

Estadística I: Análisis exploratorio de datos y muestreo

Rodrigo Zepeda-Tello

2021-08-03

Contents

1 Historia del muestro	7
2 Conceptos	9
3 Análisis Exploratorio de Datos	17
3.1 Inicio	17
3.2 Librerías	17
3.3 Base a analizar	17
3.4 Definiciones y notación	18
3.5 Estadísticos univariados	19
3.6 Ejercicio	23
3.7 Ejercicios	35
3.8 Gráficas univariadas	36
3.9 Gráficas bivariadas	42
3.10 Estadísticos bivariados	50
3.11 Ejercicio	69
3.12 Ajuste funcional	70
3.13 Ejercicios del capítulo	86
4 Muestreo Aleatorio Simple	93
4.1 Inicio	93
4.2 Librerías	93
4.3 Notación	93
4.4 Muestreo Aleatorio Simple sin Reemplazo (MAS/sR)	97
4.5 Teorema del Límite Central (Aplicación)	103
4.6 Ejemplo Resumen: Estimación de una proporción bajo muestreo aleatorio simple sin reemplazo	106
4.7 Ejemplo Resumen: Estimación del total de individuos en una fotografía	109
4.8 Ejercicio:	118
4.9 Ejemplo Resumen: Estimación de una región crítica	119
4.10 Ejemplo Resumen: Estimación del total de una población	121
4.11 Demostración del Teorema del Límite Central para Muestras Finitas	123

4.12 Muestreo Aleatorio Simple Bernoulli (BE)	123
4.13 Ejemplo Resumen: Aduana	129
4.14 Muestreo Aleatorio Simple con Reemplazo (MAS/cR)	131
4.15 Ejemplo Resumen: Proporción de trabajadores enfermos con o sin reemplazo	134
4.16 Ejemplo Resumen: Captura-Recaptura con reemplazo	136
4.17 Muestreo Aleatorio Simple Ponderado (MAS/P)	137
4.18 Ejercicios	138
5 Muestreo Aleatorio Estratificado	143
5.1 Introducción a Muestreo Aleatorio Estratificado (MAE)	143
5.2 Alocación	146
5.3 Ejercicio de clase:	150
5.4 Ejercicio en R tipo control	152
6 Intervalos de Confianza mediante bootstrap	157
6.1 Inicio	157
6.2 Intervalos asintóticos	157
6.3 Muestreo basado en modelos	159
6.4 Intervalos Bootstrap	162
7 Muestreo Aleatorio Multietápico	169
7.1 Muestreo aleatorio por clusters en una sola etapa	170
7.2 Muestreo aleatorio por clusters bietápico (en dos etapas)	173
7.3 Ejemplo: Disco duro	175
7.4 Ejemplo: Encuesta Nacional de Salud	176
A Programación en R	179
A.1 Algunas ventajas de R y cosas no tan padres	180
A.2 Bienvenidx a R, Camp Pontanezen (sí, así se llama esta versión) .	181
A.3 Instalando cosas	184
A.4 Instalación de RStudio	184
A.5 Primeros pasos en R usando RStudio	187
A.6 Cálculos numéricos	196
A.7 Variables	199
A.8 Observaciones sobre la aritmética de punto flotante	204
A.9 Leer y almacenar variables en R	207
A.10 Instalación de paquetes	211
A.11 Comentarios adicionales sobre el formato	213
B Repaso de Prob	219
B.1 Funciones indicadoras	219
B.2 Conteo	219
B.3 Espacios de probabilidad	221
B.4 Probabilidad condicional	222
B.5 Independencia	224

B.6 Variables aleatorias y función de distribución (acumulada)	224
B.7 Funciones de masa de probabilidad	226
B.8 Funciones de densidad	232
B.9 Teorema de cambio de variable unidimensional	237
B.10 Probabilidad Multivariada	239
B.11 Esperanza, varianza y covarianza	240
B.12 Condicionamiento por otra variable aleatoria	242
B.13 Funciones características	242
B.14 Convergencias	242
B.15 Ley de los grandes números	242
B.16 Teorema del límite central	243

Chapter 1

Historia del muestro

Chapter 2

Conceptos

Este libro trata sobre **datos** y **encuestas** que estudian **poblaciones** a través de **muestras**. Antes de empezar con las matemáticas es necesario definir estos conceptos para tener claro de qué se habla. El más importante es la distinción entre **encuesta** y **experimento**: todo este libro hablará sobre encuestas, dejando la estadística para experimentos para otras notas.

- **Experimento** Cualquier diseño de estudio donde la investigadora puede, potencialmente, replicar cuantas veces desee el estudio y obtener cuantas mediciones sean necesarias. El objeto de estudio NO es una población finita. Por ejemplo: alimentar ratones con una sustancia y medir su presión arterial (pueden obtenerse cuantos ratones sean necesarios), determinar la vida media de un compuesto (puede hacerse cuantos kilogramos de la sustancia se requieran) o establecer si una forma de enseñanza de estadística genera mejores resultados en los exámenes que otra (se pueden obtener tantos alumnos como se desee y enseñarles de X o Y forma hasta distinguir una diferencia si es que hay).

Nota Si has escuchado hablar de inferencia estadística usualmente esto es lo que se enseña cuando se consideran variables aleatorias independientes idénticamente distribuidas. Muchos de los teoremas (como normalidad aproximada) funcionan cuando el tamaño de tu muestra, n , tiende a infinito, $n \rightarrow \infty$. Eso funciona súper bien cuando lo que se analiza son experimentos que se pueden repetir tantas veces como sea necesario para obtener la aproximación límite. Este libro no trata sobre experimentos sino sobre encuestas y en una encuesta usualmente se considera una población finita por lo que no se pueden obtener infinitas mediciones.

- **Encuesta** Una encuesta es un diseño de estudio donde se busca determinar el estado actual o una característica de una población finita (tamaño usualmente denotado N). Una encuesta no es replicable en igualdad de

condiciones y está sujeta a que, a lo más, puede obtener tantas mediciones como el tamaño de la población. Por ejemplo: realizar un cuestionario sobre salud mental dentro de la población de estudiantes (no es replicable porque si se vuelve a hacer, la salud mental de los estudiantes o la población ya cambió); determinar el tamaño en bytes del Internet (la población es finita); medir la altura de los árboles de un bosque (finito y no replicable).

