .R donde implementen el código (comentado), archivo de word (o pdf, o lo que sea) donde hacen el análisis.
  - Parcial 1: Basado en la parte teórica del Taller 1. Fecha: 14 o 15 de Marzo. Cuadramos dependiendo de como vamos en el curso.
  - Taller 2: Vale el 5 %, totalmente teórico. Su principal herramienta para estudiar para el parcial 2. En papel y lápiz (legible, o si quieren, en word o un pdf de LaTeX). Entregar antes de Mayo 9 (Domingo).
  - Parcial 2: Basado en el taller 2. Lo hacemos en Mayo 9 o 10, dependiendo de como vayamos.

[1] `https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century`

[2] Puede que haya algo de controversia en esto. Depende de a quien le pregunten, la respuesta va a ser diferente. Yo he trabajado con científicos de datos de diferentes disciplinas (estadísticos, ingenieros matemáticos, ingenieros de producción, ingenieros físicos, matemáticos, ingenieros de sistemas), y aunque cada uno tiene una serie de habilidades diferentes, lo que los une es el uso y manejo de datos para sacar conclusiones, que es exactamente lo que se aprende en estadística.

- Trabajo práctico: Hay que entregar dos cosas: Código en R, comentado. Un reporte tipo artículo (subo puntos si es en LaTeX). Los temas para esto los cuadraremos después: yo les traigo algunas ideas, las discutimos, y ustedes me dan sus ideas. La idea es que entre todos cuadremos los temas. Fecha: TBA.
- Supletorios: Me tienen que dar una excusa válida, que tengo que aprobar yo y el jefe del área de estadística. Tienen que dármela lo antes posible, ya que hay 10 días máximo para presentar el parcial antes de la fecha pactada.

Los parciales son individuales, los trabajos son en grupos de a 2 o 3 dependiendo de cuanta gente haya en el curso.

2. Repaso: Algunos conceptos de estadística

Esto es solo un breve repaso de lo que ya deben saber para ver este curso, los conceptos que deben tener de Estadística I. No voy a irme muy a fondo en esto.

La idea de la estadística es sacar conclusiones de una población usando solo la muestra. La muestra es un subconjunto de la población. En el mundo ideal no se necesitaría la estadística: nada más analizas la población y ya. Pero hay que ser eficiente con los recursos.

La estadística está construida encima de la teoría de la probabilidad, que a su vez está construida encima de la teoría de la medida[3].

Hay varios estadísticos importantes que me ayudan a resumir una muestra. Consideremos una muestra $\{X_1, X_2, \ldots, X_n\}$, i.i.d (esta es notación importante, recuérdenla. Significa independientes e idénticamente distribuidos. Lo que significa es que cada el elemento de la muestra que tenemos es independiente de los demás, y que todos siguen la misma distribución, o sea, que todos fueron generados por el mismo proceso generador de datos). El primer estadístico que nos interesa es la media:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

que simplemente se refiere al promedio de los datos. Otra medida similar (en el sentido que las dos miden centralidad) es la mediana. Para calcular la mediana, organicemos los datos de la siguiente manera: $\{X_{[1]}, X_{[2]}, \ldots, X_{[n]}\}$, que son los mismos datos, pero ordenados de manera creciente. O sea, $X_{[1]}$ es el más chiquito, $X_{[2]}$ el segundo más chiquito, $X_{[n]}$ el más grande. Ahora, la mediana está dada por $X_{\lfloor \lfloor n/2 \rfloor \rfloor}$[4], o sea, el dato de la mitad.

Ilustremos esto con un ejemplo. Digamos que queremos estudiar la altura del curso de cálculo III de EAFIT, pero por cuestiones de privacidad, solo obtenemos 5 datos, que son 169, 173, 164, 210, 157, medidos en centímetros. La media es:

$$\bar{X} = \frac{169 + 173 + 164 + 220 + 157}{5} = 174{,}6$$

Ahora, para calcular la mediana, ordenamos: 157, 164, 169, 173, 210. El dato de la mitad es 169.

