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

Alejandro Calle Saldarriaga

27 de noviembre de 2023

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 de la clase

1.1. Pactos de clase

- 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. Hay dos talleres y un trabajo práctico final (más tarde cuadramos entre todos que hacemos acá).
- Así se va a calificar el curso:
  - Taller 1: Vale el 10 %. 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. Deben entregar los talleres en LaTeX. Me lo tienen que mandar al correo antes de el Domingo 13 de Marzo, 23.59 PM. Mandar archivo con scripts de R y pdf en un .zip. Los talleres son los ejercicios propuestos en este documento.
  - Parcial 1: Basado en la parte teórica del Taller 1. Marzo 15.
  - Taller 2: Vale el 10 %. Entregables iguales al taller 1. Entregar antes del Domingo 8 de Mayo, 23.59 PM.
  - Parcial 2: Basado en el taller 2. Lo hacemos en Mayo 10.
  - Trabajo práctico. Vale el 30 %: Hay que entregar dos cosas: la idea es hacer presentaciones en las semanas 17 y 18 si faltan, 18. Hay que entregar el proyecto (notebook de RMarkdown) antes de las presentaciones. Fecha límite, domingo mayo 22 a las 23.59 PM. Temas: por convenir. El 10 % del proyecto va a ser un plan de trabajo. Hay que entregar eso en la semana 10, Domingo abril 3 a las 23.59 PM.
- Covid-19: Ya la universidad regresa completamente a la presencialidad. Si tienen síntomas de Covid por favor no vengan a clase y deben aislarse 7 días. Si estuvieron en contacto con alguien positivo solo se deben aislar si no tienen el cuadro de vacunación completo. Ya sura no hace pruebas si no hay síntomas. Si un estudiante presenta síntomas de covid-19 y hay previstas actividades evaluativas debe solicitar una cita médica virtual a servicio.medico@eafit.edu.co para que el médico de la Universidad emita el certificado del aislamiento por el tiempo definido por el Gobierno Nacional y contactar de manera oportuna al profesor para coordinar, sin necesidad de incapacidad, la forma de evaluación. En caso que el camino sea un examen supletorio este se realizará de acuerdo con el procedimiento establecido en el reglamento. Si a mi me da covid cuadramos para reemplazar la clase presencial.
- Desarrollo estudiantil ofrece varios servicios valiosos para los estudiantes. Uno importante es la asesoría de hábitos y métodos de estudio. Otro, es la consulta psicológica. Si requieren alguno de estos servicios, manden un correo a dllo.estudiantil@eafit.edu.co o acudan al bloque 29, 5to piso.
- ¿Por qué estudiar matemáticas (y estadística)? Leer estos artículos, que son interesantes `https://www.eafit.edu.co/escuelas/ciencias/ciencias-matematicas/servicios/Paginas/porque-estudiar-matematicas.aspx`, `https://www.eafit.edu.co/escuelas/ciencias/ciencias-matematicas/servicios/Paginas/Pensamiento-matematico.aspx`, `https://www.eafit.edu.co/escuelas/ciencias/ciencias-matematicas/servicios/Paginas/Aspectos-a-tener-en-cuenta-para-el-estudio-de-las-matem%C3%A1ticas.aspx`
- Fecha límite de cancelación: 22 de Mayo, por epik.
- **Conducto regular:** El conducto regular del curso por si hay algún problema o inconformidad es este: Profesor – Coordinador de área - Jefe de Departamento – Decano – Consejo de Escuela - Consejo Académico. Si el problema está relacionado con la carrera, debe hablar con el jefe de carrera.
- **Segundo calificador**: Para pedir un segundo calificador se debe seguir el siguiente procedimiento:
  1. Realizar la solicitud de revisión en el momento mismo de la realimentación.

2. Presentar por escrito, al profesor responsable de la asignatura, dentro de los tres (3) días hábiles siguientes a la realimentación, la solicitud de revisión con una justificación clara y motivada del por qué considera que la calificación no es acertada. El profesor dispondrá de cinco (5) días hábiles para resolver el reclamo formulado e informar al estudiante la decisión correspondiente.
3. Cuando el estudiante formule una reclamación el texto de la prueba evaluativa permanecerá en poder del profesor hasta que la petición haya sido resuelta de manera definitiva.

Cuando el trabajo o la prueba hubieran sido elaborados por varios estudiantes, la solicitud de revisión deberá ser formulada por la totalidad de los interesados.

- La evaluación docente se hace en las últimas semanas del curso. Les recomiendo bastante hacerlas, es una herramienta muy importante para mi carrera profesional que debo tener en cuenta.
- Bibliografía del curso: El libro principal es (Wackerly et al., 2010). Bibliografía secundaria: (Devore, 2008; Walpole, 2007).
- Horario clases presenciales: Lunes 4PM-5.30PM, Martes 9.30AM-11AM. Aula: 34-402 para ambas.
- Clase pérdida se repone con un vídeo. Por ejemplo, para los lunes festivos, la clase del lunes la doy el martes y en esa semana lo más pronto posible les pongo el vídeo de lo que sería la clase del martes, y el lunes siguiente sigo con el tema.

Los parciales son individuales, los trabajos son en grupos de a 2 o 3. Se escojen los equipos al principio del curso. Estos grupos son para toda la duración del curso.

1.2. Motivación

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 permita 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 tablero y marcador, con ciertas pausas para hacer cositas en R. Todos los códigos de clase se los voy a colgar en la página del curso. En estas notas van a estar las cosas que les dicto con tablero/marcador, aunque un poco más profundamente, ya que el medio escrito se presta para más profundidad. Recomiendo fuertemente estudiar de la siguiente manera:

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 $X_{\lfloor (n+1)/2 \rfloor}$ para datos de tamaño impar (el dato de la mitad) o $\frac{X_{[n/2]}+X_{[n/2+1]}}{2}$ para datos de tamaño par.

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 (o sea, no tenemos la población sino una uestra, pero igual queremos sacar conclusiones sobre los 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, la mediana.

Otra medidas importantes son la varianza y la desviación estándar. La varianza está dada por[4]

[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.

[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] 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 de esos parámetros. A la población le corresponden los parámetros, a la muestra las estimaciones de estos parámetros.

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

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[5] σ^2, 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:

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

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:

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

Exercise 2. *Mostrar que $\int_a^b f(x)dx = F(b) - F(a)$, 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:

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

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

Esta función tiene propiedades bastante interesantes. La mgf es función de la variable t. Si la derivamos con respecto a t, y en la derivada insertamos $t = 0$, obtenemos el primer momento de la variable aleatoria, o sea, $E[X]$. El k-ésimo momento[6] 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 2 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 $Var[X] = E[X^2] - E[X]^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$, que es justamente la varianza de una normal, o sea, lo que esperábamos.

Exercise 3. *Suponga que X es una variable aleatoria que sigue una distribución exponencial. La función generadora de momentos para X es $m_X(t) = \frac{\lambda}{\lambda - t}$. 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

[6] 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.

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 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[7] 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[8]. Cuando instalen RStudio, ábranlo y se van a 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 está el directorio, ahí pueden ver los archivos que quieran abrir.

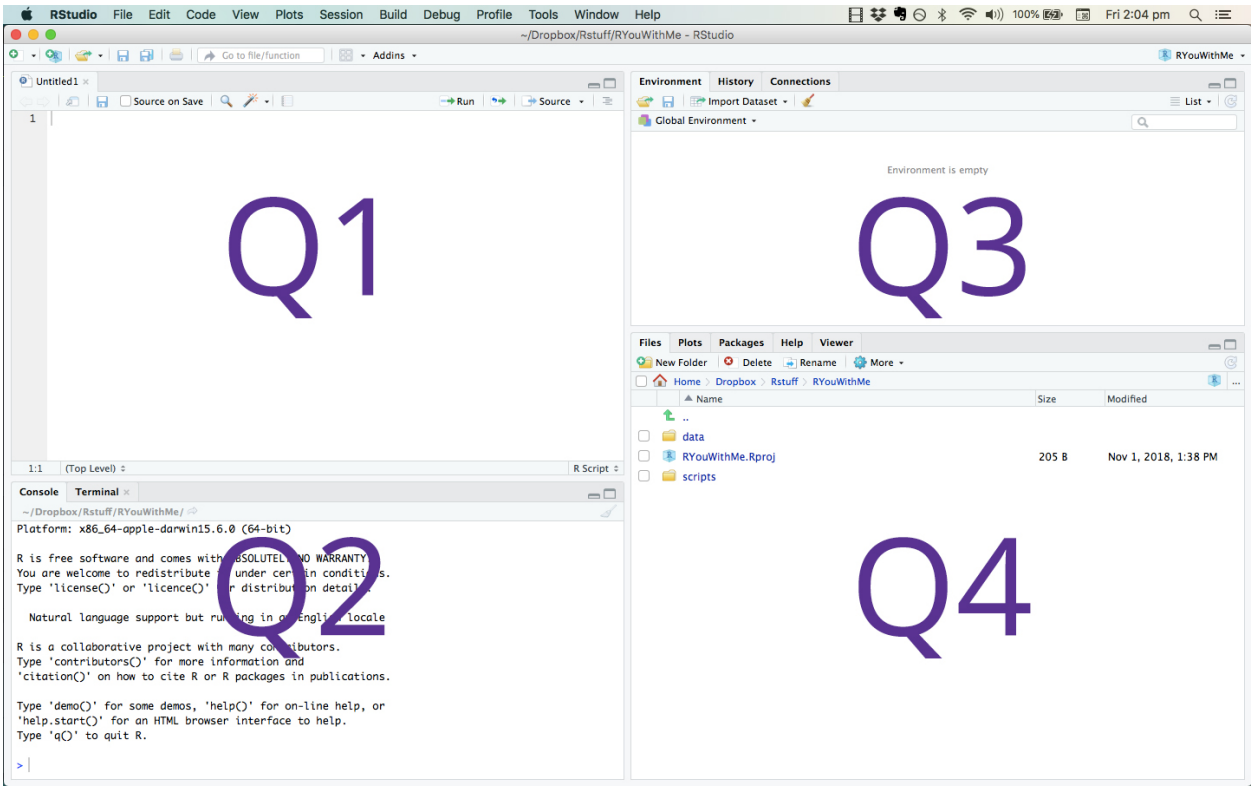


Figura 1: Cuadrantes de RStudio

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*),

[7] `https://cran.r-project.org/bin/windows/base/old/4.1.1/R-4.1.1-win.exe`
[8] `https://www.rstudio.com/products/rstudio/download/`

que es donde los estadísticos suben sus paquetes para que otra gente lo use[9].

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)`[10]. 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" $\longrightarrow$ "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 de trabajo 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 tabulares), así que le tengo que decir a R que el separador es una coma. Al correr el código, 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`. El operador $ me permite acceder a los atributos de un objeto. 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. □

[9]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/`

[10]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

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.

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[11]. 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[12]:

```
# 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
```

[11] 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.

[12] Recuerde instalar los paquetes que no tenga

```
  # remuestreo
  petal_width_resample <- sample(petal_width, replace = TRUE)
  sds[i] <- sd(petal_width_resample)
}

# dibujo el histograma
hist(sds)
```

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

Exercise 6. *Estudiar la distribución muestral de $\bar{X}$ cuando la distribución de la población es Para simular un vector de datos Weibull con $\alpha = 2$ y $\beta = 5$, utilice el comando* `x <- rweibull(n, 2, scale = 5)`. *Considere $n = 5, 10, 20, 30, 100$ y en cada caso utilice* 1000 *réplicas de la muestra. ¿Con cuál n se parece más la distribución a la normal? Ejercicio tomado de (Devore, 2008).* □

4.1.1. Guía bibliográfica

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

4.2. Distribuciones muestrales

Para desarrollar la teoría de esta sección, primero debemos recordar varios resultados de estadística I que necesitamos. Estos resultados tienen que ver con la mgf de la suma de variables aleatorias independiente, con la distribución normal estándar y con la mgf de una variable multiplicada por una constante y la mgf de una distribución χ^2, y la distribución de una estándar al cuadrado. No vamos a probar estos resultados para poder ver lo que se necesita en esta clase.

Resultado 1 (suma de mgfs independientes): Sean $X_1, X_2, \ldots, X_n$ independientes con mgf $m_{X_1}(t), \ldots, m_{X_n}(t)$, entonces $Y = X_1 + X_2 + \cdots + X_n$ tiene mgf $m_Y = m_{X_1}(t) \times m_{X_2}(t) \times \cdots \times m_{X_n}(t)$

Resultado 2 (distribución normal estándar): Si $X \sim N(\mu, \sigma^2)$ luego $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$. A Z se le llama la distribución normal estándar.

Resultado 3 (producto mgf y constante): Si X tiene mgf $m_X(t)$, luego aX tiene mgf $m_X(at)$

Resultado 4 (mgf chi cuadrado): Si $X \sim \chi^2(n)$ entonces $m_X(t) = (1-2t)^{-k/2}$ para $t < 1/2$.

Resultado 5 (distribución de estándar al cuadrado): Sea $Z \sim N(0,1)$. Luego, $Z^2 \sim \chi^2(1)$.

Recordemos que la idea de este curso es conocer las distribuciones de ciertos estadísticos de interés. Uno de los estadístico que más usamos es la media: ¿como será la distribución de la media entonces? Consideremos el caso más sencillo: cuando nuestros datos son normales. Esto nos lleva al siguiente resultado:

Teorema 1 (Distribución de la media muestral) $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ entonces $\bar{X} \sim N(\mu, \sigma^2/n)$.

Prueba. Recordemos que $m_{X_i}(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$, al ser X_i normal. Ahora, por el resultado 3, tenemos que $m_{X_i/n}(t) = e^{\frac{\mu t}{n} + \frac{\sigma^2 t^2}{2n^2}}$. Al ser X_i independientes, por el resultado 1 tenemos que

$$\begin{aligned} m_{\bar{X}(t)} &= m_{X_1/n}(t) \times m_{X_2/n}(t) \times \cdots \times m_{X_n/n}(t) \\ &= e^{\frac{\mu t}{n} + \frac{\sigma^2 t^2}{2n^2}} \times e^{\frac{\mu t}{n} + \frac{\sigma^2 t^2}{2n^2}} \times \cdots \times e^{\frac{\mu t}{n} + \frac{\sigma^2 t^2}{2n^2}} \\ &= e^{\sum_{i=1}^{n} (\frac{\mu}{n} + \frac{\sigma^2 t^2}{n^2}} \\ &= e^{n(\frac{\mu}{n} + \frac{\sigma^2 t^2}{n^2})} \\ &= e^{\mu + \frac{\sigma^2 t^2}{n}} \end{aligned}$$

Pero esta es la mgf de $N(\mu, \sigma^2/n)$, o sea que $\bar{X} \sim N(\mu, \sigma^2/n)$.

Este es un resultado bastante general: nos da la distribución de la media de datos que se distribuyen normal, y con eso podemos sacar varias conclusiones de unos datos. Veamos esto en acción con un ejemplito. Tenemos que la vida media de una máquina para elaborar pan es de 7 años, con una desviación estándar de 1 año. Sabemos que la distribución de la vida de estás máquinas es normal. ¿Cuál es la probabilidad de que la vida media de una muestra de 11 máquinas esté entre 6.4 y 7.2? ¿Cuál es el valor de $\bar{X}$ a la derecha del cual caería el 15 % de las medias calculadas de muestras de tamaño 11?

La primera pregunta es hallar la probabilidad $P(6.4 \leq \bar{X} \leq 7.2)$. Para hallar esta probabilidad, estandarizamos:

$$\begin{aligned} P(6.4 \leq \bar{X} \leq 7.2) &= P\left(\frac{6.4-7}{1/\sqrt{11}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{7.2-7}{1/\sqrt{11}}\right) \\ &= P(-1.98 \leq Z \leq 0.66) \\ &= P(Z \leq 0.66) - P(Z \leq -1.98) \\ &= 0.72 \end{aligned}$$

Para calcular $P(Z \leq 0.66)$ y $P(Z \leq -1.98)$ podemos usar el comando de R `pnorm`. El código sería `pr <- pnorm(0.66) - pnorm(-1.98)`.

Para responder la segunda pregunta, debemos notar que el valor de Z sobre el que el 15 % de la distribución esté a su derecha es el percentil 85 de los datos. Este valor se puede calcular fácilmente en R. Ahora, como $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, tenemos que $\bar{X} = z\frac{\sigma}{\sqrt{n}} + \mu = 7.312496$. En R esto es `qnorm(0.85)*(1/sqrt(11)) + 7`.

Exercise 7. *Si cierta máquina fabrica resistencias eléctricas que tienen una resistencia media de 40 ohms y una desviación estándar de 2 ohms, ¿cuál es la probabilidad de que una muestra aleatoria de 36 de estas resistencias tenga una resistencia combinada de más de 1458 ohms?* □

Ya que tenemos la distribución de la media de datos normales, centrémonos en otro estadístico bastante importante: la varianza. Recordemos que esta se llama $\hat{\sigma}^2$, y dimos su definición atrás (Ecuación 1). Muchas veces se usa S^2 para llamar la varianza muestral también. Primero, vamos a demostrar un resultado intermedio.

Teorema 2 (Suma de chi cuadradas). Sean $Z_1, Z_2, \ldots, Z_n \overset{i.i.d.}{\sim} N(0,1)$. Luego $\sum_{i=1}^n Z_i^2 \sim \chi^2(n)$.

Prueba. Por el resultado 5, tenemos que $Z_i^2 \sim \chi^2(1)$, entonces por el resultado tenemos que $m_{Z_i^2}(t) = (1-2t)^{-1/2}$. Ahora, al ser las Z_i independientes, luego los Z_i^2 son también indepentientes, y podemos usar el resultado 1. Obtenemos que:

$$\begin{aligned} m_{\sum_{i=1}^n Z_i^2}(t) &= m_{Z_1^2} \times \cdots \times m_{Z_n^2} \\ &= (1-2t)^{-1/2} \times (1-2t)^{-1/2} \times \cdots \times (1-2t)^{-1/2} \\ &= (1-2t)^{-n/2} \end{aligned}$$

que es justamente la mgf de una χ^2 con n grados de libertad.

Teorema 3 (Distribución de la varianza muestral). $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ entonces $\bar{X}$ y S^2 son independientes y $\frac{(n-1)S^2}{\sigma^2} \sim \chi(n-1)$.