Nota Una encuesta no es un cuestionario. Algunas encuestas de salud, por ejemplo, toma muestras de sangre de las personas para determinar la proporción de personas con diabetes. Esta encuesta no es un cuestionario (no tiene una pregunta) pero sí es una encuesta (busca determinar el estado actual o una característica de una población finita). Lo mismo ocurre cuando se busca determinar la cantidad de árboles en un bosque (no se les pregunta a los árboles cuántos son) o cuando se quiere establecer el peso promedio de las vacas en distintos centros ganaderos (no se les pregunta a las vacas cuánto pesan).



Figure 2.1: No todas las encuestas tienen sentido como cuestionarios

- **Población** Cualquier conjunto no vacío. Algunos ejemplos de poblaciones incluyen: las personas que viven en Guatemala (si me interesa saber algo de los guatemaltecos en general), los árboles del Amazonas (si quiero saber cosas de ecología en torno al río), los perros callejeros en Ciudad de México, los consumidores de una marca de cereal, los coches que transitan por Dubai o los granos de arena en una playa específica de Cancún.

Nota: Las poblaciones no necesariamente son de seres vivos son sólo conjuntos de cosas que se están estudiando. Las poblaciones usualmente están restringidas al tiempo y al espacio por lo que es importante tener una definición *clara* de quiénes sí están en el estudio, quiénes no y por qué.

Nota: En la mayoría de los problemas de encuestas que enfrentaremos suponemos que la población es de tamaño *finito* N . Esto en contraste con un experimento donde la población es de tamaño infinito.

- **Población objetivo** El conjunto de elementos que formarán parte del estudio. Definir la *población objetivo* es complicado en algunas situaciones; por ejemplo, si se desea saber si *los mexicanos* están a favor o en contra de legalizar la marihuana hay que establecer quiénes son *los mexicanos*. ¿Cuentan las personas con nacionalidad mexicana que residen en el extranjero? ¿Cuentan los menores de edad? ¿Qué pasa con los extranjeros que son residentes?
- **Población muestreada** Es el conjunto de elementos sobre los cuales se tomó la muestra para el análisis estadístico. *Idealmente* la población objetivo y la muestreada deberían de ser igual pero el mundo no es tan bello. En encuestas de consumo, por ejemplo, usualmente no se muestran zonas remotas o de muy bajos recursos. En encuestas de elecciones si bien la población objetivo son *todas las personas que voten el día de la elección*, como la mayoría se hacen *antes* de la elección (exceptuando las de salida) entonces se aproxima la definición de *votante* buscando incluir sólo aquellos que estén registrados en el padrón electoral o bien aquellos que al ser encuestados digan que *sí* van a votar.

A veces a la población se le conoce como *el conjunto universo* y es por esto que se denota:

$$\mathcal{U} = \{u_1, u_2, \dots\}$$

con u_i siendo sus elementos. Nosotros (hasta que se diga lo contrario) supondremos que la muestreada coincide con la objetivo y por tanto esa \mathcal{U} será sólo **la población**.

- **Muestra** Un subconjunto de la población muestreada. Si la muestra coincide con la población muestreada se dice que es un **censo**. Los “mejores” censos (para nosotros) son aquellos donde la población muestreada y la

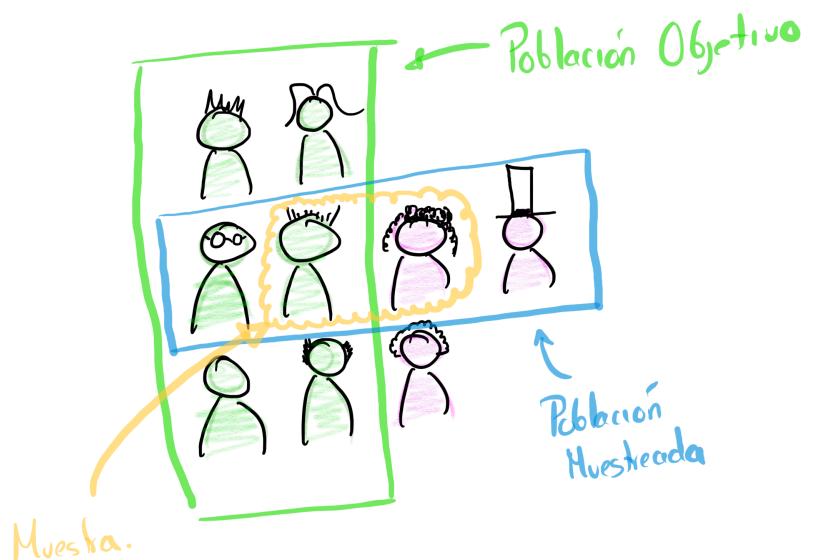


Figure 2.2: Diferencia entre población objetivo, población muestreada y muestra

población objetivo coinciden (porque ya medimos todo lo que queríamos). Para el propósito de estas notas los únicos **censos** que consideraremos son aquellos donde la población muestreada es la objetivo.

La muestra es un subconjunto de la población:

$$\mathcal{S} = \{s_1, s_2, \dots\} \subseteq \mathcal{U}$$

con s_i siendo sus elementos. Nosotros sólo consideraremos muestras que son de tamaño finito, n .

Nota: Hasta ahora no estamos hablando de *muestras aleatorias*. Esta definición habla de cualquier subconjunto no necesariamente uno que se haya obtenido por algún mecanismo aleatorio. Por ejemplo, si durante una pandemia se le pide a todas las personas de una población con apellidos de la A a la F acudan a una institución esto sí es una muestra (dado que es subconjunto de la población) pero podría argumentarse no es una muestra aleatoria (pues el proceso de selección fue determinista).

Nota A menos que se especifique lo contrario, el vacío, \emptyset es una muestra.

Me dijiste que mi presupuesto alcanzaba para 1000 muestras

La muestra vacía también es una muestra



Figure 2.3: Hay que ser muy claros para especificar que queremos una muestra con $n > 0$

- **Marco muestral** Una lista a partir de la cual se selecciona la muestra para la encuesta. Puede ser, en un salón de clases, la lista completa de

alumnos. En estudios de agricultura usualmente la lista son *parcelas de tierra* aunque interese estudiar los cultivos mientras que en poblaciones grandes de personas el marco suele ser una lista de casas (dado que no se sabe qué persona vive dónde) o bien un mapa de calles y colonias. Para el INEGI es común usar las *Áreas Geoestadísticas Básicas* (AGEB) las cuales son divisiones fijas (pequeñísimas, como una manzana) del mapa de México.