Otra medidas importantes son la varianza y la desviación estándar. La varianza está dada por[5]

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1},$$

[3] la teoría de medida es un un área de la matemática, bastante abstracta, que se desarrolló a finales del siglo XIX por Borel, Lebesegue, Radon, Fréchet, etc. Surgió en el estudió de la teoría de integración. Más tarde, en 1931, Kolmogorov, un gigante matemático soviético, se dió cuenta que podía usar la teoría de la medida para formalizar la probabilidad Kolmogorov (1950), algo que se quería hacer desde hace algún tiempo (justo era una parte del sexto problema de Hilbert `https://en.wikipedia.org/wiki/Hilbert%27s_sixth_problem.`, la lista más importante de problemas no resueltos en matemáticas). Me parece fascinante que Kolmogorov, estudiando una área tan formal y rigurosa como la teoría de la medida, haya podido ver algún vínculo con algo tan aplicado como la teoría de la probabilidad, y haya formulado sus axiomas.

[4] Esta es la función piso. Está función aproxima un número real al entero anterior. Por ejemplo, $\lfloor 1{,}5 \rfloor = 1$, $\lfloor 1{,}1 \rfloor = 1$, $\lfloor 3{,}7 \rfloor = 3$

[5] El gorrito de $\hat{\sigma}^2$ significa que es una estimación. Siempre que vean un gorrito encima de algo es que estamos estimando una cantidad de la muestra. Nos ayuda a diferenciar entre parámetros y estimaciones.

y la desviación estándar:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}.$$

Ahora, recordemos que es una variable aleatoria. Una variable aleatoria es una variable que depende de los resultados de un evento aleatorio. Antes de que hagamos los experimentos, una variable aleatoria se refiere a los *posibles* valores que puede tomar dicho evento aleatorio. Las variables aleatorias pueden ser discretas (por ejemplo, ganar o no la lotería, el número de penaltis que le van a chutar a Courtouis antes de que tape uno, el número de personas que va a llegar a hacer cola a un banco de 2 a 3 de la tarde, etc.), o pueden ser continua (la altura de estudiantes en un curso, la nota de ustedes en estadística II, el valor de la acción de sura mañana a las 11 de la mañana). Concentrémonos en las variables aleatorias continuas. La variable aleatoria continua más usada es la normal. Generalmente, describimos a las variables aleatorias continuas con su función de densidad (abreviadas generalmente *pdf*, porque en inglés se escribe *probability density function*). Para una normal con media μ y varianza σ^2[6], la función de densidad es:

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

La función de distribución de una variable aleatoria (también conocida como la acumulada, en inglés *cdf* por *cumulative distribution function*) está dada por:

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

La cdf de la normal no se puede expresar con funciones elementales. A esa cdf se le llama $\Phi(x)$ en muchos libros. Sus valores numéricos son bien conocidos y hay tablas que las reportan. Hoy en día, los calculamos usando algún software. Notar que la cdf y la pdf están muy relacionadas: la pdf es la derivada de la cdf, la cdf es la integral (en todo el dominio, hasta x) de la pdf. Si tengo una, puedo calcular la otra. Para calcular la probabilidad de que mi variable aleatoria tome valores en un intervalo (a, b), integramos la pdf:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(a) - F(b)$$

Exercise 2. *Mostrar que $\int_a^b f(x)dx = F(a) - F(b)$, donde f es la pdf de una variable aleatoria y F es la cdf de esa misma variable.* □

El valor esperado de una variable aleatoria, y más específicamente de la normal, es:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx = \mu.$$

La varianza está dada por:

$$Var[X] = E[X^2] - E[X]^2 = \sigma^2$$

Los valores esperados y las varianzas tienen bastantes propiedades interesantes. Les recomiendo leer `https://asesoriatesis1960.blogspot.com/2018/05/propiedades-de-la-esperanza-matematica.html` y `https://proyectodescartes.org/iCartesiLibri/materiales_didacticos/EstadisticaProbabilidadInferencia/Estadistica1D/5_2Varianza.html` para refrescarlas.