Prueba. Sea $W = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2$. Luego:

$$\begin{aligned} W &= \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 \\ &= \sum_{i=1}^n \left(\frac{(X_i - \bar{X}) + (\bar{X} - \mu)}{\sigma}\right)^2 \qquad \text{(sumar y restar } \bar{X}\text{)} \\ &= \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 + \sum_{i=1}^n \left(\frac{\bar{X} - \mu}{\sigma}\right)^2 + 2\left(\frac{\bar{X} - \mu}{\sigma}\right)\sum_{i=1}^n (X_i - \bar{X}) \qquad \text{(expandir binomio)} \end{aligned}$$

Notemos que $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$, entonces $\sum_{i=1}^n X_i = n\bar{X}$. Luego, del tercer término de la expresión W, tenemos que $\sum_{i=1}^n (X_i - \bar{X}) = \sum_{i=1}^n X_i - \sum_{i=1}^n \bar{X} = n\bar{X} - n\bar{X} = 0$, o sea que el tercer término de la expresión es 0. Ahora, por aparte, recordemos que

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 \\ (n-1)s^2 &= \sum_{i=1}^n (X_i - \bar{X})^2 \qquad \text{(multiplicando ambos lados por } (n-1)\text{)}. \end{aligned}$$

O sea que

$$W = \frac{(n-1)s^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2} \tag{3}$$

Al ser X_i normales, sabemos que W es una suma de $N(0,1)$ al cuadrado, o sea, por el teorema 2, $W \sim \chi^2(n)$. También sabemos que $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ Ahora, $Z^2 = \frac{(\bar{X} - \mu)^2}{\sigma^2/n} = \frac{n(\bar{X} - \mu)^2}{\sigma^2} \sim \chi^2(1)$ por el resultado 5.

Ahora, sabemos la mgf de dos de los tres términos de la Ecuación (3). Ahora, igualando mgf en 3, sabiendo que $\bar{X}$ y s^2 son independientes[13], obtenemos.

[13] Este hecho no lo demostramos porque está por fuera del alcance de este curso, pero es posible demostrarlo.

$$
\begin{aligned}
m_W(t) &= m_{\frac{(n-1)s^2}{\sigma^2}}(t) \times m_{\frac{n(\bar{X}-\mu)}{\sigma^2}}(t) \\
(1-2t)^{-n/2} &= m_{\frac{(n-1)s^2}{\sigma^2}}(t)(1-2t)^{-1/2} \\
(1-2t)^{-n/2}(1-2t)^{-1/2} &= m_{\frac{(n-1)s^2}{\sigma^2}}(t) \\
m_{\frac{(n-1)s^2}{\sigma^2}}(t) &= (1-2t)^{-(n+1)/2}
\end{aligned}
$$

Que es justo la mgf de una $\chi^2(n-1)$, lo que demuestra el teorema.

Las inferencias que hemos hecho hasta ahora, sobre la media muestral, presuponen que conocemos la varianza real de la muestra, lo cual en la práctica rara vez se cumple. Con el resultado que acabamos de mostrar, y otro resultado que mostraremos sin prueba, podemos hacer inferencia sobre la media muestral sin conocer la varianza verdadera.

Resultado 6. Sea $Z \sim N(0,1)$ y $W \sim \chi^2(v)$. Si W y Z son indepedientes, $T = \frac{Z}{\sqrt{W/v}}$ es una distribución t-student con v grados de libertad, o $t(v)$

Una consecuencia directa del teorema 3 y el resultado 6 es:

Resultado 7.

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1),$$

donde $t(n-1)$ es la distribución t de Student[14] con $n-1$ grados de libertad.

Ya avanzamos bastante la teoría, necesaria para hacer inferencia sobre una muestra sin saber el σ^2 real. Puede que las demostraciones sean algo complicadas, pero aplicar los resultados es bastante sencillo. Miremos el siguiente ejemplo:

La resistencia a la tensión para un tipo de alambre está distribuida normalmente con media desconocida μ y varianza desconocida σ^2. Seis trozos de alambre se seleccionan aleatoriamente de un rollo largo; Y_i, la resistencia a la tensión para el trozo i, se mide para $i = 1, 2, \ldots, 6$. La media poblacional μ y la varianza σ^2 pueden ser estimadas por $\bar{Y}$ y S^2, respectivamente. Como $\sigma_{\bar{Y}} = \sigma^2/n$ se deduce que $\sigma_{\bar{Y}}$ puede ser estimada por S^2/n. Encuentre la probabilidad aproximada de que $\bar{Y}$ esté dentro de $2S/\sqrt{n}$ de la verdadera media poblacional μ.

$$
\begin{aligned}
P\left(\frac{-2S}{\sqrt{n}} \leq \bar{Y}-\mu \leq \frac{2S}{\sqrt{n}}\right) &= \left(-2 \leq \sqrt{n}\left(\frac{\bar{Y}-\mu}{S}\right) \leq 2\right) \\
&= P(-2 \leq t \leq 2) \\
&= 0.8980
\end{aligned}
$$

Este último valor se puede hallar fácilmente con R, con la acumulada de una distribución t con 5 grados de libertad, usando el comando `pr <- pt(2, 5) - pt(-2, 5)`.

Exercise 8. *Un guardabosque, que estudia los efectos de la fertilización en ciertos bosques de pinos en el sureste, está interesado en estimar el promedio de área de la base de los pinos. Al estudiar áreas basales de pinos similares durante muchos años, descubrió que estas mediciones (en pulgadas cuadradas) están distribuidas normalmente con desviación estándar aproxima de 4 pulgadas cuadradas. Si el guardabosque muestrea $n = 9$ árboles, encuentre la probabilidad de que la media muestral se encuentre a no más de 2 pulgadas cuadradas de la media poblacional.* □

Exercise 9. *Suponga que, en el problema de fertilización del bosque, la desviación estándar poblacional de áreas basales no se conoce y debe estimarse a partir de la muestra. Si se ha de medir una muestra aleatoria de $n = 9$ áreas basales, encuentre dos estadísticos g_1 y g_2 tales que $P[g_1 \leq (\bar{Y}-\mu) \leq g_2] = .90$.* □

[14] Una pequeña nota histórica curiosa: está distribución fue descubierta por William Sealy Gosset, que la público con el pseudónimo de Student, ya que la compañía donde trabajaba, la cervecería Guiness en Dublin, no hacía públicas los descubrimientos internos. Gracias a este acto de rebeldía la ciencia a podido avanzar bastante y tenemos esta importante distribución.

4.2.1. Guía Bibliográfica

Leer 7.2 de Wackerly et al. (2010), 5.4 de Devore (2008) y 8.3-8.6 de Walpole (2007).

4.3. El teorema central del límite

Hasta ahora nos hemos concentrado en distribuciones de estadísticos cuando los datos son normales ¿Pero siempre podemos asegurar que los datos son normales? Es claro que no ¿Que hacemos entonces para saber la distribución de la media muestral cuando los datos no son normales, o no sabemos si son normales? Para eso usamos uno de los resultados más importantes de la estadística el teorema central del límite:

Teorema 4. (Teorema central del límite, CLT) Sean $X_1, X_2, \ldots, X_n$ independientes con $E[X_i] = \mu$ y $Var[X_i] = \sigma_2 < \infty$. Si

$$Z = \lim_{n\to\infty} \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

entonces $Z \sim N(0, 1)$.

La demostración de este teorema requiere de algunos teoremas de cálculo (el teorema de Taylor), queda por fuera de el alcance de este curso. ¡Este es un resultado demasiado notable! Me dice que si tengo un montón de variables aleatorias con la misma media y varianza (pero con posiblemente diferentes distribuciones), y las sumo, esto va a tender a una normal! Esto nos permite hacer inferencia con las técnicas que aprendimos en la sección pasada, sin si quiera conocer la distribución de nuestros datos.

Veamos el teorema central del límite en práctica. Por ejemplo, digamos que tenemos unos datos $X_1, X_2, \ldots X_{10000} \sim exp(1)$, o sea, distribuidos exponencialmente con media 1. El histograma de estos datos se puede ver en la figura 2.

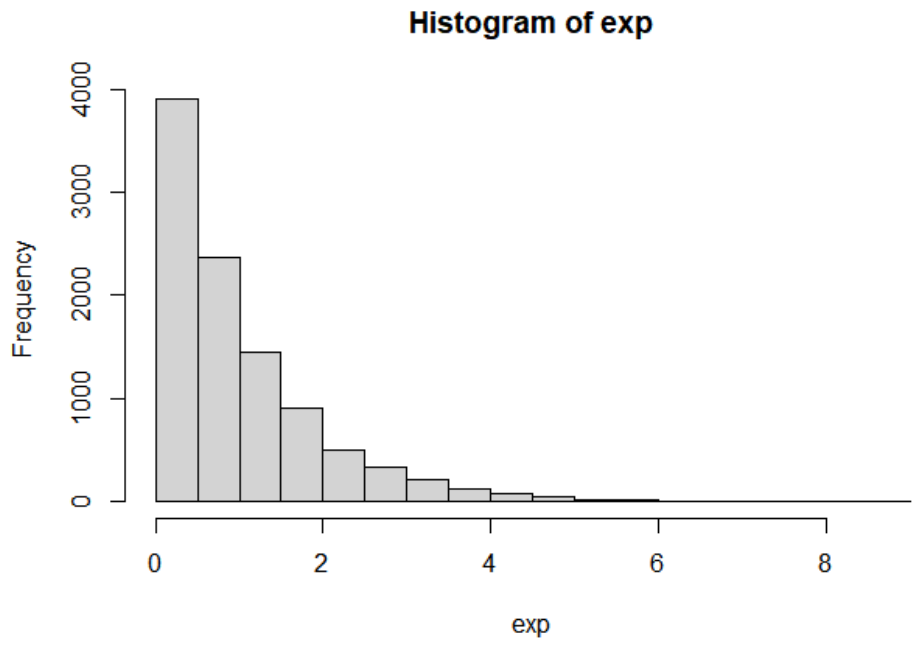


Figura 2: Histograma de muestra exponencial

Vemos que este histograma es bastante diferente de el de una normal (la campana de Gauss). ¿Pero entonces como se distribuye la suma de esos 10000 valores exponenciales? El teorema central del límite nos dice que se distribuye aproximadamente normal. Miremos el histograma de la suma de estos valores en la Figura 3.

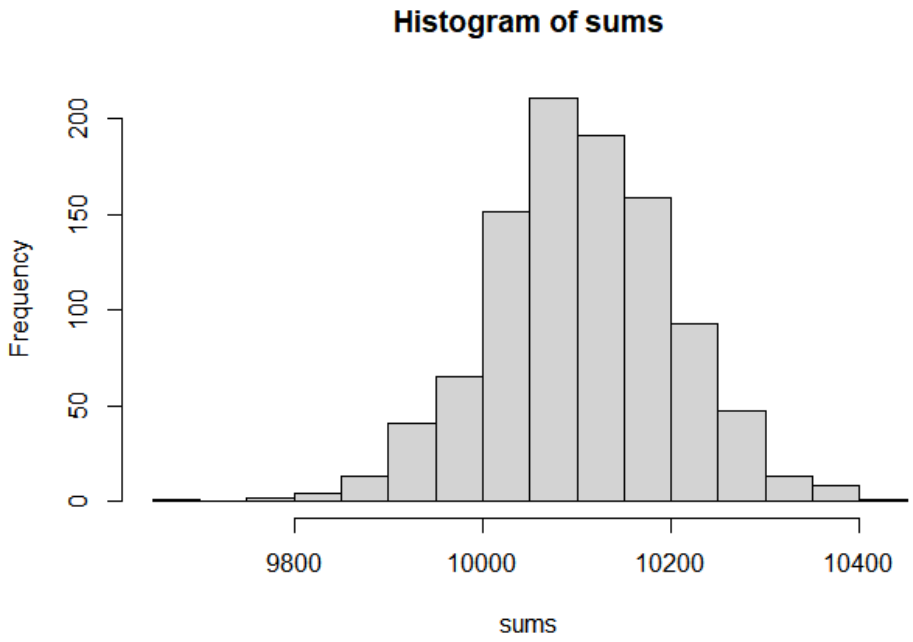


Figura 3: Histograma de la suma de exponenciales

Se parece bastante a una normal! O sea, sumando datos independientes de una distribución exponencial, que no se parece en nada a una normal, llegamos a una normal. Lo poderoso del Teorema central del límite es que funciona sea cual sea la distribución de los datos originales, siempre y cuando sean independientes y tengan la misma media y varianza. Una regla empírica: Si $n \geq 30$, podemos usar el teorema central del límite.

Ahora veamos el teorema en acción con un ejemplo. Sean Sean $X_1, X_2, ..., X_{100}$ los pesos netos reales de 100 sacos de 50 lb de fertilizante seleccionados al azar. Hallar $P(49.9 \leq \bar{X} \leq 50.1)$ si el peso esperado es 50 y la varianza es 1.

Como no sabemos la distribución de los X_i, no sabemos cual es la distribución $\bar{X}$ y no podríamos calcular esa probabilidad. Sin embargo, $n = 100 \geq 30$, o sea que, aproximadamente, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. O sea:

$$\begin{aligned} P(49.9 \leq \bar{X} \leq 50.1) &= P\left(\frac{49.9-50}{1/\sqrt{100}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{50.1-50}{1/\sqrt{100}}\right) \\ &= P(-1 \leq Z \leq 1) \\ &= 0.68 \end{aligned}$$

Exercise 10. *Suponga que $X_1, X_2, ..., X_n$ y que $Y_1, Y_2, ..., Y_n$ son muestras aleatorias independientes de poblaciones con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2. Por el teorema central del límite, la variable $Z = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2+\sigma_2^2/n}} \sim N(0,1)$. Teniendo esto en cuenta, se diseña un experimento para determinar si el operador A o el operador B obtienen el trabajo de operar una nueva máquina. Se toma el tiempo a cada operador en 50 intentos independientes que comprenden la realización de cierto trabajo usando la máquina. Si las medias muestrales para los 50 intentos difieren en más de 1 segundo, el operador con el menor tiempo medio obtiene el trabajo. De otro modo, el experimento es considerado como terminado en empate. Si las desviaciones estándar de los tiempos para ambos operadores se suponen de 2 segundos, ¿cuál es la probabilidad de que el operador A obtenga el trabajo aun cuando ambos operadores tengan igual capacidad?* □

Exercise 11. *Leer la sección 7.5 de Wackerly et al. (2010) y hacer el ejercicio 7.73* □

4.4. La prueba del teorema central del límite

Si le interesa mirar la prueba del teorema central del límite, puede leer la sección 7.4 de Wackerly et al. (2010). Es interesante notar que para hacer esa demostración, asumen que las mgf de las variables existen. Eso no necesariamente pasa. Aún así, el teorema central del límite se puede demostrar sin esa condición, pero la demostración es más extensa y complicada. También, se puede desechar el supuesto de independencia: ¡el teorema central del límite funciona también para sumas de algunos tipos de variables dependientes! Este teorema es muy poderoso y fundamental para la estadística. Es quizás su resultado más importante.

Exercise 12. *La mgf de una variable lognormal en general no existe, pero tiene media y varianza finita. Vamos a ver que el teorema central del límite funciona acá. Primero, genere* 10000 *datos de una distribución lognormal, consulte* `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Lognormal.html` *para ver como se generan datos log-normales. Dibuje el histograma de esos* 10000 *datos. Ahora, usando remuestreo, calcule* 1000 *veces la media. Dibuje el histograma de esas* 1000 *medias. ¿A qué se le parece esa distribución?* □

4.4.1. Guía Bibliográfica

Para más información sobre el teorema central del límite, pueden leer las secciones 7.3-7.5 de Wackerly et al. (2010).

5. Estimadores

Un estimador es una regla, sacada de la muestra, que nos permite conocer algo sobre la población. Ejemplos de estimadores vistos en esta clase son $\bar{X}$ y S^2, que nos permiten saber cosas sobre los parámetros poblacionales μ y σ^2. A estos estimadores que son un número único los llamamos estimadores puntuales.

5.1. Sesgo y error cuadrático medio

Siendo un poco más generales, llamemos al parámetro poblacional con el que queremos trabajar como θ. Ahora, el estimador correspondiente a ese parámetro se le llama $\hat{\theta}$, que se le llama θ estimado. Una propiedad interesante de los estimadores es su sesgo. Decimos que un estimador es insesgado si se cumple que

$$E[\hat{\theta}] = \theta,$$

o sea, el valor esperado de nuestra estimación es justamente el parámetro poblacional. El sesgo de un estimador se puede calcular con

$$B[\hat{\theta}] = E[\hat{\theta}] - \theta.$$

Se usa la letra B por que en inglés sesgo se dice *bias*. Es claro que si un estimador es insesgado luego tiene sesgo 0.

Por ejemplo, mostremos que el estimador S^2 es insesgado. Primero, recordemos que en la Sección 2 mostramos que para cualquier variable aleatoria X con media μ y varianza σ^2, $E[X^2] = \mu^2 + \sigma^2$. Una consecuencia de esto es que $E[\bar{X}^2] = \mu^2 + \sigma^2/n$. Ahora

$$\begin{aligned}
E[s^2] &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&= \frac{1}{n-1}\left\{E\left[\sum_{i=1}^{n} X_i^2\right] + nE[\bar{X}^2] - 2E[\bar{X}n\bar{X}]\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}(\sigma^2 + \mu^2) + n(\sigma^2/n + \mu^2) - 2n(\sigma^2/n + \mu^2)\right\} \\
&= \frac{1}{n-1}\{n\sigma^2 + n\mu^2 + \sigma^2 + n\mu^2 - 2\sigma^2 - 2n\mu^2\} \\
&= \frac{1}{n-1}\{n\sigma^2 - \sigma^2\} \\
&= \frac{1}{n-1}(n-1)\sigma^2 \\
&= \sigma^2,
\end{aligned}$$

lo que demuestra que el estimador es insesgado.

Exercise 13. *Muestre que el estimador $\bar{X}$ es insesgado para estimar μ en una población $X_1, X_2, \ldots, X_n \sim N(0,1)$* □

Otro concepto importante es el de error cuadrático medio, que también se le puede sacar a cualquier estimador, llamado usualmente MSE[15]. Esta definido por

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + (B(\hat{\theta}))^2 \tag{4}$$

en general, mientras menos error cuadrático tenga un estimador, mejor va a ser. Por ejemplo, si tenemos dos estimadores diferentes para un mismo parámetro poblacional, podemos comparar sus errores cuadráticos medios para ver cual es mejor. Como ejemplo, tomemos dos estimadores del párametro μ, y comparemos

[15] Por sus siglas en inglés, *mean squared error*.

sus errores cuadráticos medios. Tenemos una muestra normal $\{X_1, X_2, \ldots, X_9\} \sim N(0,1)$. Consideremos los estimadores $\frac{X_1+X_2}{2}$ y $\bar{X}$ Sabemos, por el ejercicio 13, que $B(\bar{X})$ es insesgado. También sabemos que su varianza es $\sigma^2/n = 1/9$. Solo nos falta hallar el sesgo y la varianza del otro estimador para poder encontrar su MSE. Primero el sesgo:

$$\begin{aligned} B\left(\frac{X_1+X_2}{2}\right) &= E\left(\frac{X_1+X_2}{2}\right) - \mu \\ &= \frac{1}{2}(E[X_1] + E[X_2]) - \mu \\ &= \frac{1}{2}(\mu + \mu) - \mu \\ &= \frac{1}{2}(2\mu) - \mu \\ &= \mu - \mu \\ &= 0 \end{aligned}$$

O sea que es insesgado. Ahora miremos la varianza:

$$\begin{aligned} Var\left(\frac{X_1+X_2}{2}\right) &= \frac{1}{4}(Var[X_1] + Var[X_2]) \\ &= \frac{1}{4}(1+1) \\ &= \frac{1}{4}(2) \\ &= \frac{1}{2} \end{aligned}$$

O sea que el usando la ecuación 4, el MSE de $\bar{X}$ es $1/9$, y el MSE de el otro estimador es $1/2$. Es claro que $1/9 < 1/2$, entonces el estimador $\bar{X}$ es mejor que el estimador $\frac{X_1+X_2}{2}$.