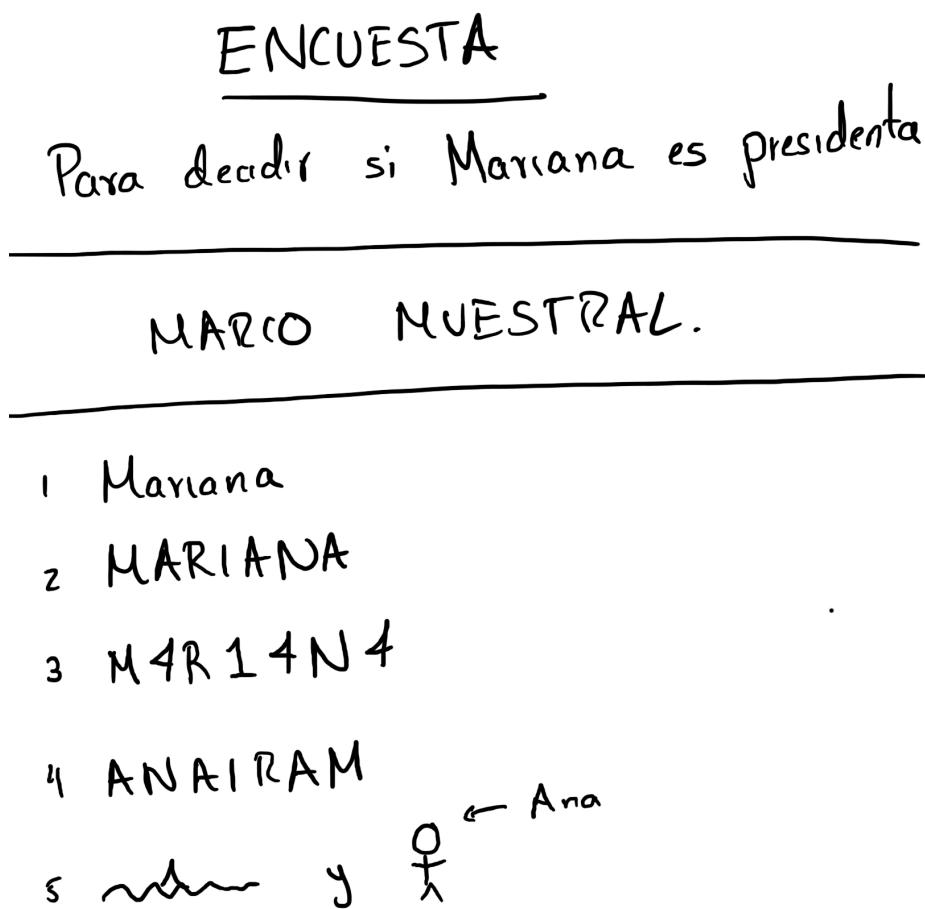


Figure 2.4: Este marco muestral está amañado

- **Unidad de muestreo** Una unidad que puede seleccionarse del marco muestral. Por ejemplo, en la lista de casas la *unidad de muestreo* sería una casa. Por otro lado, en una lista de parcelas la *unidad* es la parcela. Si se tiene una lista de estudiantes en un grupo la *unidad* serían los estudiantes. Puede que haya múltiples unidades de muestreo en la misma encuesta

por ejemplo, si se desea hacer una encuesta para evaluar alumnos de las escuelas. Suponiendo que el muestreo ocurre de la siguiente forma: 1) primero sólo se tiene la lista de escuelas por lo que se seleccionan al azar un número de escuelas; 2) una vez las investigadoras llegan a la escuela elegida, obtienen la lista de los alumnos inscritos y muestran sobre ellos. En este escenario hay dos niveles de marcos muestrales (la lista de escuelas de manera inicial y la lista de alumnos por escuela que se obtiene una vez llegas a la escuela elegida). La primera unidad de muestreo (Unidad Primaria de Muestreo) es la escuela; la segunda unidad de muestreo (Unidad Secundaria de Muestreo) son los alumnos.

- **Unidad de observación** El objeto que interesa medir. En el caso de una encuesta dirigida a personas donde se utiliza una lista de casas para encontrar a las personas, la unidad de muestreo es la casa (el objeto de la lista) pero la unidad de observación son las personas que viven dentro de la casa (lo que quiero medir). Una cosa muy parecida (pero no idéntica) es tener una lista de casas y desear estudiar propiedades de la casa (digamos, tamaño). En ese caso la unidad de muestreo y observación coinciden: son la casa.

Chapter 3

Análisis Exploratorio de Datos

3.1 Inicio

Siempre que inicies un nuevo trabajo en R ¡no olvides borrar el historial!

```
rm(list = ls()) #Clear all
```

3.2 Librerías

Para este análisis vamos a tener que llamar a las siguientes librerías previamente instaladas (por única vez) con `install.packages`:

```
library(tidyverse)
library(dplyr)
library(moments)
library(lubridate)
library(ggcorrplot)
library(ks)
```

Si no tienes una librería puedes instalarla escribiendo en la consola el `install` junto con su nombre:

```
install.packages("lubridate")
```

3.3 Base a analizar

Como ejemplo analizaremos la base de *Carpetas de Investigación de la Fiscalía General de Justicia* de la CDMX para el año 2018 y mes de Diciembre misma

que se encuentra en este link

Si el link anterior no abre ve al sitio https://datos.cdmx.gob.mx/explore/dataset/carpetas-de-investigacion-pgj-cdmx/table/?refine.ao_hechos=2018 y elige la opción de año 2018, mes diciembre y descargar como csv.

La forma más fácil en RStudio es yéndonos a **Import Dataset** en el panel derecho seguido de **From Text** y seleccionamos el archivo. En este caso hay dos opciones cualquiera de las dos opciones funciona: si en tu ordenador no sirve una, ¡prueba la otra!