Una representación alternativa a la pdf para una variable aleatoria es la función generadora de momentos (de ahora en adelante, *mgf* por el inglés *moment generating function*). Esta esta dada por:

$$M_X(t) = E[e^{tX}] = e^{\mu t + \sigma^2 t^2/2} \tag{1}$$

Esta función tiene propiedades bastante interesantes. La mgf es función de la variable t. Si la derivamos con respecto a t, y a la función insertamos $t = 0$, obtenemos el primer momento de la variable aleatoria,

[6]Las densidades, generalmente, se describen con sus parámetros. Conocer la media y varianza de una normal me hace poder describir completamente dicha normal

o sea, $E[X]$. El k-ésimo momento[7] de una variable aleatoria es $E[X^k] = \int_{-\infty}^{\infty} x^k f(x)$. Calcular esa integral puede ser engorroso. Pero hay otra forma más sencilla de calcular esto: simplemente, derivamos la mgf k veces, y introducimos $t = 0$. Hagámoslo con la normal: derivemos la ecuación 1 para encontrar la media y la varianza de una distribución normal:

$$M'_X(t) = \left[\mu + \frac{2\sigma^2 t}{2}\right][e^{\mu t + \sigma^2 t^2/2}]$$

y evaluando en 0:

$$\begin{aligned} M'_X(0) &= \left[\mu + \frac{2\sigma^2 0}{2}\right][e^{\mu 0 + \sigma^2 0^2/2}] \\ &= [\mu][e^0] \\ &= \mu \end{aligned}$$

O sea, $E[X] = \mu$, que es justamente la esperanza de la normal, lo que esperábamos. Ahora, derivemos otra vez, usando la regla del producto:

$$\begin{aligned} M''_X(t) &= [\sigma^2][e^{\mu t + \sigma^2 t^2/2}] + [\mu t + \sigma^2 t^2/2][e^{\mu t + \sigma^2 t^2/2}][\mu t + \sigma^2 t^2/2] \\ &= [e^{\mu t + \sigma^2 t^2/2}][\sigma^2 + (\mu + \sigma^2 t)^2] \\ &= [e^{\mu t + \sigma^2 t^2/2}][\sigma^2 + \mu^2 + 2\sigma^2 \mu t + 4\sigma^4 t^2] \end{aligned}$$

Evaluando en 0, tenemos que

$$\begin{aligned} M''_X(0) &= [e^{\mu 0 + \sigma^2 0^2/2}][\sigma^2 + \mu^2 + 2\sigma^2 \mu t 0 + 4\sigma^4 0^2] \\ &= e^0[\sigma^2 + \mu^2]. \end{aligned}$$

O sea, $E[X^2] = \sigma^2 + \mu^2$. Como $E[X]^2 = \mu^2$, luego $Var[X] = E[X^2] - E[X]^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$.

Exercise 3. *Suponga que X es una variable aleatoria que sigue una distribución exponencial. La función generadora de momentos para X es $m_X(t) = \frac{\lambda}{1-\lambda}$. Halle la media y la varianza de la variable usando la mgf.* □

2.1. Guía bibliográfica

Si necesitan leer un poco más sobre la distribución normal, pueden leer la sección 4.5 de Wackerly et al. (2010), la sección 4.3 de Devore (2008) o la sección 6.2 de Walpole (2007). Para ver más sobre la esperanza, pueden leer la sección 4 de Walpole (2007). Para más sobre la functión generadora de momentos, pueden leer la sección 6.5 de Wackerly et al. (2010).

3. Una Breve Introducción a R

Esta introducción requiere que tengan abierto el documento en una ventana, y en otras van haciendo lo que acá les voy diciendo.