Exercise 14. *Halle el MSE de $\bar{X}$ cuando la población es n.* □

5.1.1. Guía bibliográfica

Pueden leer 8.1-8.4 de Wackerly.

5.2. Intervalos de confianza

Hasta ahora nos hemos concentrado en estimadores puntuales: de la muestra, me sacan un solo número. Estos nos interesan en tanto que nos ayudan a conocer más o menos como son los parámetros poblacionales. ¿Pero como podemos asegurar, con cierta probabilidad, que los parámetros poblacionales están cerca de nuestras estimaciones? Para eso usamos los intervalos de confianza. Un intervalo de confianza es, en esencia, un intervalo donde sé cual es la probabilidad de qué un parámetro poblacional este en ese intervalo. En general, un intervalo de confianza me dice es un intervalo tal que

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$$

O sea, me dice que la probabilidad de que mi parámetro poblacional, θ este entre dos cosas que estimé, $\hat{\theta}_L$ y $\hat{\theta}_U$, es de $1 - \alpha$. Obviamente, debemos estimar $\hat{\theta}_L$ y $\hat{\theta}_U$ de alguna manera. Vamos a ver como hacer esto en las próximas clases, dependiendo de cual sea el θ que queremos estimar (ejemplos de θ pueden ser la media poblacional, la varianza poblacional). ¿Entonces que es α? α es la probabilidad de que el parámetro poblacional no este en el intervalo. $1 - \alpha$ es entoces el nivel de confianza que tenemos. Entre más pequeño sea el nivel de confianza, más grande va a ser el intervalo de confianza. Un nivel de confianza del 5 % nos dice que hay una probabilidad del 95 % de que el parámetro que estoy estimando esté dentro del intervalo de confianza. Este valor es bastante común, y es el que generalmente escogemos en este curso, pero el ingeniero/científico/usuario de estadística debe escoger un valor de confianza apropiado para su problema. Por ejemplo, si estamos estimando que tan bien funciona un avión, decir que el ala funciona bien el 95 % de las veces probablemente no sea adecuado, y vamos a requerir de un α mucho más pequeño, por ejemplo, $\alpha = 0.00001\,\%$, para tener una probabilidad bastante alta de que el ala vaya a funcionar. En cambio, si estamos prediciendo ventas, probablemente no se vaya a quebrar la empresa si nos equivocamos sobre las ventas en un 10 %, entonces podemos escoger $\alpha = 10\,\%$.

Por ejemplo, sea θ un parámetro poblacional que queremos estimar. Tenemos que $\hat{\theta}$ está normalmente distribuido con media θ y desviación estándar $\sigma_{\hat{\theta}}$. Recordemos que esto es bastante común: ocurre si usamos un estadístico que es una suma de una muestra de variables aleatorias normales para estimar $\hat{\theta}$ o si usamos una muestra lo suficientemente grande, por el Teorema Central del Límite. Ahora, consideremos

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim N(0, 1)$$

entonces, hay un valor $z_{\alpha/2}$ (por simetría de la normal) tal que

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

para cualquier α. Para ver como es este $z_{\alpha/2}$, observe la Figura 4.

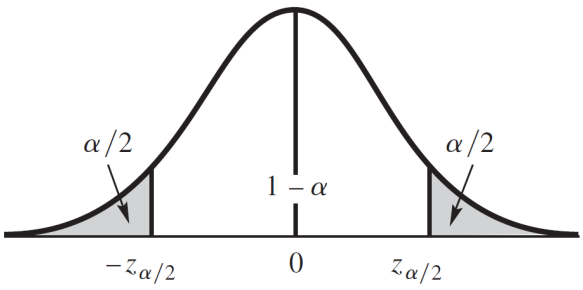


Figura 4: Área bajo la curva de una N(0,1) delimitada por los valores -$z_{\alpha/2}$ y $z_{\alpha/2}$

Juguemos un poco con esta expresión:

$$
\begin{aligned}
P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) &= P\left(-z_{\alpha/2} \leq \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \leq z_{\alpha/2}\right) \\
&= P(-z_{\alpha/2}\sigma_{\hat{\theta}} \leq \hat{\theta}-\theta \leq z_{\alpha/2}\sigma_{\hat{\theta}}) \\
&= P(-z_{\alpha/2}\sigma_{\hat{\theta}} - \hat{\theta} \leq -\theta \leq z_{\alpha/2}\sigma_{\hat{\theta}} - \hat{\theta}) \qquad \text{(restando } \hat{\theta}\text{)} \\
&= P(z_{\alpha/2}\sigma_{\hat{\theta}} + \hat{\theta} \geq \theta \geq -z_{\alpha/2}\sigma_{\hat{\theta}} + \hat{\theta}) \qquad \text{(multiplico todo por } -1\text{)} \\
1-\alpha &= P(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}) \qquad \text{(reorganizando desigualdades)}
\end{aligned}
$$

O sea, si tengo cualquier estimador $\hat{\theta}$ distribuido normal, puedo construir un intervalo de confianza con $1-\alpha$ probabilidad de que θ esté ahí con el límite inferior $\hat{\theta} - z_{\alpha/2}$ y el límite superior $\hat{\theta} + z_{\alpha/2}$!

Por ejemplo, sabemos que la media muestral $\bar{X}$ es un estimador con distribución normal (para datos normales o para datos con muestra $n \geq 30$). Entonces el intervalo de confianza con confianza $1-\alpha$ de la media muestral es $(\bar{X} - z_{\alpha/2}\sigma/n, \bar{X} + z_{\alpha/2}\sigma/n)$.

Un ejemplo práctico: Se registraron los tiempos de compra de n = 64 clientes seleccionados al azar en un supermercado local. El promedio y varianza de los 64 tiempos de compra fueron 33 minutos y 256 $minutos^2$, respectivamente. Estime μ, el verdadero promedio de tiempo de compra por cliente, con un coeficiente de confianza de $1-\alpha = .90$. O sea, debemos encontrar $\bar{X} \pm z_{\alpha/2}\sigma/n$, ya que la media muestral se distribuye normalmente con desviación estándar $\sigma/\sqrt{n}$ al ser $n = 64 \geq 30$. Notar que $\alpha = 0.1$. El intervalo es entonces:

$$
\begin{aligned}
\bar{X} \pm z_{\alpha/2}\sigma/n &= 33 \pm z_{0.1/2}\sqrt{\sigma/n} \\
&= 33 \pm z_{0.05}\sqrt{256/64} \\
&= 33 \pm 3.289707 \\
&= (29.71029, 36.28971)
\end{aligned}
$$

Para calcular $z_{0.95}$ puedo usar el comando de R, `qnorm(0.95)`. Notar que antes usábamos un valor de Z y R nos devolvía una probabilidad acumulada hasta este valor, usando `pnorm(z)`, que usa la cdf de la normal estándar. Ahora, le damos una probabilidad acumulada a R, y nos devuelve un valor de z. Esta es la función inversa de la probabilidad acumulada, llamada función cuantil.

Un pequeño aparte: si tenemos unos datos cargados en R (dígamos que están guardados en una variable llamada `x`) y queremos hallar fácilmente un intervalo de confianza del 95 % de la media muestra, simplemente corremos

```
tt <- t.test(x)
tt$conf.int
```

Si queremos otro nivel de confianza, digamos, queremos 99 % de confianza, corremos

```
tt <- t.test(x, conf.level = 0.99)
tt$conf.int
```

Exercise 15. *¿Cuál es la temperatura corporal normal para personas sanas? Una muestra aleatoria de 130 temperaturas corporales en personas sanas proporcionadas por Allen Shoemaker dio* 98.25 *grados y desviación estándar de* 0.73 *grados.* □

5.2.1. Guía Bibliográfica

Pueden leer 8.5-8.6 de Wackerly et al. (2010).

5.3. Selección de tamaño muestral

Muchas veces debemos diseñar un experimento para hacer inferencia sobre algún parámetro poblacional. Por ejemplo, queremos saber cual es el promedio real de altura en los estudiantes de EAFIT. Digamos que queremos saber, con un 99 % de confianza, que la media real va a estar en un intervalo dado. ¿Cuántos datos debemos coger? ¿Que n nos asegura esa probabilidad?

Más generalmente, supongamos que queremos saber, con una confianza de $1-\alpha$, que la media real y la media muestral solo diferirán en una cantidad B, podemos hallarlo mediante la igualdad

$$z_{\alpha/2}s/\sqrt{n} = B,$$

despejando n,

$$n = \frac{z^2_{\alpha/2}s^2}{B^2} \tag{5}$$

Este valor, en general, nos va a dar un número real. Pero no podemos muestrear un número real de individuos: los individuos son discretos. Para aliviar esto, aproxime a n por el entero que le sigue. Veamos un ejemplo:

Un servicio estatal de fauna silvestre desea calcular el número promedio de días que cada cazador con licencia se dedica a esta actividad realmente durante una estación determinada, con un límite en el error de estimación igual a 2 días de caza. Si los datos recolectados en estudios anteriores han demostrado que σ es aproximadamente igual a 10, ¿cuántos cazadores deben estar incluidos en el estudio?

Podemos usar directamente la ecuación 5:

$$n = \frac{z^2_{\alpha/2}\sigma^2}{B^2} = 96.03647 \approx 97$$

5.3.1. Guía Bibliográfica

Pueden leer 8.7 de Wackerly et al. (2010).

5.4. Intervalo de confianza para la desviación estándar muestral

Otro estimador muestral que hemos considerado bastante importante en este curso es el de la varianza o desviación estándar muestral. Nos interesara también saber, con una probabilidad dada, en que intervalo está nuestra media muestral, para por ejemplo saber que tan variables son nuestros datos.

Recordemos que el teorema 3 nos dice que

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

O sea que queremos encontrar unos valores de la distribución $\chi^2(n-1)$ tales que

$$P\left(\chi^2_L \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_U\right) = 1-\alpha$$

Para visualizar cuales serían estos valores, ver la Figura 5.

O sea, tenemos

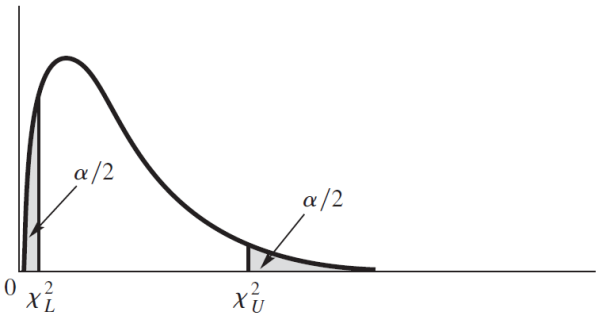


Figura 5: Área bajo la curva de una variable χ^2 delimitada por los valores -$\chi^2_{\alpha/2}$ y $\chi_{1-\alpha/2}$

$$
\begin{aligned}
1-\alpha &= P\left(\chi^2_{1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2}\right) \\
&= P\left(\frac{\chi^2_{1-\alpha/2}}{(n-1)s^2} \leq 1/\sigma^2 \leq \frac{\chi^2_{\alpha/2}}{(n-1)s^2}\right) \\
&= P\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \geq \sigma^2 \geq \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right) \qquad \text{(invirtiendo términos)} \\
&= P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) \qquad \text{(organizando)}
\end{aligned}
$$

Apliquemos esto a un ejemplo práctico en R. Encontremos un intervalo de confianza para la varianza poblacional y la desviación estándar poblacional.

Usando el siguiente código (donde `df` es una base de datos cargada previamente con los datos que `pizza_delivery.csv`, que pueden bajar de interactiva:

```
var_bill <- var(df$bill)
n <- dim(df)[1]
chisq_l <- qchisq(0.025, n-1)
var_bill_upper <- (n-1)*var_bill/chisq_l
chisq_u <- qchisq(0.975, n-1)
var_bill_lower <- (n-1)*var_bill/chisq_u
```

que nos dice que $\sigma^2 \in (116.6868, 136.3763)$ con 95 % de probabilidad. Para hallar el intervalo de confianza en el que está la desviación estándar, simplemente podemos sacarle la raíz cuadrada a estos valores, obteniendo que $\sigma \in (10.80216, 11.67803)$.

Exercise 16. *Para la misma base de datos de el ejemplo anterior,* `pizza_delivery.csv`*, halle el intervalo de confianza para la desviación estándar de la variable* `temperature`*.* □

5.4.1. Guía bibliográfica

Pueden leer la sección 8.9 de Wackerly et al. (2010).

5.5. Intervalos de confianza bootstrap

Hasta ahora, hemos encontrado intervalos de confianza para estimadores para los cuales conocemos la distribución. ¿Pero si no conocemos la distribución del estimador, que hacemos?

Podemos usar otra vez la técnica de remuestreo *bootstrap* que usamos para visualizar la distribución de estadísticos con distribución desconocida para nosotros. Simplemente debemos hacer esto: remuestrear B veces una muestra, y con eso calcular B veces el estadístico. Ahora, el límite inferior del intervalo de confianza será el número con posición (en orden ascendente) $\alpha/2 * B$ de nuestros estadísticos remuestreados. El límite superior será el dato con posición (en orden ascendente) $(1 - \alpha/2) * B$ de nuestros datos remuestreados. Un ejemplo en R, donde encuentro un intervalo de confianza para la mediana de unos datos:

```
medianas <- numeric(1000)
for (i in 1:1000){
  bill_resample <- sample(df$bill, replace = TRUE)
  medianas[i] <- median(bill_resample)
}
quantile(medianas, 0.025)
quantile(medianas, 0.975)
```

Exercise 17. *Encontrar un intervalo de confianza para el máximo de la variable* `bill`. □

5.5.1. Guía Bibliográfica

Para un estudio más detallado sobre intervalos de confianza bootstrap, pueden leer (DiCiccio and Efron, 1996).

5.6. Eficiencia Relativa

En las clases pasadas hemos propuesto diferentes estimadores, con los que podemos estimar un mismo parámetro poblacional. Vimos que podemos comparar esos estimadores de diferentes maneras, por ejemplo, comparando su MSE o su sesgo. En estás secciones daremos un análisis un poco más formal de algunos criterios de comparación de estimadores, y también algunas propiedades que deseamos que los estimadores tengan.

Recordemos que en clases pasadas, dijimos que si tenemos dos estimadores insesgados diferentes para θ, por ejemplo $\hat{\theta}_1$ y $\hat{\theta}_2$ preferimos el estimador que tenga menos varianza. Decimos que el estimador $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$ si $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$. De la misma manera, la eficiencia relativa de $\hat{\theta}_1$ con respecto a $\hat{\theta}_2$ se define como

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$$

Si $eff(\hat{\theta}_1, \hat{\theta}_2) > 1$, decimos que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$ y por ende, mejor estimador. Si $eff(\hat{\theta}_1, \hat{\theta}_2) < 1$ entonces $\hat{\theta}_2$ es más eficiente que $\hat{\theta}_1$ y por ende mejor estimador. Si $eff(\hat{\theta}_1, \hat{\theta}_2) = 1$ los dos estimadores tienen la misma varianza entonces ninguno es más eficiente que el otro.

Hagamos un ejemplo donde comparemos eficiencias de diferentes estimadores para el mismo parámetro poblacional.

Sea $Y_1, Y_2, ..., Y_n$ una muestra aleatoria de una población con media μ y varianza σ^2. Considere los siguientes tres estimadores para μ:

$$\begin{aligned}
\hat{\mu}_1 &= 1/2(Y_1 + Y_2) \\
\hat{\mu}_2 &= 1/4(Y_1) + \frac{Y_2 + \cdots + Y_{n-1}}{2(n-2)} + 1/4(Y_n) \\
\hat{\mu}_3 &= \bar{Y}
\end{aligned}$$

Primero, veamos si los estimadores son insesgados para ver si podemos calcular las eficiencias relativas[16]. Empezamos con $\hat{\mu}_1$.

$$\begin{aligned}
E[\hat{\mu}_1] &= E[1/2(Y_1 + Y_2)] \\
&= 1/2(E[Y_1 + Y_2]) \\
&= 1/2(E[Y_1] + E[Y_2]) \\
&= 1/2(\mu + \mu) \\
&= \mu
\end{aligned}$$

O sea que el estimador es insesgado. Ahora veamos como es el sesgo de $\hat{\mu}_2$.

$$\begin{aligned}
E[\hat{\mu}_2] &= E\left[1/4(Y_1) + \frac{Y_2 + \cdots + Y_{n-1}}{2(n-2)} + 1/4(Y_n)\right] \\
&= E[1/4(Y_1)] + E\left[\frac{Y_2 + \cdots + Y_{n-1}}{2(n-2)}\right] + E[1/4(Y_n)] \\
&= 1/4(E[Y_1]) + \frac{1}{2(n-2)} E[Y_2 + \cdots + Y_{n-1}] + 1/4(E[Y_n]) \\
&= 1/4(\mu) + \frac{(n-2)\mu}{2(n-2)} + 1/4(\mu) \\
&= 1/4(\mu) + 1/2(\mu) + 1/4(\mu) \\
&= \mu
\end{aligned}$$

[16]Esto es importante. La eficiencia relativa solo está definida para comparar estimadores insesgados. No podemos comparar un estimador insesgado con uno sesgado usado eficiencia. Tampoco podemos comparar dos estimadores sesgados.