En mi caso el archivo está en una carpeta que se llama **datasets** y se lee de la siguiente manera:

```
datos <- read.csv("datasets/carpetas-de-investigacion-pgj-cdmx.csv")
```

3.4 Definiciones y notación

Siguiendo la definición de Gelman et al. (2013) , denotamos el conjunto de datos observados como la matriz (*base de datos*) de $n \times \ell$

$$Z = (z_1 | z_2 | \dots | z_\ell)$$

donde $\ell \in \mathbb{N}$ con $\ell > 0$ y las z_i sin pérdida de generalidad, son vectores columna de longitud n ($z_i = (z_{i,1}, z_{i,2}, \dots, z_{i,n})^T$). Una columna z_k con $0 \leq k \leq \ell$ se le conoce como:

- *Numérica* si $z_k \in \mathbb{R}^n$. En particular es entera si $z_j \in \mathbb{Z}^n$.
- *Categórica* si cada entrada de z_k es una indicadora de pertenencia a algún conjunto (por ejemplo Hombre / Mujer ó Ingresos Altos / Ingresos Medios / Ingresos Bajos). Usualmente z_k se representa con un carácter o con un entero. Una variable categórica puede ser *lógica* si z_k es un indicador que toma alguno de los dos valores: TRUE ó FALSE.
- *Ordinal* Una variable ordinal es aquél $z_k \in \mathcal{C}$ donde sobre \mathcal{C} existe un orden total; es decir si $x, y, w \in z_k$ se tiene que:
 - a. Ocurre al menos una de las siguientes: $x \leq y$ ó $x \geq y$.
 - b. Si $x \leq y$ y $y \geq w$ entonces $x \leq w$
 - c. Si $x \leq y$ y $x \geq y$ entonces $x = y$. Variables numéricas univariadas son ordinales por el orden natural de \mathbb{R} .
- *Carácter* si z_k es un carácter o una cadena de caracteres donde los caracteres son el objeto de análisis en sí (no como pertenencia). Por ejemplo si cada entrada $z_{k,m}$ representa un Tweet.

OJO Los datos $z_{k,m}$ son variables fijas ya dadas y **NO SON ALEATORIAS**.

En el caso de nuestra base de datos podemos resumir la información contenida en la misma mediante `glimpse`:

```
datos %>% glimpse()

## # Rows: 19,861
## # Columns: 18
## # $ año_hechos      <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2~
## # $ mes_hechos       <chr> "Diciembre", "Diciembre", "Diciembre", "Diciembre"
## # $ fecha_hechos     <chr> "2018-12-13 12:00:00", "2018-12-22 19:00:00", "20~
## # $ delito            <chr> "USURPACIÓN DE IDENTIDAD", "SUSTRACCION DE MENORE~
## # $ categoria_delito  <chr> "DELITO DE BAJO IMPACTO", "DELITO DE BAJO IMPACTO~
## # $ fiscalía           <chr> "INVESTIGACIÓN EN MIGUEL HIDALGO", "INVESTIGACIÓN~
## # $ agencia            <chr> "MH-2", "59", "BJ-1", "IZP-9", "75TER", "FDS-5", ~
## # $ unidad_investigacion <chr> "UI-1SD", "UI-1CD", "UI-1SD", "UI-2SD", "3 S/D", ~
## # $ colonia_hechos     <chr> "LOMAS DE SOTELO", NA, "DEL VALLE CENTRO", "AMPLI~
## # $ alcaldia_hechos    <chr> "MIGUEL HIDALGO", "CUAUTLA", "BENITO JUAREZ", "IZ~
## # $ fecha_inicio        <chr> "2019-06-16 12:14:09", "2019-06-06 16:26:15", "20~
## # $ mes_inicio          <chr> "Junio", "Junio", "Febrero", "Abril", ~
## # $ ao_inicio            <int> 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2~
## # $ calle_hechos         <chr> "AV. CONSCRIPTO", "AVENIDFA DIEZ DE MARZO", "FELI~
## # $ calle_hechos2        <chr> ".", "HECHOS EN CUAUTLA MORELOS", "ESQUINA COYOAC~
## # $ longitud             <dbl> -99.22535, NA, -99.17088, -99.03016, -99.13423, ~~
## # $ latitud              <dbl> 19.44028, NA, 19.37207, 19.34797, 19.54788, 19.34~
## # $ Geopoint              <chr> "19.4402832543,-99.2253527208", "", "19.372068287~
```

Notamos que el vector columna `año_hechos` es una variable numérica mientras que `mes_hechos` es categórica. No hay variables lógicas en esta base. Una variable carácter es el vector columna `calle_hechos` que no denota un conjunto sino una cadena de caracteres (véanse las faltas de ortografía, por ejemplo).

Al ser la tabla de datos una matriz podemos acceder a la entrada en la fila j y columna k haciendo:

`base[j, k]`

por ejemplo:

```
datos[4, 6]
```

```
## [1] "INVESTIGACIÓN EN IZTAPALAPA"
```

NOTACIÓN Para facilitar la notación en lo que sigue de estas notas y hasta nuevo aviso, si z_k es una columna categórica de Z denotaremos a los elementos de dicha columna como $C = (c_1, c_2, \dots, c_n) = z_k^T$. Si z_k es numérica denotamos a los elementos de dicha columna como $\vec{x} = (x_1, x_2, \dots, x_n) = z_k$.

3.5 Estadísticos univariados

3.5.1 Definición [Estadístico]

Un estadístico es una función cuyo dominio es la matriz de datos observados Z o una columna de la misma. Es decir, un estadístico es cualquier función de los datos (ver Wolfe and Schneider (2017)).

A continuación veremos algunos ejemplos de estadísticos así como su interpretación.