Primero, vamos a instalar R y RStudio. A lo largo de la clase, vamos a usar R a través de RStudio. La diferencia entre R y RStudio es que el primero es un lenguaje de programación, el que hace todos los cómputos que necesitamos, y el segundo es la interfaz gráfica mediante la que vemos lo que vamos haciendo. Vamos a instalar R, vamos a instalar RStudio, vamos a instalar un paquete de R. Lo pueden pensar como

[7] El primer momento de una variable aleatoria es la esperanza (valor esperado), y el segundo está relacionado con la varianza. Esos son los momentos que generalmente consideramos, aunque hay información importante en momentos de orden superior. En particular, los momentos 3 y 4 están relacionados con la asimetría y la kurtosis de la distribución, que son medidas importantes.

si fuese un carro: R es es el motor, lo que hace que el carro ande y funcione. RSTUDIO es lo que ve uno cuando se monta a la silla de conductor: el volante, la pala de cambios, el acelerador, el freno. Es con lo que el conductor del carro interactúa. Ahora, el paquete es una adición a R. Por ejemplo, nos pareció muy anticuada la radio del carro, entonces le añadimos una pantalla táctil para usar eso más bien.

Voy a asumir que están usando Windows. Para los usuarios de macOS, sigan el siguiente tutorial: `https://datacritica.org/2021/03/19/instalacion-de-r-y-rstudio-en-macos/`. Primero, bajen el instalador[8] de R y denle doble click. Sigan las instrucciones y lo instalan. La versión que vamos a usar es la 4.1.1, llamada *Kick Things*. Todas las versiones de R tienen nombres sacados de la Charlie Brown. Luego, instalen RStudio[9]. Cuando instalen RStudio, ábranlo y se encontrar algo como en la Figura 1. En Q1 está el editor, acá pueden escribir los programas que van a usar. En el Q2 está la consola, que ahí es donde se corren los comandos que escriben en Q1 (o pueden escribir ahí directamente). En Q3 están los objetos de R, esos generalmente son los datos que carguemos y las cosas que vamos calculando. En Q4 hay está el directorio, ahí pueden ver los archivos que quieran abrir.