Este estimador también es insesgado entonces. Por último, notar que $\hat{\mu}_3 = \bar{Y}$ es insesgado. La insesgadez de la media muestral la hemos mostrado ya en el curso. Luego, podemos comparar los tres estimadores. Busquemos las varianzas de los tres estimadores:

$$\begin{aligned} Var[\hat{\mu}_1] &= Var[1/2(Y_1 + Y_2)] \\ &= 1/4(Var[Y_1 + Y_2]) \\ &= 1/4(Var[Y_1] + Var[Y_2]) \\ &= 1/4(\sigma^2 + \sigma^2) \\ &= 1/2(\sigma^2) \end{aligned}$$

$$\begin{aligned} Var[\hat{\mu}_2] &= Var[1/4(Y_1) + \frac{Y_2 + \cdots + Y_{n-1}}{2(n-2)} + 1/4(Y_n)] \\ &= Var[1/4(Y_1)] + Var\left[\frac{Y_2 + \cdots + Y_{n-1}}{2(n-2)}\right] + Var[1/4(Y_n)] \\ &= 1/16(Var[Y_1]) + \frac{1}{4(n-2)^2} Var[Y_2 + \cdots + Y_{n-1}] + 1/16(Var[Y_n]) \\ &= 1/16(\sigma^2) + \frac{(n-2)\sigma^2}{4(n-2)^2} + 1/16(\sigma^2) \\ &= 1/16(\sigma^2) + \frac{\sigma^2}{4(n-2)} + 1/16(\sigma^2) \\ &= \frac{(n-2)\sigma^2 + 4\sigma^2 + (n-2)\sigma^2}{16(n-2)} \\ &= \frac{n\sigma^2 - 2\sigma^2 + 4\sigma^2 + n\sigma^2 - 2\sigma^2}{16(n-2)} \\ &= \frac{2n\sigma^2}{16(n-2)} \\ &= \frac{n\sigma^2}{8(n-2)} \end{aligned}$$

Notar que $Var[\hat{\mu}_3] = Var[\bar{Y}] = \sigma^2/n$. Ya tenemos todas las varianzas, podemos comparar eficiencias:

$$\begin{aligned} eff(\hat{\mu}_1, \hat{\mu}_2) &= \frac{Var(\hat{\mu}_2)}{Var(\hat{\mu}_1)} \\ &= \frac{\frac{n\sigma^2}{8(n-2)}}{\frac{\sigma^2}{2}} \\ &= \frac{2\sigma^2}{8\sigma^2(n-2)} \\ &= \frac{1}{4(n-2)} \end{aligned}$$

Si $n \geq 3$, entonces $eff(\hat{\mu}_1, \hat{\mu}_2) < 1$, entonces $\hat{\mu}_2$ es más eficiente que $\hat{\mu}_1$. Notar que por la definición de $\hat{\mu}_2$, n tiene que ser por lo menos 3, o sea que el $\hat{\mu}_2$ siempre es más eficiente que $\hat{\mu}_1$. Examinamos otra eficiencia:

$$
\begin{aligned}
eff(\hat{\mu}_2, \hat{\mu}_3) &= \frac{Var(\hat{\mu}_3)}{Var(\hat{\mu}_2)} \\
&= \frac{\frac{\sigma^2}{n}}{\frac{n\sigma^2}{8(n-2)}} \\
&= \frac{8\sigma^2(n-2)}{n^2\sigma^2} \\
&= \frac{8(n-2)}{n^2}
\end{aligned}
$$

Si $n \geq 3$, entonces $eff(\hat{\mu}_2, \hat{\mu}_3) < 1$, entonces $\hat{\mu}_3$ es más eficiente que $\hat{\mu}_2$. Luego $\hat{\mu}_3$ es el estimador más eficiente (entre estos tres) para μ.

Exercise 18. *Considere $Y_1, Y_2, Y_3 \sim exp(\theta)$. Consideremos los siguientes estimadores para θ: $\hat{\theta_1} = Y_1$, $\hat{\theta_2} = \frac{Y_1+Y_2}{2}$, $\hat{\theta_3} = \frac{Y_1+2Y_2}{3}$, $\hat{\theta_4} = \bar{Y}$. Halle las eficiencias relativas necesarias (para los estimadores para los cuales es posible). Diga cual es el mejor estimador basado en las eficiencias relativas que halló.* □

5.6.1. Guía bibliográfica

Pueden leer 9.1-9.2 de (Wackerly et al., 2010).

5.7. Consistencia

La consistencia es una propiedad de un estimador que me dice que pasa cuando la muestra crece y crece. ¿Que pasaría por ejemplo con la media muestral? ¿Que pasa con $\bar{X} = \frac{1}{n}\sum_{i=1^n} X_i$ cuando $n \to \infty$? Pues lo ideal es que, entre más datos tenga, más se parezca mi estimador $\bar{X}$ a lo que estoy intentando estimar, μ. ¿Pero qué significa que se parezcan $\bar{X}$ y μ? Podemos decir que se parecen cuando están muy cerquita en probabilidad, comparación que hemos hecho bastantes veces en este curso. La consistencia es entonces que estemos seguros de que μ y $\bar{X}$ están cerquita cuando $n \to \infty$. ¿Pero en estadística cuando está uno seguro de algo? Cuando la probabilidad de que eso ocurra es 1. Siendo un poquito más generales, consideremos un parámetro θ que vamos a estimar con $\hat{\theta}_n$ (o sea, $\hat{\theta}_n$ depende de n). El estimador $\hat{\theta}_n$ es consistente cuando, para cualquier número $\epsilon > 0$:

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \leq \epsilon) = 1$$

o de forma equivalente

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

Una forma equivalente de decir esto es decir que θ_n converge en probabilidad a θ.

Resultado 8. Un resultado que se puede verificar de está definición, usando el Teorema de Chevycheff (que no vimos en este curso, entonces no lo voy a demostrar) es que sí $\hat{\theta}_n$ es insesgado, entonces θ_n es consistente si

$$\lim_{n\to\infty} Var(\hat{\theta}_n) = 0$$

O sea, ahí tienen dos herramientas para ver si un estimador es consistente: una para estimadores insesgados, y una más general, para cualquier tipo de estimador.

Ahora, enunciemos otros resultados importantes de teoría asintótica[17]:

Resultado 9. Sea $\hat{\theta}_n$ un estimador consistente de θ y $\hat{\theta}'_n$ un estimador consistente de θ'. Entonces se cumplen:

[17]La teoría asintótica en estadística se refiere a cuando investigamos como se comporta una variable aleatoria cuando $n \to \infty$. Esto significa que por ejemplo, el teorema central del límite y la consistencia son conceptos de la teoría asintótica de la estadística

$i)\hat{\theta}_n + \hat{\theta}'_n$ es un estimador consistente de $\theta + \theta'$,

$ii)\hat{\theta}_n\hat{\theta}'_n$ es un estimador consistente de $\theta\theta'$,

$iii)$Si $\theta' \neq 0, \hat{\theta}_n/\hat{\theta}'_n$ es un estimador consistente de θ/θ',

$iv)$Si $g(\cdot)$ es continua entonces $g(\hat{\theta}_n)$ es un estimador consistente de $g(\theta)$.

Resultado 10. Suponga que U_n tiene una función de distribución que converge en una función de distribución normal estándar cuando $n \to \infty$. Si W_n converge en probabilidad a 1, entonces la función de distribución de U_n/W_n converge a la distribución normal estándar.

Usemos estos resultados para mostrar que si tengo una muestra $Y_1, \ldots, Y_2$ con media μ y varianza $\sigma^2 < \infty$, entonces $\sqrt{n}\left(\frac{\bar{Y}_n - \mu}{S_n}\right)$ converge a una distribución normal estándar.

En el tablero mostramos que S_n^2 converge en probabilidad a σ^2. Ahora, $g(x) = +\sqrt{x/c}$ es continua para x y c positivos, entonces por el resultado $9iv$ tenemos que $\sqrt{S_n^2/\sigma^2}$ converge en probabilidad a $\sqrt{\sigma^2/\sigma^2} = \sqrt{1} = 1$. Por el resultado 10, tenemos que la función de distribución de $\frac{\sqrt{n}\frac{\bar{Y}_n - \mu}{\sigma}}{S_n/\sigma} = \left(\frac{\bar{Y}_n - \mu}{S_n}\right)$ converge a una normal estándar.

Este resultado es bastante importante: nos dice cual es la distribución asintótica de $\left(\frac{\bar{Y}_n - \mu}{S_n}\right)$ sin importar la distribución de la muestra. En clases pasadas, habíamos visto que $\left(\frac{\bar{Y}_n - \mu}{S_n}\right)$ seguía una distribución t con $n-1$ grados de libertad. Lo que nos dice esto es que para n suficientemente grandes, podemos aproximar esa t con una normal estándar.

Exercise 19. *Muestre, computacionalmente, que $\bar{X}$ es un estimador consistente para μ. Genere $X_1, \ldots, X_n$ con dos distribuciones diferentes y muestre una gráficas que exhiba consitencia para las dos.* □

5.7.1. Guía bibliográfica

Leer 9.3 de (Wackerly et al., 2010).

5.8. Suficiencia

Hasta ahora hemos usado estimador que, intuitivamente, creemos que funciona (y hemos mostrado algunas propiedades buenas de estos, como la insesgadez y consistencia de $\bar{X}$ y s^2). Una de las funciones de los estimadores son resumirnos la muestra: tenemos un montón de datos, y sacamos conclusiones generales sobre la población usando no todos los datos, sino los estimadores de los parámetros poblacionales. ¿Pero como sabemos si mis estimadores me resumen correctamente la muestra?

Para esto se usa la suficiencia. Consideremos una muestra $Y_1, Y_2, ..., Y_n$ de una distribución de probabilidad con parámetro desconocido θ. Entonces se dice que el estadístico $U = g(Y_1, Y_2, \ldots, Y_n)$ es suficiente para θ si la distribución condicional de $Y_1, Y_2, \ldots, Y_n$, dado U, no depende de θ.

Lo que significa esto intuitivamente es que si usamos un U para estimar θ, pero ese U no depende para nada de como sea θ, entonces ningún otra función de la muestra (otro estadístico) proporcionará más información de como funciona θ. En ese sentido, U contiene toda la información posible de θ, o sea, es bueno para resumir θ, o en otras palabras, es suficiente para resumir enteramente a θ.

Esta condición es bastante difícil de comprobar en la práctica, entonces vamos a usar otra condición que es equivalente. Para eso, primero tenemos que definir la verosimilitud de la muestra: Sean $x_1, \ldots, x_n$ datos independientes muestreados de unas variables $X_1, \ldots, X_n$, con densidades $f(X_1|\theta), \ldots, f(X_n|\theta)$ que dependen de un parámetro θ. La verosimilitud entonces es la densidad conjunta de $x_1, \ldots, x_n$, o sea,

$$L(x_1, \ldots, x_n|\theta) = f(x_1|\theta) \times \cdots \times f(x_n|\theta) = L(\theta)$$

Ahora, esto nos permite enunciar el siguiente resultado:

Resultado 10. Sea $U = f(x_1, \ldots, x_n)$ un estadístico de la muestra. Luego U es suficiente si $L(\theta)$ se puede factorizar en dos funciones no negativas

$$L(\theta) = g(U, \theta)h(x_1, \ldots, x_n)$$

donde $g(U, \theta)$ depende solamente de U y θ, y $h(x_1, \ldots, x_n)$ no depende de θ.

Ilustremos el resultado 10 con un ejemplo: Sea $Y_1, Y_2, ..., Y_n$ una muestra aleatoria en la que Y_i posee la función de densidad de probabilidad

$$f(y_i|\theta) = \begin{cases} (1/\theta)e^{-y_i/\theta} & \text{si } 0 \leq y_i < \infty \\ 0 & \text{de otro modo} \end{cases}$$

Hallemos la verosimilitud:

$$\begin{aligned} L(\theta) =& f(x_1|\theta) \times \cdots \times f(x_n|\theta) \\ &= (1/\theta)e^{-y_1/\theta} \times \cdots \times (1/\theta)e^{-y_n/\theta} \\ &= \frac{e^{-\sum y_i/\theta}}{\theta^n} \\ &= \frac{e^{-n\bar{Y}}/\theta}{\theta^n} \end{aligned}$$

Si escogemos $g(\bar{Y}, \theta) = \frac{e^{-n\bar{Y}}/\theta}{\theta^n}$ y $h(y_1, \ldots, y_n) = 1$, tenemos la factorización necesaria para el resultado 10, o sea, $\bar{Y}$ es suficiente para θ.

Exercise 20. *Sea $Y_1, Y_2, ..., Y_n$ una muestra aleatoria de una distribución normal con media μ y varianza σ^2. Si μ es desconocida y σ^2 conocida, muestre que $\bar{Y}$ es suficiente para μ. Ahora, si μ es conocida y σ^2 desconocida, muestre que $\sum_{i=1}^{n}(Y_i - \mu)^2$ es suficiente para σ^2.* □

5.8.1. Guía bibliográfica

Leer 9.4 de Wackerly et al. (2010).

6. Métodos de estimación

6.1. Estimadores insesgados de mínima varianza

Con todo lo que hemos discutido hasta ahora sobre estimadores, lo que preferiríamos es que tuviéramos un estimador insesgado, consistente, suficiente y con la varianza más chiquita posible. Para encontrar un estimador con la mayoría de estas propiedades, podemos usar uno de los teoremas más importantes en la teoría de la estimación:

El teorema de Rao-Blackwell. Sea $\hat{\theta}$ un estimador insesgado para θ, tal que $Var(\hat{\theta}) < \infty$. Si U es suficiente para θ, definamos $\hat{\theta}^* = E(\hat{\theta}|U)$. Entonces, para todo θ, se cumple que $E(\hat{\theta}^*) = \theta$ y $Var(\hat{\theta}^*) \leq Var(\hat{\theta})$.

Primero miremos intuitivamente que nos dice este resultado. Si tenemos un estimador insesgado $\hat{\theta}$ y un estadístico suficiente U, podemos mejorar a $\hat{\theta}$ (convertirlo en otro estimador con menor o igual varianza, que sea también insesgado) usando U. En general, para las distribuciones y estadísticos que consideramos en este curso, los estadísticos U que consideramos en este curso nos asegurar que el $\hat{\theta}^*$ que encontramos en función de U y $\hat{\theta}$ son los estadísticos insesgados de mínima varianza para θ. O sea, son los *mejores* estadísticos posibles para θ. Estos estimadores son llamados MVUE (*Minimum Variance Unbiased Estimator)*. Ahora, dos resultados que necesitamos para la demostración del teorema:

Resultado 11. Si Y_1 y Y_2 son dos variables aleatorias, entonces $E(Y_1) = E[E(Y1|Y2)]$.

Resultado 12. Si Y_1 y Y_2 representan variables aleatorias, entonces $Var(Y_1|Y_2) = E(Var(Y_1|Y_2)) + Var(E(Y_1|Y_2))$.

Demostración: Como U es suficiente para θ, la distribución condicional de cualquier estadístico (incluyendo $\hat{\theta}$), dada U, no depende de θ. Entonces, en particular, $\hat{\theta}^* = E(\hat{\theta}|U)$ no depende de θ, y por tanto es solo función de la muestra, o sea, es un estadístico.

$$\begin{aligned} E(\hat{\theta}^*) &= E(E(\hat{\theta}^*|U)) \qquad \text{Resultado 11} \\ &= E(\hat{\theta}) \qquad U \text{ no afecta a } \hat{\theta} \\ &= \theta. \end{aligned}$$

O sea $\hat{\theta}^*$ es insesgado.

Luego, por resultado 12, tenemos que

$$\begin{aligned} Var(\hat{\theta}) &= Var(\hat{\theta}|U) \qquad U \text{ no afecta a } \hat{\theta} \\ &= E(Var(\hat{\theta}|U)) + Var(E(\hat{\theta}|U)) \\ &= Var(E(\hat{\theta}|U)) + E(Var(\hat{\theta}|U)) \qquad \text{intercambiando E y Var} \\ &= Var(\hat{\theta}^*) + E(Var(\hat{\theta}|U)) \qquad \text{definición de } \hat{\theta}^* \end{aligned}$$

Como la varianza nunca es negativa, $E[Var(\hat{\theta}|U)] \geq 0$, o sea que $Var(\hat{\theta}) \geq Var(\hat{\theta}^*)$.

Un resultado que nos ayudará a hacer un ejemplo es este:

Resultado 12. Suponga que Y, tiene una distribución Weibull con parámetro θ. Entonces Y^2 sigue una distribución exponencial con parámetro θ.

Ahora, hagamos un ejemplo: Encontremos un estimador MVUE para una muestra $Y_1, \ldots, Y_n$ que sigue una Weibull con parámetro θ. Recordemos que la función de densidad de una Weibull es:

$$f(y_i|\theta) = \begin{cases} \frac{2y_i}{\theta} e^{-y_i^2/\theta} & \text{si } 0 \leq y_i < \infty \\ 0 & \text{de otro modo} \end{cases}$$

Para hallar el estimador MVUE, debemos hallar un estadístico suficiente. Para eso, usemos el criterio de factorización de la función de verosimilitud.

$$\begin{aligned} L(\theta) &= f(y_1|\theta) \times \cdots \times f(y_n|\theta) \\ &= \frac{2y_1}{\theta} e^{-y_1^2/\theta} \times \cdots \times \frac{2y_n}{\theta} e^{-y_n^2/\theta} \\ &= \left(\frac{2}{\theta}\right)^n e^{-\frac{1}{\theta}\sum_{i=1}^n y_1^2} \prod_{i=1}^n y_i \end{aligned}$$

Luego, tenemos una factorización $L(\theta) = g(\sum_{i=1}^n y_i^2, \theta)h(y_1, \ldots, y_n)$ con $h(y1, \ldots, y_n) = \prod_{i=1}^n y_i$ y $g(\sum_{i=1}^n y_i^2, \theta) = \left(\frac{2}{\theta}\right)^n e^{-\frac{1}{\theta}\sum_{i=1}^n y_i^2}$, o sea que $\sum_{i=1}^n y_i^2$ es un estadístico suficiente para θ.

Ahora, debemos encontrar un estimador insesgado que sea función de $\sum_{i=1}^n y_i^2$. Por el resultado 12, sabemos que $y_i \sim exp(\theta)$, entonces $E(Y_i^2) = \theta$, o sea que $E(\sum_{i=1}^n Y_i^2) = n\theta$. Luego $E(\frac{1}{n}\sum_{i=1}^n Y_i^2) = \theta$, entonces $\frac{1}{n}\sum_{i=1}^n Y_i^2$ es el estadístico MVUE que estamos buscando.

Exercise 21. *Sea $Y_1, Y_2, ..., Y_n$ una muestra aleatoria de una distribución normal con media μ y varianza 1. Muestre que el MVUE de μ^2 es $\hat{\mu}^2 = \bar{Y}^2 - 1/n$.* □

6.1.1. Guía Bibliográfica

Leer 9.5 de Wackerly et al. (2010).

6.2. Método de los momentos

El método de los momentos es uno de los métodos de estimación más antiguos. Viene de la idea intuitiva de que los momentos muestrales son buenas estimaciones para los momentos poblacionales. Recordemos que el k-ésimo momento poblacional de una variable aleatoria Y es denominado como $\mu_k = E(Y^k)$, y el k-ésimo momento muestral de una muestra $Y_1, \ldots, Y_n$ es $m_k = \frac{1}{n}\sum_{i=1}^n Y_i^k$.

El método de los momentos funciona así: supongamos que tenemos una variable con t parámetros diferentes. Para estimar esos t parámetros, plantee el sistema de t ecuaciones y t incógnitas (los parámetros poblacionales) $\mu_1 = m_1, \mu_2 = m_2, \ldots, \mu_t = m_t$. Para saber las estimaciones de los parámetros, despeje. Hagamos un ejemplo:

Una muestra aleatoria de n observaciones, $Y_1, Y_2, \ldots, Y_n$ se selecciona de una población en la que Y_i, posee una función de densidad de probabilidad uniforme en el intervalo $(0, \theta)$ donde θ es desconocida. Use el método de momentos para estimar el parámetro θ.