1. Media poblacional Dado un vector de datos numéricos $\vec{x} = (x_1, x_2, \dots, x_n)^T$ definimos la media poblacional como:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad x_i \in \mathbb{R}$$

En el caso de nuestros datos podemos **calcular el promedio de delitos por día** como sigue. Primero necesitamos especificar a R que la `fecha_hechos` es una fecha. Esto lo hacemos mediante la función `ymd_hms` (year-month-day_hour-minute-second) del paquete de `lubridate` y la función `mutate` (que cambia una columna de la base de datos). El siguiente código le indica a R que cambie la columna `fecha_hechos` volviéndola a leer como fecha:

```
datos <- datos %>% mutate(fecha_hechos = ymd_hms(fecha_hechos))
```

Para mantener sólo la fecha y eliminar la hora de `fecha_hechos` podemos generar una nueva columna como sigue:

```
datos <- datos %>% mutate(fecha = date(fecha_hechos))
```

Finalmente podemos contar (`tally`) observaciones agrupadas (`group_by`) por día mediante la combinación de ambas funciones:

```
conteo_delitos <- datos %>% group_by(fecha) %>% tally()
```

```
## # A tibble: 6 x 2
##   fecha      n
##   <date>    <int>
## 1 2018-12-01  674
## 2 2018-12-02  584
## 3 2018-12-03  790
## 4 2018-12-04  640
## 5 2018-12-05  724
## 6 2018-12-06  718
```

Hay distintas formas de calcular la media. La primera es tomando la columna directo, para acceder a una columna utilizamos el signo de pesos \$ como sigue:

```
base$columna
```

En nuestro caso:

```
mean(conteo_delitos$n)
```

```
## [1] 640.6774
```

O bien podemos usar la función `summarise` integrada en `dplyr`:

```
conteo_delitos %>% summarise(mean(n))
```

```
## # A tibble: 1 x 1
##   `mean(n)`
##   <dbl>
## 1     641.
```

NOTA Una media como está escrita arriba no aplica para datos circulares. Por ejemplo, si queremos determinar el mes promedio en el que ocurren las lluvias dentro de los años se sabe que después del mes 12 continúa el mes 1 del próximo año. Una media tradicional no considera datos que pueden ser descritos mediante aritmética modular (como los meses). Para ello se utiliza la media circular:

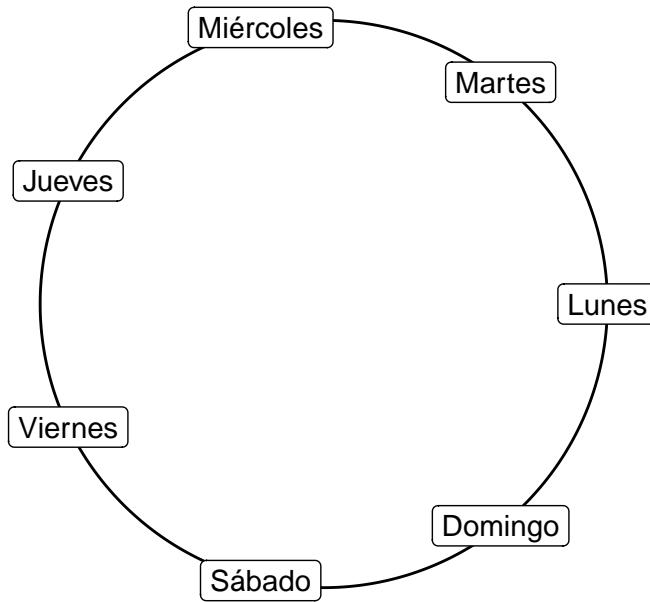
1.1 Media circular Consideremos el problema de determinar el día promedio de la semana en que más ocurren delitos (de Lunes a Domingo). Podemos resumir los eventos usando la función `weekdays`:

```
datos      <- datos %>% mutate(`Día de la Semana` = weekdays(fecha))
conteo.dia <- datos %>% group_by(`Día de la Semana`) %>% count()
```

de donde se tiene el conteo:

Día de la Semana	n
Monday	3251
Saturday	3197
Friday	2833
Sunday	2722
Wednesday	2701
Thursday	2679
Tuesday	2478

Para obtener el día promedio representamos cada uno de los días en el círculo usando coordenadas polares. Nota que el radio es irrelevante en este caso: sólo el ángulo importa; de ahí que tomemos $r = 1$:



Para un conjunto de mediciones con ángulos $(\theta_1, \theta_2, \dots, \theta_n)^T$ el centro de masa asociado a dichas mediciones es el punto (\bar{c}, \bar{s}) donde

$$\bar{c} = \frac{1}{n} \sum_{i=1}^n \cos(\theta_i) \quad \text{y} \quad \bar{s} = \frac{1}{n} \sum_{i=1}^n \sin(\theta_i)$$

La dirección media se define como la solución $\bar{\theta}$ (si $\bar{r} > 0$) a:

$$\bar{c} = \bar{r} \cos \bar{\theta} \quad \text{y} \quad \bar{s} = \bar{r} \sin \bar{\theta}$$

donde \bar{r} se conoce como *la longitud resultante promedio*. Si $\bar{r} = 0$ no existe dirección media. De manera explícita, por geometría tenemos que:

$$\bar{r} = (\bar{c}^2 + \bar{s}^2)^{1/2}$$

y que:

$$\bar{\theta} = \text{atan2}(\bar{s}/\bar{c}) = \begin{cases} \text{atan}(\bar{s}/\bar{c}) & \text{si } \bar{c} \geq 0 \\ \text{atan}(\bar{s}/\bar{c}) + \pi & \text{si } \bar{c} < 0 \end{cases}$$

donde el caso $\bar{c} = 0$ se interpreta como el límite por la derecha (respectivamente por la izquierda) de la arcotangente de acuerdo con el signo de \bar{s} .

Ahora bien, si tomamos

$$c = \sum_{i=1}^n \cos(\theta_i) \quad \text{y} \quad \bar{s} = \sum_{i=1}^n \sin(\theta_i)$$

podemos definir *la longitud resultante* como $r = (c^2 + s^2)^{\frac{1}{2}}$, entonces, es perfectamente aceptable tomar *la longitud resultante promedio* previamente definida como:

$$\bar{r} = (\bar{c}^2 + \bar{s}^2)^{1/2} = \frac{r}{n}$$

>**NOTA** Muchos son los casos donde las coordenadas polares son más útiles. Sabemos que las coordenadas polares piden la especificación de un ángulo ϕ y una recta de longitud r . Las coordenadas polares pueden ser utilizadas para especificar puntos en cualquier lugar de un plano, no solamente en el círculo unitario. En lugar de las coordenadas cartesianas (x_i, y_i) , las coordenadas polares (ϕ_i, r_i) especifican dónde se encuentra cada punto en términos de su dirección y distancia con respecto al origen. Así, nos dan la clara ventaja de separar información tanto de distancia como de dirección en el análisis de datos, mientras que las cartesianas confunden aspectos espaciales.