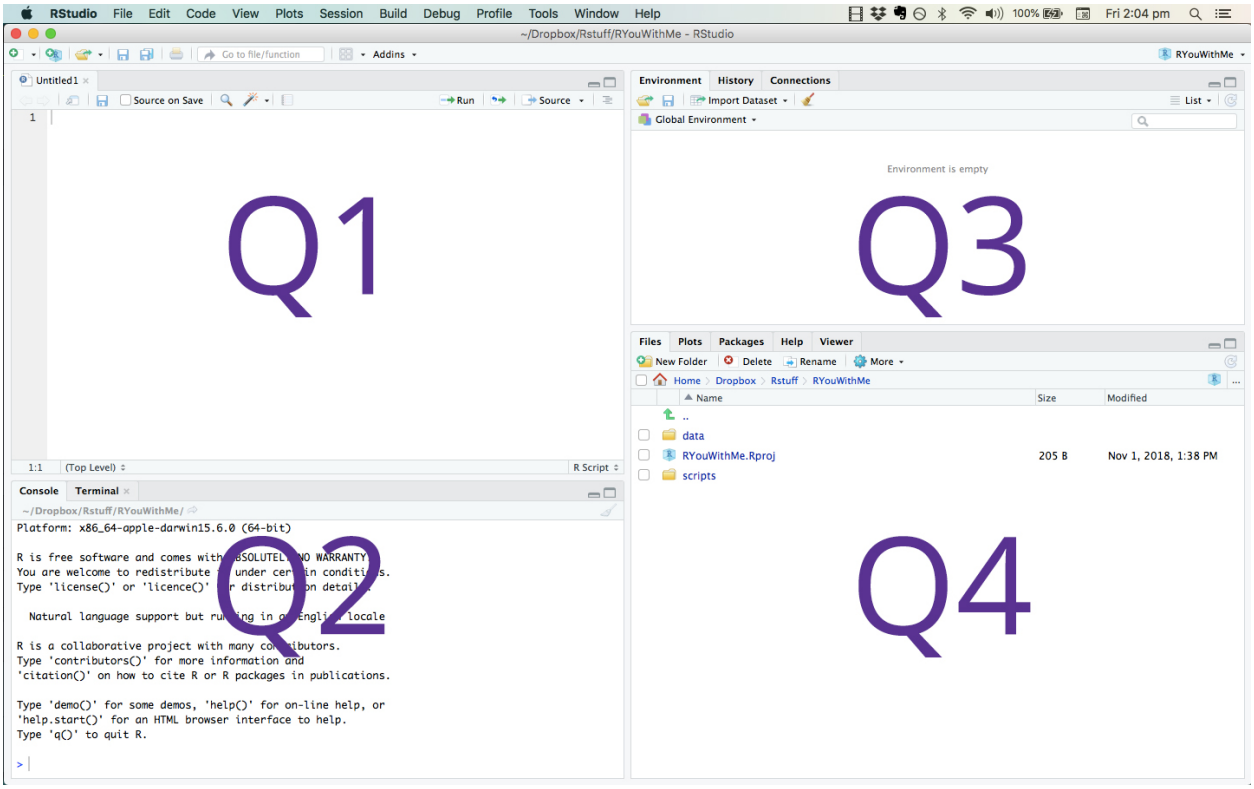

Figura 1: Cuadrantes de RStudio

Hagamos un pequeño ejemplo: vamos a instalar un paquete de R, vamos a computar algunas cosas sencillas, y vamos a usar el paquete que instalamos. El paquete que vamos a instalar se llama GGPLOT2, un paquete muy usado en la comunidad estadística para hacer gráficos bonitos (Wickham, 2016).

Para instalar un paquete, escriban en la consola

```
install.packages("ggplot")
```

En general, se escribe `install.packages("paquete")`, con el nombre del paquete correspondiente. Los paquetes que se pueden instalar así son los paquetes de CRAN (*The Comprehensive R Archive Network*), que es donde los estadísticos suben sus paquetes para que otra gente lo use[10].

[8] `https://cran.r-project.org/bin/windows/base/old/4.1.1/R-4.1.1-win.exe`

[9] `https://www.rstudio.com/products/rstudio/download/`

[10] Todo esto es Software libre! O sea, no hay que pagar nada para usarlos. Pueden ver varios paquetes que hay ahí acá `https://cran.r-project.org/`

Ahora, bajen los datos que están acá `https://data.mendeley.com/datasets/7xwsksdpy3/1/files/29227286-d2f0-40cc-8ff0-dfb9f3e461bb`. Vamos a abrirlo con R. Acá es donde le tienen que poner cuidado a Q4: para poder abrir el archivo, el archivo tiene que estar en esa carpeta. Si no está ahí, R no es capaz adivinar en que parte del computador lo tienen. Yo, personalmente, les recomiendo crear una carpeta del curso, y meter ahí todos los archivos y scripts que vayan usando. Usen ese explorador para encontrar la carpeta que crearon, y una vez estén en ella, denle click a la ruedita que dice "More", y le dan click a la opción "Set as working directory".

Ahora, en el editor (Q1), escriban `library(ggplot2)`[11]. Denle guardar, y pónganle al archivo como deseen. En el directorio entonces van a tener dos archivos: los datos y este script. Ahora denle click a esta serie de cosas: "File" $\longrightarrow$ "New Project" $\longrightarrow$ "Existing Directory" $\longrightarrow$ "Existing Directory" $\longrightarrow$ "Create project". Eso les crea un nuevo archivo, con terminación `.Rproj`. Ahora cada vez que quieran trabajar en R, doble click a ese icono y eso les va a poner de directorio el directorio que escogieron ahorita. Si no hacen este paso, les toca cambiar de directorio cada vez que quieran trabajar.