Sabemos que para una uniforme, $E[Y_i] = \mu = \theta/2$. El primer momento muestral es $\bar{Y}$. Entonces igualamos, y $\theta/2 = \bar{Y}$, entonces $\theta = 2\bar{Y}$ es un estimador para θ.

Exercise 22. *Sean $Y_1, Y_2, ..., Y_n$ variables aleatorias uniformes independientes y distribuidas idénticamente en el intervalo $(0, 3\theta)$. Deduzca el estimador del método de momentos para θ.* □

6.2.1. Guía Bibliográfica

Leer la 9.6 de Wackerly et al. (2010).

6.3. Método de máxima verosimilitud

La idea de la estimación por máxima verosimilitud también es bastante intuitiva. Digamos que tenemos la función de verosimilitud de una muestra $Y_1, \ldots, Y_n$. O sea, tenemos una función $L(Y_1, Y_2, \ldots, Y_n|\theta_1, \ldots, \theta_k) = f(y_1|\theta_1, \ldots, \theta_k) \times \cdots \times f(y_n|\theta_1, \ldots, \theta_k)$, y queremos escoger los valores de $\theta_1, \ldots, \theta_k$ de alguna manera basado en esta función. Ya tenemos las observaciones, ¿entonces como escogemos los $\theta_1, \ldots, \theta_k$? Pensemos. Ya que tenemos las observaciones $Y_1, Y_2, \ldots, Y_n$, sería inteligente escoger los parámetros $\theta_1, \ldots, \theta_k$ que hacen que mis observaciones sean más probables. O sea, escogemos $\theta_1, \ldots, \theta_k$ de tal manera que maximizan $L(\theta)$. Este es el método de máxima verosimilitud. Recordemos como se maximiza una función con respecto a una variable: se deriva respecto a ese variable, y se iguala esa derivada a 0, y se despeja el parámetro de interés. Un tip importante para las funciones de verosimilitud: generalmente son intratables, o sea, son muy difíciles de

derivar. Pero la función de log-verosimilitud (aplicarle un logaritmo a la función de verosimilitud) facilita mucho el problema. Un parámetro que maximice la función de log-verosimilitud también maximiza la función de verosimilitud. Primero, recordemos algunas propiedades esenciales de los logaritmos:

Propiedades de los logaritmos:

$$\begin{aligned} i)log(MN) &= log(M) + log(N) \\ ii)log(M/N) &= log(M) - log(N) \\ iii)log(M^p) &= plog(M) \end{aligned}$$

Veamos todo esto en práctica: Sean $Y_1, \ldots, Y_n \sim Gamma(\alpha, \theta)$ con $\alpha > 0$ conocida. Hallar el MLE $\hat{\theta}$ de θ. Hallemos la función de verosimilitud:

$$\begin{aligned} L(\theta) &= \frac{1}{\Gamma(\alpha)\theta^\alpha} y_1^{\alpha-1} e^{y_1/\theta} \times \cdots \times \frac{1}{\Gamma(\alpha)\theta^\alpha} y_n^{\alpha-1} e^{y_n/\theta} \\ &= \frac{e^{-\sum_{i=1}^n y_i/\theta}}{(\Gamma(\alpha)^n)\theta^{\alpha n}} \prod_{i=1}^n y_i^{\alpha-1} \end{aligned}$$

Lo que parece bastante difícil de derivar. Encontremos entonces la log-verosimilitud:

$$log(L(\theta)) = -\sum_{i=1}^n y_i/\theta - nlog\Gamma(\alpha) - n\alpha log(\theta) + (\alpha - 1)\sum_{i=1}^n log(y_i)$$

y derivamos con respecto a θ

$$dlog(L(\theta))/d\theta = \frac{-\sum_{i=1}^n y_i}{\theta^2} \times (-1) - \frac{n\alpha}{\theta}$$

e igualamos a 0:

$$\begin{aligned} 0 &= \frac{\sum_{i=1}^n y_i}{\theta^2} - \frac{n\alpha}{\theta} \\ &= \frac{\sum_{i=1}^n y_i - \theta n\alpha}{\theta^2} \\ 0 &= \sum_{i=1}^n y_i - \theta n\alpha \\ \theta n\alpha &= \sum_{i=1}^n y_i \\ \hat{\theta} &= \frac{\sum_{i=1}^n y_i}{n\alpha} \end{aligned}$$

Entonces $\hat{\theta} = \frac{\sum_{i=1}^n y_i}{n\alpha}$ es el estimador MLE para θ.

Exercise 23. *Sea $Y_1, Y_2, ..., Y_n$ una muestra aleatoria de la función de densidad dada por*

$$f(y_i|\theta) = \begin{cases} (1/\theta) r y_i^{r-1} e^{-y_i^r/\theta} & si\ \theta > 0, y_i > 0 \\ 0 & de\ otro\ modo \end{cases}$$

donde $r > 0$ es una constante conocida. Encontrar un estadístico suficiente para θ. Encontrar el MLE para θ. ¿Es el estimador encontrado por MLE el MVUE?

□

Exercise 24. *Suponga que $Y_1, Y_2, \ldots, Y_n$ constituyen una muestra aleatoria de tamaño n de una distribución exponencial con media λ. Encuentre un intervalo de confianza de $100(1-a)\,\%$ para $t(\theta) = \theta^2$. Hint: lea la sección 9.8 del libro de Wackerly et al. (2010).* □

Exercise 25. *Sea $Y_1, Y_2, \ldots, Y_n$ una muestra aleatoria de la función de densidad de probabilidad*

$$f(y_i|\theta) = \begin{cases} (1+\theta)y_i^{\theta} & si\ \theta > -1, 0 < y_i < 1 \\ 0 & de\ otro\ modo \end{cases}$$

□

Encuentre los estimadores por el método de los momentos (Hint: para esto debe encontrar $E[Y] = \int_0^1 f(y_i) y_i dy_i$). Encuentre el MLE de θ. Compare los estimadores. Para eso, halle la factorización de la función $L(\theta)$, y observe que uno de estos estimadores es función de un estadístico suficiente. Teniendo esto en cuenta, diga cual estimador es mejor.

6.3.1. Guía bibliográfica

Puede leer 9.7-9.8 de Wackerly et al. (2010).

7. Pruebas de hipótesis

7.1. Pruebas de media: muestras grandes

Recordemos que la idea de la estadística es hacer inferencia. En el caso de este curso, hacemos inferencia sobre parámetros aunque se puede hacer inferencia sobre cosas más generales, como conjuntos de parámetros o sobre distribuciones. Esa inferencia se puede dar de dos maneras: con estimaciones o con pruebas de hipótesis, que es el tema que veremos ahora. Las pruebas de hipótesis son lo que le da vida al método científico. Creemos que alguna teoría es real, y usamos los datos que recogemos para ver si tenemos evidencia a favor o en contra de esa teoría. Si los datos no concuerdan con la teoría, rechazamos la hipótesis. Las pruebas de hipótesis se usan en todos los campos donde la teoría y las observaciones se puedan comparar, lo cual es increíblemente amplio. En estadística tenemos parámetros poblacionales, y la idea en este curso es probar una hipótesis que se relacione al valor de uno o más de esos parámetros.

Ahora, seamos un poquito más concretos. Digamos que tenemos una muestra $Y_1, \ldots, Y_n$, que dependen de un θ. Consideremos un estimador $\hat{\theta}$ con distribución muestral normal. Recordemos que por el teorema central del límite, muchos estimadores tienen distribución asintótica normal (como $\bar{Y}$), entonces esta teoría que desarrollaremos es bastante general. Sea θ_0 un valor específico de θ que queremos probar. O sea, tenemos la hipótesis que $\theta = \theta_0$. Esto es lo que en estadística llamamos hipótesis nula, o H_0. Se escribe: $H_0 : \theta = \theta_0$. Pero también necesitamos una hipótesis alternativa, algo que aceptar si tenemos evidencia en contra de nuestra H_0. Las partes esenciales de una prueba estadística son el estadístico de prueba y una región de rechazo asociada. El estadístico de prueba (al igual que un estimador) es una función de las mediciones muestrales. La región de rechazo, que de aquí en adelante estará denotada por RR, específica los valores del estadístico de prueba para el cual la hipótesis nula ha de ser rechazada a favor de la hipótesis alternativa. Si, para una muestra particular, el valor calculado del estadístico de prueba cae en la región de rechazo RR, rechazamos la hipótesis nula H_0 y aceptamos la hipótesis alternativa H_a. Si el valor del estadístico de prueba no cae en la RR, aceptamos H_0. Una definición importante: Se comete un error tipo I si H_0 es rechazada cuando H_0 es verdadera. La probabilidad de un error tipo I está denotada por α. El valor de α se denomina tamaño o nivel de la prueba. Este valor generalmente se escoge a priori, dependiendo del contexto del problema. Recordemos en intervalos de confianza, cuando escogíamos un nivel de confianza α a priori. O sea, debemos escoger las veces que estamos dispuestos a rechazar H_0 aún cuando es verdadera. Estos procedimientos están relacionados por esos valores, y luego veremos que son muy similares en muchos sentidos, y en ciertos casos, buscan la misma cosa. Se comete un error tipo II si H_0 es aceptada cuando H_a es verdadera. La probabilidad de un error tipo II está denotada por β, que se llama la potencia de la prueba. En estadística, todo tiene su precio. Tenemos que aceptar que vamos a cometer algún error tipo I para poder cometer pocos errores tipo II. O sea, mientras más chiquito sea α, más grande será β.

Por ejemplo, las pruebas médicas están diseñadas como pruebas de hipótesis en principio. Consideremos una prueba para saber si una persona tiene covid o no, o sea, una prueba con H_0 : La persona tiene covid vs. H_a : La persona no tiene Covid. Acá, α debería ser lo más bajito posible (en medicina, a $1 - \alpha$ se le llama la especificidad de la prueba), lo cual va a afectar a β (que en medicina llaman la sensibilidad de la prueba). Cuesta mucho más equivocarse en H_a que en H_0: si le decimos a una persona que no tiene Covid, pero en realidad tiene Covid, entonces esa persona va a seguir su vida normal y es posible que contagie a más personas, lo cual no es beneficioso para el sistema de salud (y obviamente para las personas tampoco). Las pruebas de Covid deben tener entonces una sensibilidad grande, lo cual va a disminuir la especificidad. Las pruebas deberían entonces estar diseñadas de tal forma que es más fácil encontrarse con un falso positivo (persona que sin Covid que sale positivo en la prueba) que con un falso negativo (personas con Covid que salen negativas a la prueba). Hay que tener un compromiso entre esos dos valores: las pruebas infalibles no existen.

Volvamos a la prueba. Tenemos θ como parámetro. Tenemos el estadístico de prueba $\hat{\theta}$. Tenemos la hipótesis nula que $H_0 : \theta = \theta_0$. Tenemos la hipótesis alternativa que $H_a : \theta > \theta_0$. Luego nuestra región de rechazo es $RR = \{\hat{\theta} > k\}$ para un k dado. Una ilustración de esto está en la Figura 6.

¿Como encontramos la región de rechazo? ¿O sea, como elegimos k? Pues primero, debemos fijar un α del error tipo I, llamado también el tamaño de la prueba. Si H_0 es verdadera, tenemos que $\hat{\theta} \sim N(\theta_0, \sigma_{\hat{\theta}})$, entonces $k = \theta_0 + z_\alpha \sigma_{\hat{\theta}}$. Luego la región de rechazo es

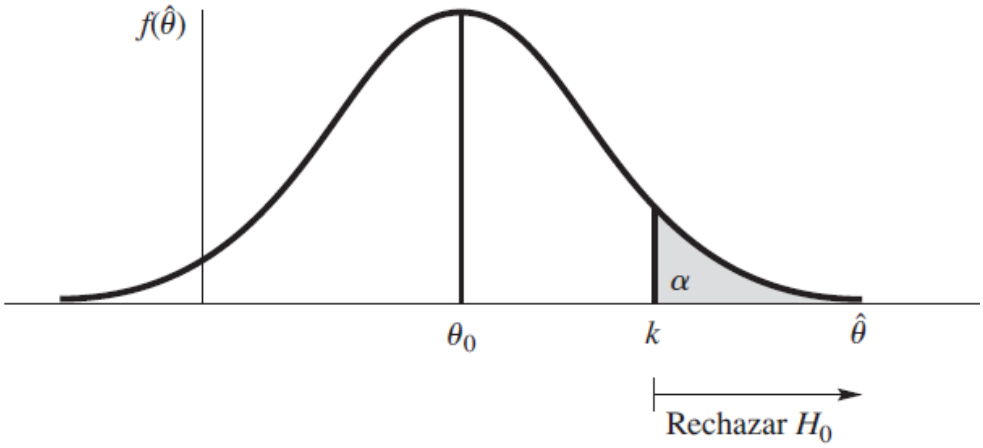


Figura 6: Ilustración prueba de hipótesis

$$\begin{aligned} RR &= \{\hat{\theta} : \hat{\theta} > \theta_0 + z_\alpha \sigma_{\hat{\theta}}\} \\ &= \left\{\hat{\theta} : \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} > z_\alpha\right\} \end{aligned}$$

O sea que podemos tomar $Z = \frac{\hat{\theta}-\theta_0}{\sigma_{\hat{\theta}}}$ como estadístico de prueba y rechazar si $Z > z_\alpha$.

Hagamos un ejemplo concreto. El vicepresidente de ventas de una gran empresa afirma que los vendedores están promediando no más de 15 contactos de venta por semana. (Le gustaría aumentar esta cantidad.) Como prueba de su afirmación, aleatoriamente se seleccionan n = 36 vendedores y se registra el número de contactos hechos por cada uno para una sola semana seleccionada al azar. La media y varianza de las 36 mediciones fueron 17 y 9, respectivamente. ¿La evidencia contradice lo dicho por el vicepresidente? Use una prueba con nivel $\alpha = 0.05$.

Estamos interesados en la hipótesis de que el vicepresidente está equivocado, o sea, nos interesa ver si $H_a : \mu > 15$, y nuestra hipótesis nula será que $H_0 : \mu = 15$. Sabemos que $\bar{Y} \sim N(\mu, \sigma/\sqrt{n})$, o sea que podemos elegir el estadístico de prueba

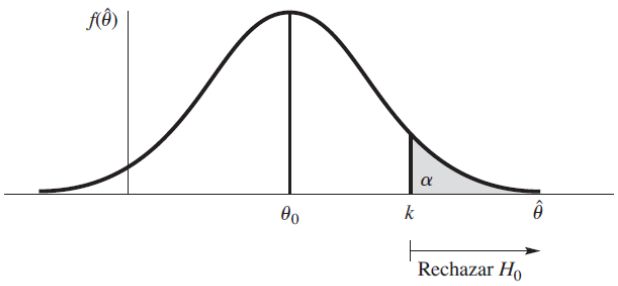
$$Z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1),$$

y tenemos la región de rechazo $RR = \{z > z_\alpha\} = \{z > 1.644854\}$. O sea que rechazaremos la hipótesis nula si nuestro estadístico de prueba es mayor que 1.644854. Ahora, calculemos el estadístico de prueba.

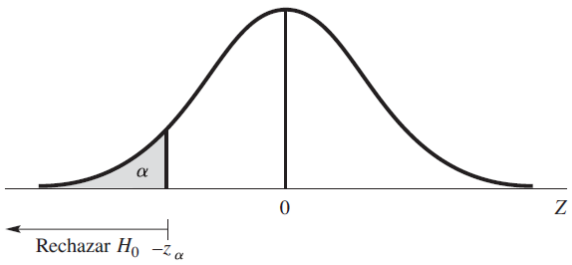
$$Z = \frac{17 - 15}{\sqrt{9/36}} = 4$$

Ahora, $4 > 1.644854$ se encuentra en la región de rechazo, entonces rechazamos $H_0 : \mu = 15$, Entonces, al nivel de significancia $\alpha = 0.05$, la evidencia es suficiente para indicar que la afirmación del vicepresidente es incorrecta y que el número promedio de contactos de ventas por semana es mayor que 15.

Hasta ahora solo hemos hecho pruebas del tipo $H_0 : \theta = \theta_0$ vs $H_a : \theta > \theta_0$, donde rechazamos la hipótesis nula para valores grandes de nuestro estadístico de prueba. También se pueden hacer pruebas de tipo $H_0 : \theta = \theta_0$ vs $\theta < \theta_0$. Acá se rechaza la hipótesis nula para valores pequeños del estadístico de prueba.

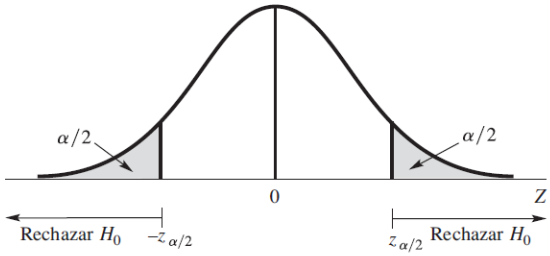


(a) $H_0 : \theta = \theta_0$ vs $H_a : \theta > \theta_0$



(b) $H_0 : \theta = \theta_0$ vs $\theta < \theta_0$

(a)



(b)

$H_0 : \theta = \theta_0$ vs $H_a : \theta \neq \theta_0$

Figura 7: Diferentes pruebas de hipótesis con estadísticos de prueba distribuidos normalmente.

Por último, podemos tener pruebas del tipo $H_0 : \theta = \theta_0$ vs $H_a : \theta \neq \theta_0$, donde se rechaza la hipótesis nula para valores muy grandes o muy pequeños de el estadístico de prueba. En resumen:

$$H_0 : \theta = \theta_0$$

$$H_a : \begin{cases} \theta > \theta_0, & \text{alternativa de cola superior} \\ \theta < \theta_0, & \text{alternativa de cola inferior} \\ \theta \neq \theta_0, & \text{alternativa de dos colas} \end{cases}$$

$$\text{Estadístico de prueba} : Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

$$\text{RR} : \begin{cases} \{Z > z_\alpha\}, & \text{RR de cola superior} \\ \{Z < -z_\alpha\}, & \text{RR de cola inferior} \\ \{|Z| > z_{\alpha/2}\}, & \text{RR de dos colas} \end{cases}$$

que podemos ver resumido en la Figura 7.