1.2 Varianza Angular(Circular) Pewsey, Neuhäuser, and Ruxton (2013) define la varianza circular como:

$$\text{Var} = 1 - \bar{r}$$

donde $\bar{r} = (\bar{c}^2 + \bar{s}^2)^{1/2}$ es el resolvente explicado anteriormente. Es importante notar que Var puede tomar valores en $[0, 1]$ y su interpretación es similar a la de la varianza de datos lineales, i.e. “entre más pequeño sea el valor de la varianza circular, más conectrada es la distribución de los datos” Fisher (1995).

1.3 Dirección Mediana Circular

Denotamos a la *dirección mediana circular* por $\tilde{\theta}$. Se puede determinar de manera muy similar al caso del cálculo de la mediana para datos lineales. De tal manera que $\tilde{\theta}$ es cualquier ángulo ϕ tal que la mitad de los datos se encuentran en el arco $[\phi, \phi + \pi]$ y la mayoría de los puntos están más cerca de ϕ que de $\phi + \pi$ (Mardia y Jupp). Ahora si n es impar, la mediana es uno de los datos, en caso contrario la mediana pasa por el punto medio de dos datos adyacentes.

1.4 Desviación Estándar Circular Al igual que con datos lineales, la desviación estándar está ítimamente relacionada con la varianza, definimos a la *desviación estándar circular* como

$$v = (-2\ln(1 - V))^{\frac{1}{2}} = (-2\ln(\bar{r}))^{\frac{1}{2}}$$

Para más información sobre estadística circular puedes consultar Pewsey, Neuhäuser, and Ruxton (2013), Fisher (1995), Mardia and Jupp (2009).

3.6 Ejercicio

Utiliza la función `atan2` de R junto con `cos` y `sin` para seno y coseno para estimar el día promedio en el que ocurren más delitos según la base `conteo.dia`.

2. Total poblacional (ver Särndal, Swensson, and Wretman (2003)) Dado un vector de datos numéricos $\vec{x} = (x_1, x_2, \dots, x_n)^T$ definimos el total poblacional como:

$$t_{\vec{x}} = \sum_{i=1}^n x_i, \quad x_i \in \mathbb{R}$$

En este caso de las carpetas de investigación el total nos daría todas las carpetas abiertas durante diciembre. Para ello calculamos el total sumando todos los elementos:

```
sum(conteo_delitos$n)
```

```
## [1] 19861
```

O bien (y esto es una de las cosas interesantes de `tidyverse`) agregándolo a los cálculos previos:

```
conteo_delitos %>% summarise(mean(n), sum(n))
```

```
## # A tibble: 1 x 2
##   `mean(n)` `sum(n)`
##     <dbl>     <int>
## 1       641.    19861
```

3. Varianza poblacional (no ajustada) Dado un vector de datos numéricos $\vec{x} = (x_1, x_2, \dots, x_n)^T$ definimos la *varianza poblacional* como¹:

$$\sigma_{\vec{x}}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad x_i \in \mathbb{R}$$

Misma que podemos calcular con el comando `var` ya sea directamente en la columna:

```
var(conteo_delitos$n)
```

```
## [1] 10046.23
```

O bien a través del `summarise` integrando con el anterior:

```
conteo_delitos %>% summarise(mean(n), sum(n), var(n))
```

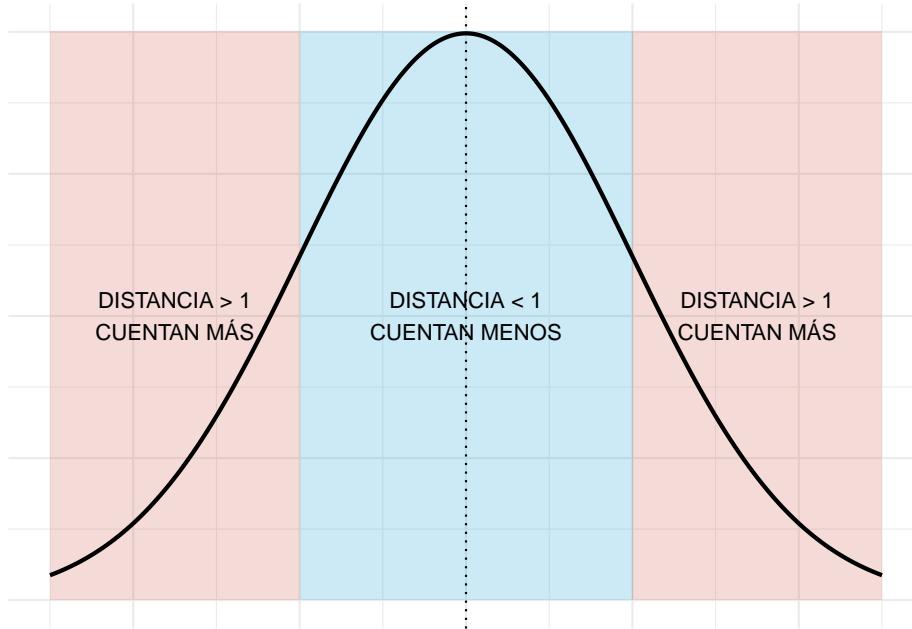
```
## # A tibble: 1 x 3
##   `mean(n)` `sum(n)` `var(n)`
##     <dbl>     <int>     <dbl>
## 1       641.    19861    10046.
```

¹En Panaretos (2016) se conoce como `sample variance` pero para nosotros la muestral tendrá un $n - 1$ en el denominador

La raíz cuadrada de la varianza se conoce como **desviación estándar** y se denota como sigue:

$$\sigma_{\bar{x}} = \sqrt{\sigma_x^2}$$

Recuerda que la varianza se interpreta como la distancia cuadrática promedio a la que están los datos. En particular la varianza casi no considera valores que están a menos de 1 de distancia de \bar{x} (pues $(x_i - \bar{x})^2 < 1$ en ese caso) pero le da mayor peso a valores que están muy lejanos (donde $(x_i - \bar{x})^2 \gg 1$ si x_i está muy lejos de \bar{x}). Gráficamente:



Si nos interesaría que todos los valores (tanto los cercanos a \bar{x} como los lejanos) pesaran de manera idéntica entonces usaríamos el MAD:

3.1 Varianza angular (circular) En el caso de datos circulares, Pewsey, Neuhäuser, and Ruxton (2013) define la varianza circular como:

$$\text{Var} = 1 - \bar{r}$$

donde $\bar{r} = (\bar{c}^2 + \bar{s}^2)^{1/2}$ es el resolvente explicado anteriormente.