Momento de leer los datos en R. En el editor, escriban `iris <- read.csv("iris-write-from-docker.csv")`. Acá le estoy diciendo a R que me lea esos datos, y me los guarde en una matriz llamada iris. Los datos están separados por comas (manera muy usual de guardar datos), así que le tengo que decir a R que el separador es una coma. Al correr el código esto, les debe salir en el Q3 un objeto llamado iris. Para correr el código, seleccionan el texto en el editor y le dan control + enter o seleccionan el texto y clickean en la parte que dice Run (En Q1 arriba a la derecha).En Q3 debe salir algo llamado iris, que dice 150 obs. of 5 variables. Si no les sale así, probablemente tuvieron un error. Pueden ver la matriz que está en iris dándole click.

Ahora, digamos que queremos ver una variable específica de estos datos. Por ejemplo, hay una variable llamada `petal_width`. Vamos a mirar específicamente esta variable en este ejercicio. Para extraerla a un vector, escriban `petal_width <- iris$petal_width`. Eso les crea una variable nueva con los datos que queremos. Ahora, miremos la media y la desviación estándar de esa variable. Para la media, usamos `mean(petal_width)` y para la desviación estándar usamos `sd(petal_width)`. Si corren todo el código, en la consola (Q2) les debe aparecer

```
> mean(petal_width)
[1] 1.198667
> sd(petal_width)
[1] 0.7631607
```

O sea, la media es 1,198667 y la desviación estándar es 0,7631607. Ahora, hagamos un histograma usando GGPLOT2. Escriban en el script `ggplot(iris, aes(x=petal_width)) + geom_histogram()`. Cuando lo corran, en el Q4 les debe aparecer el histograma. Hasta acá la pequeña introducción a R. Acá (`https://acallesalda.github.io/files/ejemplo_inicial.R`) pueden bajarse el archivo de script de R. Notar que el código está comentado (las frases que empiezen con # no las lee R, son solo para que las mire el programador. Comentar los scripts es muy importante: nos permite volver a ellos en el futuro para saber que es lo que hicimos, y nos facilita el trabajo en equipo. Los archivos de script de R siempre terminan en `.R` Sigan este tutorial al pie de la letra y nada raro debe pasar. Si no fueron capaz de seguirlo, si les pasó algo, por favor, lo antes posible, háblenme. En todo el curso vamos a estar usando R y si no lo tienen instalado va a ser complicado seguir el curso. Además, no podrán hacer los ejercicios prácticos.

Exercise 4. *Una de las métricas por la que se miden los académicos es por las citas. A los académicos les interesa que los citen, y una manera de ser citado es publicar software libre, para que los que usen el paquete los citen. Averigüen como se cita un paquete en R (hay un comando que me entrega la cita del paquete en un formato llamado BibTex). Corran el comando para* GGPLOT2 *y copie y pegue lo que les sale en la consola de* R. □

3.1. Guía bibliográfica

Pueden encontrar una introducción bastante sencilla a R en `https://rladiessydney.org/courses/ryouwithme/01-basicbasics-1/`. También, Venables and Smith (2009) da una excelente y más completa introducción.

[11] Cada que vayamos a usar un paquete de CRAN, es necesario cargarlo al principio del script. Solo hay que instalar los paquetes una vez, pero hay que cargarlos cada que se quieran usar

4. Estadísticos muestrales y sus distribuciones.

4.1. Muestreo y estadísticos

Uno de los conceptos centrales de la estadística es el del estadístico. Un estadístico (además de ser una persona que se graduó de estadística) es una función de los elementos de la muestra. Por ejemplo si tenemos la muestra $\{X_1, X_2, \ldots, X_n\}$, un estadístico sería

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

la media muestral. Eso quiere decir que si conocemos la distribución de $\{X_1, X_2, \ldots, X_n\}$, podemos usar algunas técnicas para encontrar la distribución de $\bar{X}$. Otro estadístico sería $X_{[1]}$, que es el elemento más pequeño de la muestra. Otro podría ser $X_1 + X_n$. Cualquier función de la muestra es un estadístico, y en teoría, podemos hayar (o aproximar) su distribución. La distribución de un estadístico es llamada generalmente la distribución muestral. Los estadísticos son usados para hacer inferencia sobre los datos, o sea, para tomar decisiones o para estimar los parámetros poblacionales. Recordemos: usamos la estadística para tomar conclusiones sobre la población, aunque solo tengamos la muestra, que debe ser representativa.