Exercise 26. *El voltaje de salida para un circuito eléctrico es de* 130. *Una muestra de 40 lecturas independientes del voltaje para este circuito dio una media muestral de* 128.6 *y desviación estándar de* 2.1. *Pruebe la hipótesis de que el promedio de voltaje de salida es* 130 *contra la alternativa de que es menor a* 130. *Use una prueba con nivel* $\alpha = 0.05$. □

Exercise 27. *En general, podemos considerar una prueba* $H_0 : \theta = 0$ *y* $H_A : \theta \neq 0$. *Digamos que tenemos una muestra* $-0.50728211, -0.54436458, 1.80807224, 0.98948854, 0.77142487, -0.90587370, -1.64153712$,

$-0.24158265, 1.46621748, -0.59944409, 1.15798962, 2.18176256, -1.01127618, -0.08012787, 0.19019399,$
$-1.42439202, 0.39362387, 0.76955192, -1.31770944, -1.09688725$ *con distribución desconocida, y que podemos estimar el parámetro con* $\hat{\theta} = \frac{1}{\sum_{i=1}^{n} Y_i^3}$*, que no sabemos como se distribuye. Aproxime la distribución de* $\hat{\theta}$ *con la técnica Bootstrap de remuestreo. Construya la región de rechazo usando la distribución remuestrada* 1000 *veces,* $\hat{\theta}^*$*. Ordene los valores de* $\hat{\theta}^*$*. La región de rechazo está entre* $\hat{\theta}^*_{1000\alpha/2}$ *y* $\hat{\theta}^*_{1000(1-\alpha/2)}$*, donde los subíndices se refieren a las posiciones de los* $\hat{\theta}^*$ *ordenados. ¿Rechaza o no la hipótesis de que* $H_0 : \theta = 0$*?* □

7.1.1. Guía Bibliográfica

Pueden leer 10.1-10.3 de Wackerly et al. (2010).

7.2. Pruebas de media: muestras pequeñas.

Recordemos que si tenemos una muestra $Y_1, Y_2, \ldots Y_n$ con $n < 30$, tenemos una muestra pequeña, y el estadístico

$$T = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

se distribuye t con $(n-1)$ grados de libertad. Ahora, podemos considerar una prueba similar a la de medias grandes, pero ahora usando los valores de la $t(n-1)$ para comparar con el estadístico de prueba. En resumen:

$$H_0 : \mu = \mu_0$$

$$H_a : \begin{cases} \mu > \mu_0, & \text{alternativa de cola superior} \\ \mu < \mu_0, & \text{alternativa de cola inferior} \\ \mu \neq \mu_0, & \text{alternativa de dos colas} \end{cases}$$

$$\text{Estadístico de prueba} : T = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

$$\text{RR} : \begin{cases} \{t > t_\alpha\}, & \text{RR de cola superior} \\ \{t < z_\alpha\}, & \text{RR de cola inferior} \\ \{|t| > t_{\alpha/2}\}, & \text{RR de dos colas} \end{cases}$$

El origen de la distribución t de Student se debe a un británico, William Sealy Gosset, bajo el pseudónimo Student. El trabajaba en la cervecería Guiness, de Dublín. Se dio cuenta que las pruebas de hipótesis con muestras grandes no funcionaban muy bien para las muestras pequeñas que el tenía, y desarrolló la teoría para hacer pruebas de medias con muestras pequeñas. Lo hizo bajo un pseudónimo porque Guiness no quería que sus competidores conocieran su método superior de pruebas de hipótesis para muestras pequeñas. Muy egoístas. Gracias a Dios Student era un científico con ganas de avanzar la sociedad con sus descubrimientos. Ahora, veamos un ejemplo.

Digamos que tenemos unas velocidades iniciales de ocho balas probadas con una nueva pólvora, junto con la media muestral y la desviación muestral estándar, $\bar{y} = 2959$ y $s = 39.1$. El fabricante dice que la nueva pólvora produce un promedio de velocidad de no menos de 3000 pies por segundo. ¿Los datos muestrales aportan suficiente evidencia para contradecir lo afirmado por el fabricante en el nivel de significancia de $\alpha = 0.025$?

Suponiendo que las velocidades iniciales están distribuidas normalmente en forma aproximada, podemos usar la prueba. Debemos probar $H_0 : \mu = 3000$ vs $H_a : \mu < 3000$. El estadístico que debemos usar es

$$\begin{aligned} T &= \frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t(n-1) \\ &= \frac{2959 - 3000}{39.1/\sqrt{8}} \\ &= -2.96587 \end{aligned}$$

Ahora el valor a comparar es $t_\alpha = -2.364624$. O sea que $T < t_\alpha$, o sea que cae en la región de rechazo, y rechazamos la hipótesis nula. O sea, no podemos concluir que la nueva pólvora produce un promedio de velocidad de no menos de 3000.

Ahora miremos como se comparan las medias de dos grupos con muestras pequeñas. Digamos que queremos comparar dos grupos, A y B, i.e., $Y_{1,A}, \ldots, Y_{n_A,A}$ y $Y_{1,B}, \ldots, Y_{1,n_B}$ a ver si tienen medias iguales o diferentes, asumiendo que tienen una varianza común σ^2. Esta prueba se llama la prueba t-test de Student. O sea, tenemos $H_0 : \mu_1 - \mu_2 = 0$ contra $H_a : \mu_1 - \mu_2 \neq 0$ con el estadístico de prueba

$$T = \frac{\bar{Y_A} - \bar{Y_B}}{\sqrt{s_A^2/n_A + s_B^2/n_B}}$$

que tiene distribución $t(n_A + n_B - 2)$. Podemos usar el valor de $t_{-\alpha/2}$ y $t_{\alpha/2}$ para ver si hay evidencia a favor o en contra de la hipótesis. Un ejemplo sencillo en R se muestra en clase y se deja en interactiva el código.

¿Pero que pasa si no sabemos que las varianzas de los dos grupos son iguales? Esta prueba se llama la prueba t-test de Welch. Debemos usar el mismo estadístico de prueba, pero con otros grados de libertad estimados por

$$df = \left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right) \Big/ \left(\frac{s_A^4}{n_A^2(n_B - 1)} + \frac{s_B^2}{n_B^2(n_A - 1)}\right).$$

Notar que df puede que no sea un número entero, entonces debemos aproximarlo por el próximo número entero. Un ejemplo de esto se muestra en R, ver archivo en interactiva y ejemplo en clase.

Exercise 28. *Sean* $38.9, 61.2, 73.3, 21.8, 63.4, 64.6, 48.4, 48.8, 48.5$ *unos pesos de mujeres con media real* μ_A. *Sean* $67.8, 60, 63.4, 76, 89.4, 73.3, 67.3, 61.3, 62.4$ *unos pesos de hombres con media real* μ_B. *Vamos a probar si los pesos de hombres y mujeres son diferentes, o sea, si* $H_0 : \mu_A = \mu_B$ *con* $H_A : \mu_A \neq \mu_B$. *Asuma que las varianzas de los dos conjuntos son iguales, calcule* T *y mire si está en la región de región de rechazo, mire si rechaza o no. Ahora, asuma que las varianzas son diferentes, y calcule el nuevo df y la nueva región de rechazo, mire si rechaza o no. ¿Que conclusiones puede sacar sobre el supuesto de varianzas iguales? ¿Si tiene unos datos cualquiera, de los cuales no sabe mucho, cual prueba usaría?* □

Exercise 29. *Asuma que tiene dos muestras con varianzas distintas, pero del mismo tamaño. Halle una versión simplificada de df para la prueba de Welch.* □

7.3. P-valores

Los p-valores son números asociados a una prueba estadística. Son valores que van de 0 a 1 (probabilidades), que me dicen que tan probable es conseguir los datos que tengo dado que mi hipótesis nula fue cierta. O sea, entre más chiquito sea el p-valor, menos probable es que los datos que tengo vengan de la hipótesis nula, y más evidencia hay en contra de esta. Para pruebas de hipótesis, calculamos un p-valor, y elegimos un nivel de significancia α (que, generalmente, es 0.05). Rechazamos la hipótesis nula si $p < \alpha$. Veamos un ejemplito.

Se realizó un estudio psicológico para comparar los tiempos de reacción de hombres y mujeres a un estímulo. En el experimento se emplearon muestras aleatorias independientes de 50 hombres y 50 mujeres. La media estimada para hombres es 3.6 y para mujeres es 3.8. La desviación estándar estimada para hombres es 0.18 y para mujeres es 0.14. ¿Los datos presentan evidencia para sugerir una diferencia entre los tiempos medios de reacción verdaderos para hombres y mujeres? Encuentre el p-valor, use un nivel de significancia de $\alpha = 0.05$

Podemos calcular el estadístico de prueba de esta prueba de medias, que es

$$\begin{aligned} Z &= \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2/n_2}} \\ &= \frac{3.6 - 3.8}{\sqrt{0.18/50 + 0.14/50}} \\ &= -2.5 \end{aligned}$$

Como tenemos una prueba de dos colas, el valor p es la probabilidad de que $Z \leq -2.5$ o $Z \geq 2.5$. Y $P(Z \geq 2.5) = P(Z \leq -2.5) = 0.0062$, entonces $p = 2 \times 0.0062 = 0.0124$. Al ser $\alpha = 0.05$ mayor que 0.0124, decidimos no rechazar H_0 en favor de H_a, entonces dedujimos que no hay evidencia de una diferencia de tiempo en reacciones para hombres y mujeres.

Los p-valores son de las herramientas más usadas (y abusadas) de la inferencia estadística. Hay artículos que hasta dicen, debido a el abuso tan grande que hay en la ciencia de p-valores, que la mayoría de los estudios científicos son falsos (Ioannidis, 2005). Algunos journals, en especial de psicología, han prohibido su uso, abogando por otros métodos como intervalos de confianza. ¿Pero si es sensato prohibir del todo esta herramienta? Algunos autores estadísticos no están de acuerdo, ver (Wasserstein et al., 2019), y promueven una filosofía que se reduce a las siglas ATOM:

- A: Accept uncertainty (acepta la incertidumbre).
- T: be toughtfull (se concienzudo).
- O: be open (se abierto y sincero).
- M: be modest (se modesto).

En resumen, cuando estés haciendo un análisis estadístico, no descartes los p-valores, son una medida útil. Pero no es el santo grial. Si te da un p-valor menor que 0.05 no cantes victoria y concluyas los resultados que querías concluir: el p-valor solo es el inicio del análisis. La asociación americana de estadísticos da las siguientes recomendaciones en cuanto a los p-valores

- Los p-valores pueden indicar como de incompatible es tu data con tu modelo estadístico específico, y ese no necesariamente es el único modelo existente.
- Los p-valores no miden la probabilidad que la hipótesis nula sea verdadera, o la probabilidad de que los datos fueron producidos solo por azar.
- Las conclusiones científicas o de negocio o de política no deben estar basadas solamente en p-valores.
- La inferencia requiere un reporte completo y transparente.
- Los p-valores no miden el tamaño de un efecto o la importancia de un resultado.

- Por si solos los p-valores no me dan una buena cantidad de evidencia con respecto a un modelo o a una hipótesis.

Otra cuestión: los p-valores muchas veces se interpretan como *efectos*, especialmente en estudios de economía, donde se usan modelos lineales. Cuando un p-valor es mayor que 0.05, los economistas tienden a concluir que hay efecto de una variable en otra. *Esto no es necesariamente verdad.* Tener cuidado con este tipo de conclusiones. Los p-valores se interpretan en contexto, no como una medida absoluta de evidencia.

Exercise 30. *Asuma que tenemos una prueba t con* $H_0 : \mu \leq 0$ *y* $H_a : \mu > 0$*. Simule unos datos normales de tamaño* 5*,* $N(0.5, 1)$ *(en R,* `x <- rnorm(5, 0.5, 1)`*. Haga la prueba de hipótesis* 10000 *veces distintas, con* 10000 *conjuntos de datos simulados diferentes. Haga la prueba de hipótesis (en R,* `t <- t.test(x, mu=0, alternative="greater")`*). Guarde los* 10000 *p-valores. ¿Cuantos de estos p-valores son menores que* 0.05*? ¿Que puede concluir de esto respecto a la verdad de* H_0 *y* H_a*?* □

7.3.1. Guía Bilbiográfica

Puede leer las secciones 10.6-10.7 de Wackerly et al. (2010).

7.4. Pruebas de varianza

Hasta ahora hemos hecho pruebas de hipótesis relacionadas a la media de una población dada. ¿Pero que pasa si queremos saber como se comporta la varianza? Digamos que tenemos una muestra normal $Y_1, \ldots, Y_n$ con media μ y varianza σ^2. ¿Qué hacemos si creemos que la varianza de nuestra muestra es igual a un valor dado, o sea, tenemos que $H_0 : \sigma^2 = \sigma_0^2$ vs $H_a : \sigma^2 > \sigma_0^2$? Pensemos que estadístico podríamos usar para construir esta prueba. Tiene que ser un estadístico que estime la varianza, y que conozcamos su distribución. El estadístico obvio sería s^2, la varianza muestral. Recordemos que

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Ahora, ¿como usamos esto para encontrar la prueba de hipótesis? Si H_0 es verdadera, tenemos que $\sigma^2 = \sigma_0^2$, entonces

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

Si H_a es verdadera y el valor real de σ^2 es mayor que σ_0^2, o se que esperaríamos que s^2, que estima el valor poblacional σ^2 sea mayor que σ_0^2. O sea, entre más grande s^2, más evidencia encuentro de que H_a es verdadera. Pero si s^2 es grande, también χ^2_{n-1} es grande. O sea, podemos rechazar la prueba con $RR = \{\chi^2 > \chi^2_\alpha\}$. Razonamientos similares me llevan a encontrar regiones de rechazo parecidas para $H_0 : \sigma^2 = \sigma_0^2$ vs $H_a : \sigma^2 < \sigma_2^2$ y para $H_0 : \sigma^2 = \sigma_0^2$ vs $H_a : \sigma^2 \neq \sigma_2^2$. En resumen, tenemos que:

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_a : \begin{cases} \sigma^2 > \sigma_0^2, & \text{alternativa de cola superior} \\ \sigma^2 < \sigma_0^2, & \text{alternativa de cola inferior} \\ \sigma^2 \neq \sigma_0^2, & \text{alternativa de dos colas} \end{cases}$$

$$\text{Estadístico de prueba} : \chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

$$\text{RR} : \begin{cases} \{\chi^2 > \chi^2_\alpha\}, & \text{RR de cola superior} \\ \{\chi^2 < \chi^2_{1-\alpha}\}, & \text{RR de cola inferior} \\ \{\chi^2 > \chi^2_{\alpha/2} \text{ o } \chi^2 < \chi^2_{1-\alpha/2}\}, & \text{RR de dos colas} \end{cases}$$

Hagamos un ejemplo: Un fabricante de cascos de seguridad para trabajadores de la construcción está interesado en la media y la varianza de las fuerzas que sus cascos transmiten a quienes los usan cuando se someten a una fuerza externa normal. El fabricante desea que la fuerza media transmitida por los cascos sea de 800 libras (o menos), bastante abajo del límite legal de 1000 libras, y desea que σ sea menor que 40. Se ejecutaron pruebas en una muestra aleatoria de $n = 40$ cascos, encontrando que la media muestral y la varianza eran iguales a 825 libras y 2350 $libras^2$, respectivamente. ¿Los datos aportan suficiente evidencia para indicar que σ^2 excede de 40?

Primero, tenemos que definir cual es σ_0^2. En este caso, nos dice el ejercicio que σ_0^2. Ahora, debemos plantear las hipótesis. Como queremos saber si la varianza real es mayor que 40, tenemos $H_0 : \sigma^2 = 40$ vs $H_a : \sigma^2 > 40$. Hallamos el estadístico de prueba correspondiente,

$$\begin{aligned} \chi^2 &= \frac{(n-1)s^2}{\sigma_0^2} \\ &= \frac{(40-1)2350}{40^2} \\ &= 57.28125 \end{aligned}$$

Ahora, debemos hallar el χ^2_α con el que nos compararemos, para una χ^2 con $n-1 = 39$ grados de libertad. En R, esto se puede hacer con `qchisq(0.05, 39)`. Esto da $\chi^2_\alpha = 25.69539$. Como $\chi^2 > \chi^2_\alpha$, mi χ^2 está en

la región de rechazo, por lo que podemos rechazar H_0 y decir que hay suficiente evidencia para pensar que $\sigma > 40$.

Exercise 31. *Un fabricante de máquinas para empacar jabón en polvo afirma que su máquina podría cargar cajas con un peso dado y una varianza de no más de* .01 *onzas. Se encontró que la media y la varianza de una muestra de ocho cajas fue de* 3.1 *y* .018*, respectivamente. Pruebe la hipótesis de que la varianza de la población de mediciones de peso es* $\sigma^2 = .01$ *contra la alternativa de que* $\sigma^2 > .01$*. ¿Que supuestos necesita sobre la muestra para hacer esta prueba? Utilice una significancia* $\alpha = 0.05$*.* □

Pero a veces, como con pruebas de medias, deseamos comparar las varianzas de dos grupos de datos con distribuciones normales. Supongamos que tenemos dos grupos, $Y_{1,1}, \ldots, Y_{1,n_1}$ y $Y_{2,1}, \ldots Y_{2,n_2}$ distribuidos normalmente, con medias desconocidas y varianzas σ_2^2 y σ_2^2, y deseamos probar $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_a : \sigma_1^2 > \sigma_2^2$. O sea, vamos a rechazar H_0 cuando σ_1^2 sea mucho más grande que σ_2^2. Esto lo podemos medir mediante el estadístico

$$F = \frac{S_1^2}{S_2^2}$$

Si S_1^2 es mucho más grande que S_2^2, F va a ser grande, entonces rechazamos la prueba para valores grandes de F. Pero resulta que bajo H_0 ese estadístico F sigue una distribución llamada la distribución F, con $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ grados de libertad en el denominador. O sea que la región de rechazo es $RR = \{F > F_\alpha\}$. Si queremos probar la alternativa $H_a : \sigma_1^2 < \sigma_2^2$, simplemente podemos intercambiar los índices arbitrarios de las poblaciones, o sea, cualquiera de los dos puede ser la población 1. Así, nuestra hipótesis alternativa vuelve a ser $H_a : \sigma_1^2 > \sigma_2^2$, y obtenemos la misma región de rechazo.

Para probar $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_a : \sigma_1^2 \neq \sigma_2^2$, hay que dar otro pasito. Sea F_b^a una variable aleatoria con distribución F que tiene a grados de libertad en el numerador y b grados de libertad en el denominador. Ahora, la región de rechazo de la prueba que estamos mencionando es $RR = \{F > F^{n_1-1}_{n_2-1,\alpha/2}$ o $F < (F^{n_2-1}_{n_1-1,\alpha/2})^{-1}\}$

Ahora, hagamos un ejemplo. Un experimento publicado en The American Biology Teacher estudió la eficacia de usar 95 % de etanol y 20 % de blanqueador como desinfectantes para eliminar contaminación por bacterias y hongos cuando se cultivan tejidos de plantas. El primer experimento se repitió 17 veces, y el segundo se repitió 15 veces, usando berenjenas como el tejido de planta cultivado. Las observaciones reportadas fueron el número de cortes de berenjena no contaminados después de 4 semanas de almacenamiento. La varianza de contaminados para etanol estimada es $S_1^2 = 2.78$ y para blanqueador es $S_2^2 = 0.17143$. ¿Qué concluiría usted, con $\alpha = .02$?