4. Desviación Media Absoluta (MAD) Dado un vector de datos numéricos $\vec{x} = (x_1, x_2, \dots, x_n)^T$ definimos la desviación media absoluta, MAD, como (Panaretos (2016)):

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

Misma que se puede calcular en R como:

```
mad(conteo_delitos$n, center = mean(conteo_delitos$n))
```

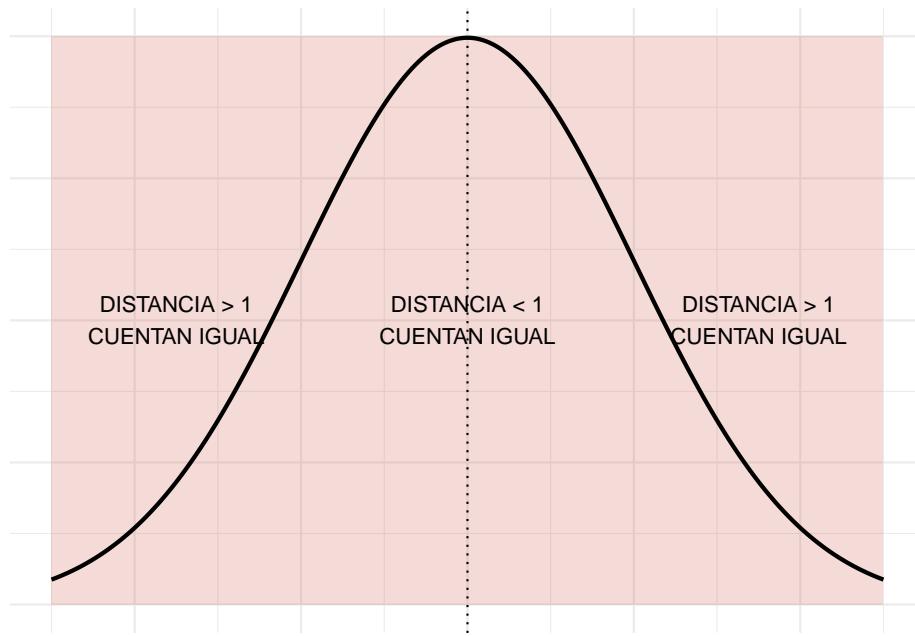
```
## [1] 114.6385
```

o bien dentro del `summarise`:

```
conteo_delitos %>% summarise(mean(n), sum(n), var(n),
                                mad(n, center = mean(n)))
```

```
## # A tibble: 1 x 4
##   `mean(n)` `sum(n)` `var(n)` `mad(n, center = mean(n))` 
##     <dbl>      <int>     <dbl>                <dbl>
## 1       641.     19861    10046.               115.
```

La MAD también es una forma de medir distancia pero en este caso se tiene que todos aportan por igual los muy alejados y los que no:



Para pensarle: En el caso de una variable que se supone que es uniforme y no interesa penalizar valores lejanos de la media. ¿cuál sería una mejor manera de cuantificar la dispersión MAD ó varianza? ¿en qué casos importaría la otra?

Las siguientes dos definiciones son con base en conceptos de proba. ¿Los recuerdas?

5. Coeficiente de asimetría Dado un vector de datos numéricos $\vec{x} =$

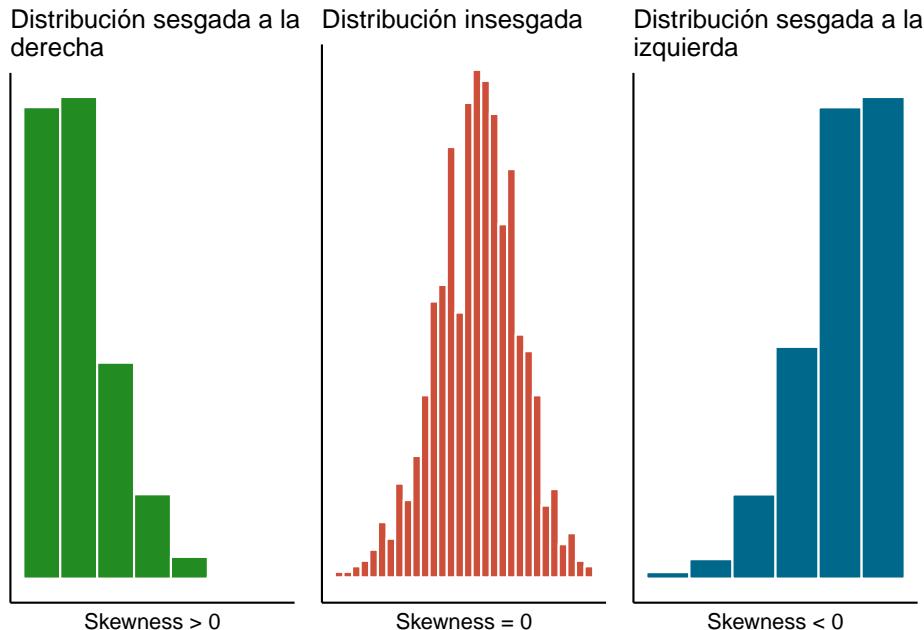
$(x_1, x_2, \dots, x_n)^T$ definimos el coeficiente de asimetría de Fisher (*skewness*) como:

$$\text{Skewness}_{\bar{x}} = \frac{1}{n\sigma_{\bar{x}}^3} \sum_{i=1}^n (x_i - \bar{x})^3$$

Para más referencias ver Panaretos (2016). A fin de interpretar el coeficiente de asimetría podemos dividir esa suma en dos pedazos (olvidándonos de la constante):

$$\sum_{i=1}^n (x_i - \bar{x})^3 = \underbrace{\sum_{\substack{i=1 \\ x_i > \bar{x}}}^n (x_i - \bar{x})^3}_{A} + \underbrace{\sum_{\substack{i=1 \\ x_i < \bar{x}}}^n (x_i - \bar{x})^3}_{B}$$

Notamos que si $|A| > |B|$ la mayor parte de las x_i (o las que se alejan más de la media) son mayores a \bar{x} y por tanto los datos van a estar *sesgados a la derecha*. Por otro lado si $|B| > |A|$ significa que hay más x_i (o con mayor peso) del lado izquierdo de la media que del lado derecho de la misma y por tanto los datos están *sesgados a la izquierda*. Datos *insesgados* son aquellos donde $\text{Skewness}_{\bar{x}} = 0$.