En este curso usaremos varias herramientas matemáticas para *deducir* de forma lógica las distribuciones de varios estadísticos[12]. Pero otra forma de hallar estas distribuciones es *simular* con la ayuda de algún software, en nuestro caso, R. Volvamos a los datos que utilizamos antes. Recordemos que calculamos la desviación estándar de una variable que teníamos. ¿Cual será la distribución de esta desviación estándar?

Para calcular esto, podemos remuestrear la muestra. Remuestrear es un proceso mediante el cual construimos una muestra diferente con nuestra muestra original. Recordemos que nuestra muestra tenía 150 valores. Para remuestrear (con reemplazo), nada más voy escogiendo 1 a 1 de esos 150 valores, al azar y con la misma probabilidad, hasta tener 150 otra vez. Note que si hacemos este proceso sencillo, lo más probable es que tengamos datos repetidos en nuestra muestra construida. Ahora, a esta muestra nueva, le calculamos la desviación estándar, y la guardamos por ahí. Repetimos este remuestreo muchas veces, digamos 1000, y calculamos 1000 veces la desviación estándar. Si hacemos un histograma de estas 1000 desviaciones estándar, vamos a obtener una aproximación de la distribución de la desviación estándar de nuestros datos. Este sencillo proceso es conocido como bootstrapping (Efron, 1979), y es uno de los métodos más usados y poderosos de la estadística computacional. En R es bastante sencillo. El siguiente script implementa este ejercicio[13]:

```
# cargar paquetes
library(ggplot2)

# leer datos
iris <- read.csv("iris-write-from-docker.csv")
petal_width <- iris$petal_width

#inicializo vector vacío
sds <- numeric(1000)

# Repito mil veces
for (i in 1:1000){
  # guardo en la posición i la desviación estándar correspondiente a este
  # remuestreo
  petal_width_resample <- sample(petal_width, replace = TRUE)
  sds[i] <- sd(petal_width_resample)
}

# dibujo el histograma
```

[12] Esto se puede hacer siempre y cuando las distribuciones de nuestros datos sean conocidas y suficientemente sencillas, y el estadístico sea una función sencilla de los datos. Si no se cumple esto, estás deducciones son complejas.

[13] Recuerde instalar los paquetes que no tenga

```
hist(sds)
```

Exercise 5. *Simule la distribución muestral de la mediana de la variable que acabamos de considerar, y grafique el histograma.* □

Exercise 6. *Haga el ejercicio 44 de Devore (2008). Para simular un vector de datos Weibull con* $\alpha = 2$ *y* $\beta = 5$*, utilice el comando* `x <- rweibull(n, 2, scale = 5)`*, donde n son los valores que están en el libro.* □

5. Guía bibliográfica

Para muestreo y una introducción a estadísticos, leer 8.2 de Walpole (2007) y 5.3 de Devore (2008).

Referencias

Devore, J. (2008). *Probabilidad y estadística para ingeniería y ciencias.* Matemáticas (International Thomson). International Thomson.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* (1), 1–26.

Kolmogorov, A. N. (1950). *Foundations of the theory of probability.* New York: Chelsea Publishing Co.

Venables, W. N. and D. M. Smith (2009). *An Introduction to R* (2nd ed.). Network Theory Ltd.

Wackerly, D., W. Mendenhall, and R. Scheaffer (2010). *Estadística matemática con aplicaciones.* Grupo Editorial Iberoamérica.

Walpole, R. (2007). *Probabilidad Y Estadistica Para Ingenieria Y Ciencias.* Pearson Educación.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.