Primero, planteemos la hipótesis nula, que nos dice $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_a : \sigma_1^2 \neq \sigma_2^2$. Ahora, calculemos el estadístico de prueba

$$F = \frac{2.78}{0.17143} = 16.21653$$

Los valores con los que nos debemos comparar son $F^{16}_{14,0.01}$ y $(F^{14}_{16,0.01})^{-1}$. El primero, lo podemos sacar con el comando de R `qf(0.01, 16, 14)`, que es 0.2763437 y el segundo lo podemos sacar con `qf(0.01, 14, 16)^-1`, que es 3.618682. Ahora, rechazamos ya que $F > F^{16}_{14,0.01}$, o sea que hay evidencia de que las varianzas sean distintas.

Exercise 32. *Los exámenes de aptitud deben producir calificaciones con una gran cantidad de variación para que un administrador pueda distinguir entre personas con baja aptitud y otras de elevada aptitud. El examen estándar empleado por cierta industria ha producido calificaciones con desviación estándar de* 10 *puntos. Este exámen fue aplicado a* 23 *empleados Un nuevo examen se aplica a* 20 *posibles empleados y produce una desviación estándar muestral de* 12 *puntos. ¿Las calificaciones del nuevo examen son considerablemente más variables que las del examen estándar? Use* $\alpha = 0.01$ □

Exercise 33. *Considere los datos del ejercicio 28. Haga una prueba que mire que las varianzas de los pesos de mujeres es mayor a la varianza de los pesos de los hombres. Use R.* □

7.4.1. Guía Bibliográfica

Puede leer la sección 10.9 de Wackerly et al. (2010).

7.5. Relación de pruebas de hipótesis con intervalos de confianza

Consideremos una prueba de un parámetro normal con dos colas, o sea, $H_0 : \theta = \theta_0$ vs. $H_a : \theta \neq \theta_0$. Recordemos que debemos usar el estadístico

$$Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

y la región de rechazo $RR = \{|Z| > Z_{1-\alpha/2}\}$. Rechazamos cuando Z cumple eso. Equivalentemente, podemos decir que *no rechazamos* cuando

$$-z_{1-\alpha/2} \leq \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} \leq z_{1-\alpha/2}$$

O sea que H_0 no es rechazada a nivel α si

$$\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta_0 \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}$$

que es justamente el intervalo de confianza para un parámetro normal.

Ahora un ejemplo. Estamos midiendo resistencia al corte, obtenidas de pruebas de compresión no confinada para dos tipos de suelos. Para el suelo I, tenemos que $n_1 = 30$, $\bar{y}_1 = 1.65$ y $s_1 = 0.26$. Para el suelo II, tenemos que $n_2 = 35$, $\bar{y}_2 = 1.43$ y $s_2 = 0.22$. Construyamos un intervalo de confianza del 99 % para las diferencias en resistencias medias al corte para los dos tipos de suelo. ¿El valor $\mu_1 - \mu_2 = 0$ está en este intervalo? Con base en el intervalo, ¿debe ser rechazada la hipótesis nula de que $\mu_1 - \mu_2 = 0$?

El intervalo de $\mu_1 - \mu_2$ está dado por

$$\bar{y}_1 - \bar{y}_2 \pm z_{0.005}\sqrt{s_1^2/n_1 + s_2^2/n_2} = 1.65 - 1.43 \pm 2.56\sqrt{0.26^2/30 + 0.22^2/30} = 0.22 \pm 0.16$$

O sea, el intervalo de confianza es $(0.06, 0.38)$. El valor $\mu_1 - \mu_2 = 0$ no está en este intervalo, entonces rechazamos la hipótesis nula de que las resistencias son iguales con un $\alpha = 0.01$

Exercise 34. *Una muestra de* 40 *lecturas independientes del voltaje de un circuito dio una media muestral de* 128.6 *y desviación estándar de* 2.1. *Construya un intervalo de confianza para la media. ¿Aceptaría la hipótesis de que el voltaje es* 130*?* □

Exercise 35. *Haga un pequeño resumen de la sección 10.7 de (Wackerly et al., 2010).* □

7.5.1. Guía Bibliográfica

Puede leer la sección 10.5 de (Wackerly et al., 2010).

7.6. Potencia y Lema de Neyman-Pearson

Hasta ahora hemos considerado ciertas hipótesis que creemos son interesantes, y escogemos algún estadístico para el que hallamos ciertas regiones de rechazo basadas en que sabemos como se distribuyen esos estadísticos. ¿Pero como sabemos que elegimos bien esos estadísticos? Sabemos que el α lo escogemos a priori, entonces por lo menos tenemos esa garantía de que tenemos el error tipo I controlado de cierta forma. ¿Pero como sabemos si nuestras pruebas son buenas en potencia, i.e., que vamos a rechazar la hipótesis nula cuando es falsa?

Para esto, primero introduzcamos una definición de un concepto nuevo. Decimos que una hipótesis es **simple** si me define la población de una manera única, y decimos que es compuesta si no cumple esto. Por ejemplo, la hipótesis $H_0 : \mu = 3$ es simple, ya que me dice que la población está definida únicamente con el parámetro $\mu = 3$. En cambio la hipótesis $H_0 : \mu > 3$ es compuesta, ya que me da todos los posibles valores de $\mu > 3$, o sea, puedo definir la población con cualquiera de esos valores posibles, entonces es compuesta. Ya con esto podemos definir el Lema de Neyman-Pearson.

Lema de Neyman-Pearson. Suponga que deseamos probar la hipótesis nula simple $H_0 : \theta = \theta_0$ vs. la hipótesis alternativa simple $H_a : \theta = \theta_a$ ocn base en la muestra aleatoria $Y_1, \ldots, Y_n$ de una distribución con párametro θ. Sea $L(\theta)$ la verosimilitud de la muestra cuando el valor del párametro es θ. Entonces, para un α dado, la prueba que maximiza en potencia en θ_a tiene una región de rechazo determinada por

$$\frac{L(\theta_0)}{L(\theta_a)} < k$$

donde el valor de k se escoge de modo que la prueba tenga el valor deseado para el α escogido. Esta es la prueba de máxima potencia en el nivel α para H_0 vs. H_a.

Calcular los valores específicos de k para una prueba dada puede ser bastante tedioso. En clase muestro un ejemplo y pueden ver como de tedioso es. La idea de esto es simplemente que sepan, que dada cualquier prueba de hipótesis que tengan, es posible encontrar una región de rechazo que sabemos que es la mejor por el Lemma de Neyman-Pearson.

7.6.1. Guía Bibliográfica

Leer sección 10.10 de (Wackerly et al., 2010).

7.7. Pruebas de bondad y ajuste

Las pruebas que hemos hecho hasta ahora se han concentrado en ver si un parámetro específico de una distribución se comporta como creemos. Por ejemplo, los t-test miran el comportamiento de las medias, los F-test el de las varianzas, etc. ¿Pero como hacer para comparar distribuciones como tal? ¿Como podemos ver si nuestros datos siguen una distribución específica?

Consideremos un lanzamiento de un dado. Si el dado es justo (o sea, sin carga para un lado en específico), sabemos que las probabilidades de de que salga cualquier cara es 1/6. Si lanzamos el dado muchas veces, podemos comparar el número de cada cara esperado versus el número de veces que en verdad salió cada cara. Por ejemplo, pensemos que lanzamos un dado 120 veces. Cada cara tiene $120/6 = 20$ número de veces esperado que va a salir. Digamos que los resultados del dado para cada cara son $12, 25, 30, 11, 10, 32$. Resulta que si tenemos el estadístico

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

Donde o_i son las observaciones y e_i el número de ocurrencias esperadas para cada i. Este estadístico se distribuye chi cuadrado con $k - 1$ grados de libertad. Para nuestro caso, calculamos

$$\chi^2 = \frac{(12-20)^2}{20} + \frac{(25-20)^2}{20} + \frac{(30-20)^2}{20} + \frac{(11-20)^2}{20} + \frac{(10-20)^2}{20} + \frac{(32-20)^2}{20} = 25.7$$

Si calculamos $\chi^2_{0.95} = qchisq(0.95, 5) = 11.0705$, vemos que nuestro estadístico es mayor que el valor de referencia al 5 % entonces rechazamos la hipótesis nula de que el dado no esta cargado, hay evidencia a favor de que esté cargado.

Exercise 36. *Las calificaciones de un curso de estadística para un semestre específico fueron las siguientes:* 14 - *A,* 18 - *B,* 32 - *C, D* - 20, 16 - *E. Pruebe la hipótesis nula de que todas las calificaciones tienen la misma probabilidad, con un nivel de significancia de* 0.05. □

Para probar si unos datos dados tienen una distribución teórica dada, podemos usar la prueba de Kolmogorov-Smirnov. La prueba de Kolmogorov-Smirnov me sirve para probar si una muestra viene de una distribución particular o no. En resumen, mira que tan diferentes son la distribución acumulada empírica de los datos y la distribución acumulada teórica. Si son bastante diferentes, se rechaza la hipótesis nula de que los datos se distribuyen con la distribución teórica propuesta. Para usar esto, en R escribimos `ks.test(x, "pdist", params)`, donde pdist es el nombre de una distribución en R, x es el vector con nuestros datos, y params son los parámetros de pdist. Por ejemplo, si queremos ver si unos datos que tenemos se distribuyen $N(0, 1)$ podemos escribir `ks.test(x, "pnorm", 0, 1)`. Si queremos en general ver que nuestros datos se distribuyen normal, podemos usar la media y la desviación estándar estimada, o sea, podemos usar `ks.test(x, "pnorm", mean(x), sd(x))`. Si queremos ver por ejemplo si nuestros datos siguen una distribución Gamma con párametros 10 y 1/3 podemos escribir `ks.test(x, "pgamma", 10, 1/3)`. Podemos usar cualquier distribución que esté en R, que para cuestiones prácticas, son todas las que hemos visto en este curso.

Exercise 37. *Si tengo dos vectores de datos diferentes, y quiero ver si vienen de la misma distribución, puedo usar la prueba de Kolmogorov-Smirnov. En R es bastante sencillo, solo* `ks.test(x,y)`. *Haga la prueba en R de que los pesos de hombres y mujeres se distribuyen igual, con los datos del ejercicio 28. Reporte p-valor. ¿Que puede concluir de esto?.* □

7.7.1. Guía Bibliográfica

Pueden leer 10.11 de Walpole (2007) para la prueba con estadístico χ^2. Para la prueba de Kolmogorov-Smirnov, pueden ver el capítulo 6 de (Conover, 1971).

7.8. Prueba de independencia (datos categóricos)

En esta sección discutimos pruebas relacionadas a una tabla de contingencia. Una tabla de contingencia esta dada por

	Nivel de ingreso			
Reforma fiscal	Bajo	Medio	Alto	Total
A favor	182	213	203	598
En contra	154	138	110	402
Titak	336	351	313	1000

Esto se llama una tabla de contingencia 2×3, con 2 renglones y 3 columnas. En general, podemos hacer esta prueba para cualquier tabla $r \times c$, donde tengamos c categorías de cosas que pueden elegir entre r opciones. Podemos estimar la probabilidad de que cada evento ocurra (i.e., que una persona tenga nivel de ingreso bajo, o que este a favor de la reforma. Todas las probabilidades son:

- L: Una persona tiene ingresos bajos. $P(L) = 336/1000 = 0.336$.
- M: Una persona tiene ingresos medios. $P(M) = 351/100 = 0.351$.
- H: Una persona tiene ingresos altos. $P(H) = 313/1000 = 0.313$.
- F: Una persona está a favor de la reforma. $P(F) = 598/1000 = 0.598$.
- A: Una persona está en contra de la reforma. $P(A) = 402/1000 = 0.402$.

Si H_0 es verdadera, y en realidad la reforma fiscal y el nivel de ingreso son independientes, tenemos que:

$$\begin{aligned}
P(L \cap F) &= P(L)P(F) = 0.336 \times 0.598 = 0.2 \\
P(L \cap A) &= P(L)P(A) = 0.336 \times 0.402 = 0.135 \\
P(M \cap F) &= P(M)P(F) = 0.351 \times 0.598 = 0.209 \\
P(M \cap A) &= P(M)P(A) = 0.351 \times 0.402 = 0.141 \\
P(H \cap F) &= P(H)P(F) = 0.313 \times 0.598 = 0.187 \\
P(H \cap A) &= P(H)P(A) = 0.313 \times 0.402 = 0.126
\end{aligned}$$

Las frecuencias esperadas se pueden obtener multiplicando estas probabilidades totales por el número total de encuestados. Poniendo entre paréntesis estas frecuencias, obtenemos la tabla

	Nivel de ingreso			
Reforma fiscal	Bajo	Medio	Alto	Total
A favor	182 (200)	213 (209)	203 (187)	598
En contra	154 (135)	138 (141)	110 (126)	402
Titak	336	351	313	1000

Ahora, el número de grados de libertad que tiene la tabla está dado por $v = (r-1)(c-1)$ que en nuestro caso es $v = (2-1)(3-1) = 2$, y usamos el siguiente estadístico:

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

En nuestro caso:

$$\chi^2 = \frac{(182-200)^2}{200} + \frac{(213-209)^2}{209} + \frac{(203-187)^2}{187} + \frac{(154-135)^2}{135} + \frac{(138-141)^2}{141} + \frac{(110-126)^2}{126} = 7.85$$

Y comparando con $\chi^2_{0.95} = 5.91$ con $v = 2$, tenemos que nuestro estadístico de prueba es más grande que el valor de referencia, entonces obtenemos evidencia en contra de la hipótesis nula de independencia, o sea que la opinión del votante no es independiente a su nivel de ingresos.

Exercise 38. *Las prueas de independencia no funcionan muy bien cuando tenemos una tabla de contingencia* 2×2. *Para eso, se usa la versión corregida del chi-cuadrado:*

$$\chi^2 = \sum_i \frac{(|o_i - e_i| - 0.5)^2}{e_i}$$

Teniendo esto en cuenta, considere que tenemos una muestra aleatoria de 90 *adultos clasificados respecto al género y al número de horas dedicadas a ver redes sociales cada semana. Se dice que* 15 *hombres usan redes más de* 25 *horas a la semana, y* 27 *usan redes menos de* 25 *horas a la semana. Para mujeres,* 29 *usan redes más de* 25 *horas a la semana, y* 19 *usan redes menos de* 25 *horas a la semana. Construya la tabla de contingencia y pruebe la hipótesis de que usar redes más de* 25 *horas a la semana y el género son independientes a un nivel del* 0.01 □

7.8.1. Guía Bibliográfica

Pueden leer 10.12 de Walpole (2007).

8. Regresión Lineal

8.1. Regresión lineal simple

Hasta acá hemos tenido muestras $Y_1, Y_2, \ldots, Y_n$, asumidas i.i.d. Una implicación de esto es que $E(Y_i) = \mu$ es un número constante que no depende de ninguna otra variable. Esto, generalmente, no es verdad en muchos problemas inferenciales. Por ejemplo, si tenemos los pesos de estudiantes de EAFIT, sabemos que la media de esos pesos va a depender de otras cosas, como por ejemplo, de la altura de cada persona (entre más alto más peso en promedio).

El modelo de regresión lineal nos permite modelar una variable en términos de otra, nos permite asociar variables en una relación lineal. Para nuestro ejemplo, podríamos escribir el modelo:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

donde los pesos son Y_i, las alturas son x_i, ϵ_i es un error (una variable aleatoria, que luego diremos que distribución debe seguir) y β_0, β_1 son parámetros desconocidos que debemos estimar para darnos cuenta como es la relación entre el peso y la altura. Por ejemplo, si tenemos los pesos medidos en Kg y las alturas medidas en cm, si obtenemos un $\hat{\beta}_1 = 0.3$ entonces podemos decir que un aumento de un centímetro en una persona le aumenta en promedio 0.3 Kg. Lo que se hace al estimar β_0 y β_1 es estimar la recta que mejor se ajuste a los datos.

Para estimar la recta que mejor se ajuste a los datos, usamos un procedimiento que se llama mínimos cuadrados ordinarios. ¿Como ajustar esto? Pues es bastante intuitivo, simplemente deseamos escoger la recta de tal manera que las diferencias entre los valores observados y los valores correspondientes en la recta sean lo más pequeños posibles. Así, para nuestro ejemplo, consideremos el modelo estimado.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

donde $\hat{Y}_i$ es el valor pronosticado de Y_i cuando $x = x_i$, entonces el error del valor de Y_i a partir de $\hat{Y}_i$ pueden verse como la diferencia entre estos valores al cuadrado, o sea, debemos minimizar la suma de las diferencias cuadradas de todos estos estimadores:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

Exercise 39. *¿Por qué creen que usamos las diferencias cuadradas y no las diferencias y ya? ¿Que otro tipo de diferencias se les ocurriría usar? ¿Habría algún problema con usar las diferencias de los valores absolutos?* □

¿Pero como minimizamos esta ecuación? Si la SSE tiene un mínimo, ocurrirá para valores de β_0. Recordando cálculo III, podemos hallar estos valores derivando e igualando a 0, o sea, como las soluciones de el conjunto de ecuaciones $\partial SSE/\partial\hat{\beta}_0 = 0$ y $\partial SSE/\partial\hat{\beta}_1 = 0$. Derivando:

$$\frac{\partial SSE}{\partial \hat{\beta}_0} = -\sum_{i=1}^{n} 2[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]$$

$$0 = -2\left(\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i\right)$$

y

$$\frac{\partial SSE}{\partial \hat{\beta}_1} = -\sum_{i=1}^{n} 2[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]x_i$$

$$0 = -2\left(\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2\right)$$

Resolviendo simultáneamente, obtenemos (esto se muestra como hacer en clase)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Exercise 40. *Mostar que*

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

□

Exercise 41. *Asuma que* $Y_1, \dots Y_n \sim N(\mu_1, \sigma_1)$ *y que* $X_1, \dots, X_n \sim N(\mu_2, \sigma_2)$. *Halle* $E[\hat{\beta}_0]$ *y* $E[\hat{\beta}_1]$. *Son insesgados y consistentes?* □

Hay 4 supuestos importantes del modelo de regresión lineal, que podemos checkear que se cumplen de una manera u otra, que veremos más tarde. Es importante tener en cuenta que si se violan estos supuestos, las conclusiones que saquemos de la regresión no son válidas, entonces en cualquier estudio serio y riguroso hay que confirmar estos supuestos. Por ahora, solo voy a listar los supuestos:

- **Linealidad:** x y y se relacionan de manera lineal. Formalmente

  $$E[Y_i] = \beta_0 + \beta_1 X_i$$

  Si las relaciones entre x y y son no-lineales, podemos subsanar esto aplicando transformaciones no lineales a x, o añadiendo más variables independientes.

- **Independencia:** Los residuales son independientes. Formalmente:

  $$E[\epsilon_i \epsilon_j] = 0 \quad \text{para} \quad i \neq j.$$

  Este problema se da principalmente en problemas de series de tiempo.

- **Homocedasticidad**: Esto es más difícil de escribir que lo que es de entender. Se refiere a que la varianza de los residuales tiene que ser constante, independiente de el valor de x. Formalmente:

  $$E[\epsilon_i \epsilon_i] = \sigma^2 \quad \text{para todo } i$$

  Para subsanar esto se pueden hacer transformaciones en las variables dependientes e independientes, o hacer una regresión ponderada (luego veremos como hacer esto).

- **Normalidad**: Los residuales están distribuidos normalmente con media 0. Formalmente

  $$\epsilon_i \sim N(0, \sigma^2)$$

  Este supuesto se puede violar si hay muchos outliers. Se puede subsanar con un análisis de outliers o con transformaciones no lineales a las variables.

Estos supuestos son bastante restrictivos a la hora de modelar un problema. Por eso se han introducido modelos de regresión con menos supuestos, como la regresión no paramétrica.

8.2. Regresión Lineal Múltiple

En la sección pasada quisimos explicar una variable mediante una relación lineal con otra variable (o sea, queremos ver si una variable aleatoria es función lineal de otra), y llamamos a esta metodología regresión lineal simple. Pero hay un modelo algo más general: podemos explicar una variable dependiente como una combinación lineal de el número de variables independientes que tengamos. Digamos que tenemos el modelo

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon.$$

O sea, cada observación se puede escribir como

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_k x_{ik} + \epsilon_i \tag{6}$$

para cada observación $i = 1, 2, \ldots, n$, donde x_{ij} es la j-ésima variable independiente en la i-ésima observación, y_i es la i-ésima observacion dependiente, y β_j es el coeficiente asociado a cada variable j con $j = 1, \ldots, k$. En resumen: tenemos n observaciones de la variable dependiente y de las k variables independientes. Ahora, definiendo las siguientes matrices

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

,

que es una matriz columna, de tamaño $n \times 1$,

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}$$

,

que es una matriz de tamaño $n \times (k+1)$, donde la primera columna es una simple construcción para estimar β_0,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

,

que es una matriz columna de tamaño $(k+1) \times 1$ y

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

,

que es una matriz columna de tamaño $n \times 1$. Con estas matrices, podemos reescribir la Ecuación 6 como

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$$

Usando un poquito de álgebra lineal, y minimizando la suma de cuadrados (que en este caso está dada por la operación matricial $\epsilon'\epsilon$) usando un poco de cálculo vectorial, es sencillo ver que podemos estimar β como

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

Ahora, veamos algunas propiedades de los estimadores por mínimos cuadrados

1. **Insesgadez**: $E[\hat{\beta}_i] = \beta_i$ para todo $i = 1, 2, \ldots, k$.
2. $Var(\hat{\beta}_i) = c_{ii}\sigma_2$, donde c_{ii} es el elemento en la fila y columna i (empezando en 0) de $(\boldsymbol{X'X})^{-1}$.
3. $Cov(\hat{\beta}_i, \hat{\beta}_j) = c_{ij}\sigma^2$ donde c_{ij} es el elemento en la fila i y columna j (empezando en 0) de $(\boldsymbol{X'X})^{-1}$.
4. Un estimador insesgado de σ^2 es
$$S^2 = \frac{\boldsymbol{Y'Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X'Y}}{n - (k+1)}$$
5. Si ϵ_i está distribuído normalmente, tenemos que
   a) Cada $\hat{\beta}_i$ tiene distribución normal con media y varianza descritas anteriormente.
   b) La variable aleatoria
   $$\frac{n - (k+1)S^2}{\sigma^2}$$
   tiene distribución χ^2 con $n - (k+1)$ grados de libertad.
   c) S^2 y $\hat{\beta}_i$ son independintes para todo $i = 0, 1, 2, \ldots, k$

Todas estas propiedades son importantes para poder hacer inferencia sobre los β_i, que propondremos en la próxima sección.

8.3. Inferencias sobre funciones lineales de los parámetros estimados

Digamos que tenemos un modelo de regresión lineal

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \epsilon$$

donde el vector $\boldsymbol{\beta}$ tiene $k+1$ parámetros. Queremos hacer inferencia sobre la función lineal

$$a_0\beta_0 + a_1\beta_1 + \cdots + a_k\beta_k \tag{7}$$

donde $a_0, \ldots, a_k$ son constantes, posiblemente iguales a 0. Podemos escribir matricialmente

$$\boldsymbol{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}$$

y podemos escribir

$$\boldsymbol{a'\beta} = a_0\beta_0 + a_1\beta_1 + a_2\beta_2 + \cdots + a_k\beta_k$$

Estimando esta cantidad, tenemos que

$$\widehat{\boldsymbol{a'\beta}} = a_0\hat{\beta}_0 + a_1\hat{\beta}_1 + a_2\hat{\beta}_2 + \cdots + a_k\hat{\beta}_k.$$

La esperanza de esta cantidad es:

$$\begin{aligned} E[\widehat{\boldsymbol{a'\beta}}] &= E[a_0\hat{\beta}_0 + a_1\hat{\beta}_1 + a_2\hat{\beta}_2 + \cdots + a_k\hat{\beta}_k] \\ &= a_0E[\hat{\beta}_0] + a_1E[\hat{\beta}_1] + a_2E[\hat{\beta}_2] + \ldots a_kE[\hat{\beta}_k] \\ &= a_0\beta_0 + a_1\beta_1 + a_2\beta_2 + \ldots a_kE\beta_k \qquad \text{(Por la propiedad 1 de la sec. anterior)} \\ &= \boldsymbol{a'\hat{\beta}} \end{aligned}$$

o sea que $\hat{\boldsymbol{a'\beta}}$ es insesgado. Ahora, veamos como es su varianza

$$\begin{aligned}
Var[\widehat{\boldsymbol{a'\beta}}] &= Var[a_0\hat{\beta}_0 + a_1\hat{\beta}_1 + a_2\hat{\beta}_2 + \cdots + a_k\hat{\beta}_k] \\
&= a_0^2 Var[\hat{\beta}_0] + a_1^2 Var[\hat{\beta}_1] + a_2^2 Var[\hat{\beta}_2] + \ldots a_k^2 Var[\hat{\beta}_k] \\
&+ 2a_0a_1 Cov[\hat{\beta}_0, \hat{\beta}_1] + 2a_0a_2 Cov[\hat{\beta}_0, \hat{\beta}_2] + \cdots + 2a_{k-1}a_k Cov[\hat{\beta}_{k-1}, \hat{\beta}_k] \\
&= a_0^2 c_{00}\sigma^2 + a_1^2 c_{11}\sigma^2 + a_2^2 c_{22}\sigma^2 + \ldots a_k^2 c_{kk}\sigma^2 + 2a_0 a1 c_{01}\sigma^2 + 2a_0a_2 c_{02}\sigma^2 + \cdots + 2a_k a_{k-1} a_k c_{k-1,k}\sigma^2 \\
&= \sigma^2[\boldsymbol{a'}(\boldsymbol{X'X})^{-1}\boldsymbol{a}].
\end{aligned}$$

Ahora, recordemos que $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ están distribuídos normalmente (propiedad 5), y como $\boldsymbol{a'\hat{\beta}}$ es una combinación lineal de $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$, entonces es una combinación lineal de variables aleatorias normales, entonces es también normal con media y varianza descritas anteriormente. Luego, podemos proponer el estadístico

$$\begin{aligned}
Z &= \frac{\boldsymbol{a'\hat{\beta}} - \boldsymbol{a'\beta}}{\sqrt{Var(\boldsymbol{a'\hat{\beta}})}} \\
&= \frac{\boldsymbol{a'\hat{\beta}} - \boldsymbol{a'\beta}}{\sigma\sqrt{\boldsymbol{a'}(\boldsymbol{X'X})^{-1}\boldsymbol{a}}}
\end{aligned}$$

que tiene una distribución normal estándar, para lo cual podemos probar la hipótesis $H_0 : \boldsymbol{a'\beta} = (\boldsymbol{a'\beta})_0$, donde $(\boldsymbol{a'\beta})_0$ es una especificación de $\boldsymbol{a}$. Pero hay un problema con esto: generalmente no conocemos el valor de σ^2, sino que solo lo podemos estimar de los datos como S^2 (propiedad 4). Ahora, si sustituimos σ por S en el estadístico anterior, obtenemos que

$$T = \frac{\boldsymbol{a'\hat{\beta}} - \boldsymbol{a'\beta}}{S\sqrt{\boldsymbol{a'}(\boldsymbol{X'X})^{-1}\boldsymbol{a}}}$$

tiene una distribución t de Student con $n - (k+1)$ grados de libertad, y proporciona un estadístico de prueba para comprobar la hipótesis que planteamos arriba. De la misma manera, podemos encontrar un intervalo de confianza con un α dado como

$$\boldsymbol{a'\hat{\beta}} \pm t_{\alpha/2} S\sqrt{\boldsymbol{a'}(\boldsymbol{X'X})^{-1}\boldsymbol{a}}.$$

Esto nos permite construir una gran cantidad de pruebas estadísticas. En particular, si $a_i = 1$ para un i y 0 para los demás, podemos ver la significancia de de un párametro particular $\hat{\beta}_i$. O sea, podemos testear la hipótesis de que $H_0 : \beta_i = 0$ vs $H_a : \beta_i \neq 0$. Si rechazamos esta hipótesis, podemos decir que la variable x_i tiene influencia lineal en y. Si no podemos rechazar esta hipótesis, tenemos algo de evidencia a que la variable x_i no aporta nada (linealmente) a y.

8.4. Prueba de significancia de modelo

Muchas veces, tenemos una variable dependiente que queremos explicar y un montón de variables dependientes usadas para explicarla, pero ni si quiera sabemos si tiene sentido plantear un modelo lineal, o sea, si las variables que tenemos para explicar realmente explican algo de la variación de la variable que queremos explicar.

O sea, queremos ver si el modelo con solo el intercepto

$$Y = \beta_0 + \epsilon_i$$

realmente es diferente del modelo completo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

y deseamos testear la hipótesis nula $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$ vs la alternativa de que al menos alguno de los β es diferente. Llamemos al modelo con solo intercepto el modelo R, y al modelo completo el modelo C. A ambos modelos podemos calcularle SSE. El del modelo R lo llamamos SSE_R y al del modelo C lo llamamos SSE_C. La resta $SSE_R - SSE_C$ debe ser bastante grande cuando H_0 es falsa, debido a que si las variables en verdad logran explicar Y, tendremos que el SSE_C se vuelve muy pequeño ya que las variables logran explicar Y y hay mucho menor error que el modelo que solo usa el intercepto. Ahora, podemos definir estas dos cantidades

$$\begin{aligned}\chi^2_{k-1} &= \frac{SSE_R - SSE_C}{\sigma^2} \\ \chi^2_{n-(k+1)} &= \frac{SSE_C}{\sigma^2 2}\end{aligned}$$

que son dos variables aleatorias chi cuadrado con $k-1$ y $n-2$ grados de libertad respectivamente. Ahora, podemos construir el estadístico

$$\begin{aligned}F &= \frac{\chi^2_{k-1}/(k-1)}{\chi^2_{n-2}/n-2} \\ &= \frac{(SSE_R - SSE_C)/(k-g)}{(SSE_C)/(n-(k+1))}\end{aligned}$$

que es una razón de chi cuadrados, que sabemos que tiene distribución F con $v_1 = k-1$ grados de libertad en el numerador y $v_2 = n-(k+1)$ grados de libertad en el denominador. Como valores grandes de $SSE_R - SSE_C$ nos llevan a rechazar H_0, si deseamos una prueba con error tipo I igual a α, debemos rechazar cuando $F > F_\alpha$.

Exercise 42. *En R, haga un modelo lineal como el que hicimos en clase para los datos de* `pizzadelivery.csv`*, usando la variable* `time` *como Y y todas las demás variables como explicativas. ¿Hay modelo? (Haga una prueba F). ¿Cuales son las variables que son significativas? ¿Que relaciones interesantes encuentra?* □

8.5. Bondad de ajuste de un modelo lineal

Consideremos el valor SSE_R, que es el valor de la suma de errores cuadrados cuando consideramos un modelo con solo el intercepto. Este también se escribe como S_{yy} Ahora, considere el estadístico

$$R^2 = \frac{S_{yy} - SSE_C}{S_{yy}}$$

esto se denomina como el coeficiente de el determinación de la regresión. Este es un valor tal que $0 \leq R^2 \leq 1$. Se puede interpretar como la cantidad de variación de la variable dependiente que es explicada por las variables independientes. Entre más cercano este a 1, podemos decir que tenemos mejor modelo, y entre más cercano este a 0, menos podemos explicar usando las variables dependientes.

Exercise 43. *Describa matemáticamente, y en palabras, que debe pasar para que $R^2 = 1$ y para que $R^2 = 0$* □

Exercise 44. *¿Como se distribuye R^2? Para hacer ejercicio, usemos lo que sabemos de clases pasadas: podemos estimar la distribución de cualquier estadístico remuestreando la muestra original. Para este ejercicio, usaremos los datos de* `pizzadelivery.csv`*, donde la variable dependiente será* `time` *y la independientes serán todas las demás. Remuestree la matriz de datos* 1000 *veces. En cada uno de esos remuestreos, ajuste un modelo lineal y calcule R^2. Dibuje un histograma de estos* 1000 *R^2. Obtenga un intervalo de confianza del* 95 % *para R^2.* □

Exercise 45. *Demuestre que el estadístico para la prueba de si hay modelo o no también se puede escribir como*

$$F = \frac{n-(k+1)}{k}\left(\frac{R^2}{1-R^2}\right)$$

□

Exercise 46. *Investigue que es un R^2 ajustado y cual es su formula. Explique por que y para que se usa. ¿Prefiere el R^2 usual o el ajustado? Justifique su preferencia.* □

8.6. Selección de modelos

En la practica, tenemos una variable dependiente Y y un montón de variables independientes $x_1, x_2, \ldots x_k$. ¿Pero serán todos los x_k necesarios? ¿Siempre es mejor usar todas las variables?

Pues resulta que no. Agregar variables indiscriminadamente hace que se aumente la dimensionalidad del problema, lo cual generalmente no es bueno. Y hace que nuestro modelo sea más complejo, y por ende más difícil de interpretar, lo cual nunca es bueno. Para esto, se han creado diferentes criterios de bondad de ajuste de modelos. Dos de ellos que hemos visto son el R^2 y el $\tilde{R^2}$ (la versión ajustada, introducida en el ejercicio 46). Introduzcamos otros dos criterios:

$$AIC = 2k - 2ln(\hat{L})$$
$$BIC = kln(n) - 2ln(\hat{L})$$

donde $\hat{L}$ es la estimación de la función de verosimilitud. Entre más grandes sean R^2 y $\tilde{R^2}$, tenemos mejor modelo, pero entre más pequeños sean AIC y BIC tenemos mejor modelos. Los valores de AIC y BIC pueden ser negativos, entonces hay que tener cuidado cuando esto ocurra (un error usual es, por ejemplo, si tenemos dos modelos, uno con $AIC = -200$ y otro con $AIC = -300$ es escoger el que tiene $AIC = -200$. Pero el mejor sería $AIC = -300$ porque es más negativo, i.e., $-300 < -200$).

Exercise 47. *Para los datos de* `pizzadelivery.csv` *haga un proceso de selección de modelo. Para esto, primero, empiece con un modelo lineal con solo el intercepto. Esto en R se puede hacer con* `lm(y ~ 1)`*. Mida AIC, BIC, R^2 y $\tilde{R^2}$. Luego, añada una variable cualquiera, y mida otra vez AIC, BIC, R^2 y $\tilde{R^2}$. Haga esto hasta que se quede sin variables, midiendo AIC, BIC, R^2 y $\tilde{R^2}$ cada vez que añada una variable nueva. Escoja el mejor modelo según los cuatro criterios. ¿Que puede interpretar de esto? En R, el AIC y BIC se puede sacar usando* `AIC(model)` *y* `AIC(model)`*, donde* `model` *es una variable que contiene el modelo lineal correspondiente.* □

Exercise 48. *Haga una regresión para* `pizzadelivery.csv` *con la variable* `time` *como variable independiente y todas las demás como variables independientes. Interprete los valores de β, interprete las significancias de los β.* □

8.7. Resumen regresión lineal

Como vimos, la regresión lineal es un método bastante interesante para analizar relaciones lineales entre variables. Es una de las herramientas más usadas de la estadística, pero también una de las peor entendidas y peor utilizadas. Muchas veces, ni si quiera se checkean los supuestos que deben seguir los datos, y se sacan conclusiones erróneas por esto. En esta sección les voy a proponer un workflow para trabajar regresión lineal sin tener estos problemas. Supongamos que tenemos una variable dependiente Y, que queremos estimar con una serie de variables independientes $x_1, x_2, \ldots, x_k$.

1. Hacer un análisis exploratorio de los datos. Con base en esto, podemos proponer nuevas variables independientes como transformaciones de las variables originales o interacciones entre ellas. (Ejemplo: tenemos unos datos de rendimiento de fútbol, y queremos ver como afectan el peso y la altura ese rendimiento. Pero tenemos la hipótesis de que el peso y la altura actúan conjuntamente, no individualmente al rendimiento. Podemos proponer una nueva variable que sea la multiplicación de peso y altura, y así podemos ver como afecta la interacción de estas al rendimiento). En el análisis exploratorio, ver si hay variables colineales (que hacen que la matriz de covarianza y por ende los β sean mal estimados), descartar si es el caso. Eso hace viendo las nubes de puntos entre cada par de variables (en R hacer esto es bastante fácil).

2. Hacer un proceso de selección de modelo para saber exactamente cuales variables vamos a utilizar. (Usar criteriro AIC o BIC).

3. Checkear si hay homocedasticidad: Para eso, hay que hacer un gráfica de $\hat{\epsilon}_i$ vs $\hat{y}_i$. Esto es fácil en R también. Tiene que dar una nube de puntos donde no haya una tendencia marcada. Si no se cumple eso, hay que estimar los βs usando mínimos cuadrados ponderados.

4. Checkear que $\hat{\epsilon}_i$ es normal. Para esto, plotear histograma, plotear qq-plot, prueba de Kolmogorov-Smirnov.

5. Cuando todos los supuestos se cumplen (homocedasticidad, normalidad residuales, no colinealidad), ver el valor de F, decidir si hay modelo o no.

6. Si hay modelo: ver los valores de β, mirar si son significativos, interpretar significancias.

7. Interpretar los valores de β significativos.

En interactiva voy a montar un ejemplo completo donde haga todo esto. También lo muestro en clase.

Exercise 49. *Haga un análisis de regresión, con todos los pasos mencionados acá, para la variable Wine quality en el excel* `datos.xls`*, en la hoja llamada Wine Quality.* □

Exercise 50. *Haga un análisis de regresión, con todos los pasos mencionados acá, para la variable Peso en el excel* `datos.xls`*, en la hoja llamada Medifis.* □

Referencias

Conover, W. (1971). *Practical Nonparametric Statistics*. Wiley.

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

DiCiccio, T. J. and B. Efron (1996). Bootstrap confidence intervals. *Statistical Science 11*(3), 189–212.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1), 1–26.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine 2*(8), e124.

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.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar (2019). Moving to a world beyond "$p < 0.05$". *The American Statistician 73*(sup1), 1–19.

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