En el caso de las carpetas podemos calcular la asimetría que no se encuentra preprogramada en R como sigue:

```
#Estimación de la desviación estándar
desv.est <- sd(conteo_delitos$n)
```

```
#Estimación del x barra
x.barra <- mean(conteo_delitos$n)

#Obtención de la n (longitud del vector)
n.longitud <- length(conteo_delitos$n)

#Cálculo de la asimetría
(1/desv.est^3)*mean((conteo_delitos$n - x.barra)^3)
```

```
## [1] -0.4528209
```

¿Qué implica el resultado anterior?

6. Curtosis Dado el mismo vector \vec{x} que en el enunciado anterior el coeficiente de curtosis se define como

$$\text{Curtosis}_{\vec{x}} = \frac{1}{n\sigma_{\vec{x}}^4} \sum_{i=1}^n (x_i - \bar{x})^4$$

La interpretación de la curtosis es similar a la que hicimos de la varianza en el sentido que el elevar a la cuarta va a magnificar los efectos de aquellos valores que estén a más de σ de distancia de la media pues podemos reescribir la suma como:

$$\frac{1}{n\sigma_{\vec{x}}^4} \sum_{i=1}^n (x_i - \bar{x})^4 = \underbrace{\frac{1}{n\sigma_{\vec{x}}^4} \sum_{\substack{i=1 \\ |x_i - \bar{x}| < \sigma}}^n (x_i - \bar{x})^4}_A + \underbrace{\frac{1}{n\sigma_{\vec{x}}^4} \sum_{\substack{i=1 \\ |x_i - \bar{x}| > \sigma}}^n (x_i - \bar{x})^4}_B$$

Notamos que la única parte importante que apota a la curtosis es la dada por **B** que es la que capta las *colas* de la distribución (pues ese lado es $\gg 1$). De ahí que podamos decir que, entre dos vectores de datos, uno tiene colas más pesadas que el otro si su curtosis es mayor. En este caso podemos analizar la **latitud y longitud** de los datos a través de la curtosis:

```
datos %>% summarise(kurtosis(latitud, na.rm = T), kurtosis(longitud, na.rm = T))
```

```
##   kurtosis(latitud, na.rm = T) kurtosis(longitud, na.rm = T)
## 1                      2.857934                      3.045037
```

donde se agregó el comando **na.rm = T** para eliminar los valores de no respuesta (missing) marcados como **NA**. Del análisis notamos que la longitud tiene colas más pesadas que la latitud.

NOTACIÓN Dado un vector $\vec{x} = (x_1, x_2, \dots, x_n)^T$ de valores numéricos denotamos el j -ésimo valor muestral ($1 \leq j \leq n$) como

$x_{(j)}$ tal que $x_{(1)} = \min\{x_1, x_2, \dots, x_n\}$ y

$$x_{(j)} = \min\{x_1, x_2, \dots, x_n\} \setminus \{x_{(1)}, x_{(2)}, \dots, x_{(j-1)}\}$$

Es decir $x_{(j)}$ es el valor en orden j al momento de ordenar la muestra. Como nota adicional se define $x_{(0)} = 0$ y $x_{(n+1)} = 0$.

Nota La curtosis a veces se define con un denominador distinto (en términos de las n) como en Myatt and Johnson (2007).

7. Mediana Dado un vector de valores numéricos $\vec{x} = (x_1, x_2, \dots, x_n)^T$ definimos la mediana como (Panaretos (2016)):

$$\text{Mediana}_{\vec{x}} = \frac{x_{(\lfloor \frac{n+1}{2} \rfloor)} + x_{(\lceil \frac{n+1}{2} \rceil)}}{2}$$

La mediana puede calcularse fácilmente haciendo:

```
median(conteo_delitos$n)
```

```
## [1] 646
```

8. Cuantil Dado un vector de valores numéricos $\vec{x} = (x_1, x_2, \dots, x_n)^T$ el α -ésimo cuantil está dado por:

$$\text{Cuantil}_{\vec{x}}(\alpha) = \frac{x_{(\lfloor \alpha \cdot (n+1) \rfloor)} + x_{(\lceil \alpha \cdot (n+1) \rceil)}}{2}$$

donde $x_{(0)} = x_{(n+1)} = 0$. R no calcula los cuantiles de manera exacta sino que por velocidad los aproxima mediante la función `quantile`. Por ejemplo en el cálculo de los cuantiles $\alpha = 0.1$ y $\alpha = 0.66$:

```
conteo_delitos %>% summarise(quantile(n, c(0.1, 0.66)))
```

```
## # A tibble: 2 x 1
##   `quantile(n, c(0.1, 0.66))`<dbl>
## 1 501
## 2 707
```

La función `summary` también es bastante útil resumiendo múltiples observaciones de la base:

```
summary(conteo_delitos$n)
```

```
##      Min. 1st Qu. Median Mean 3rd Qu. Max.
## 397.0 568.0 646.0 640.7 721.0 790.0
```

Ésta incluye los **cuartiles** los cuales corresponden a los cuantiles asociados a $\alpha = 0.25, 0.5, 0.75$ y 1 .

Nota Hay múltiples definiciones de cuantil (ver Hyndman and Fan (1996) para un intento de homologación). En particular R utiliza

una distinta y tus cómputos no van a coincidir si lo haces con esta definición y con la de R. Si quieres saber más de R consulta `?quantile`

9. Rango intercuartílico Definimos el rango intercuartílico (Panaretos (2016)) para valores numéricos $\vec{x} = (x_1, x_2, \dots, x_n)^T$ como la distancia entre el cuartil 0.75 y el 0.25 (primer y tercer cuartil):

$$\text{IQR}_{\vec{x}} = \text{Cuartil}_{\vec{x}}(0.75) - \text{Cuartil}_{\vec{x}}(0.25)$$

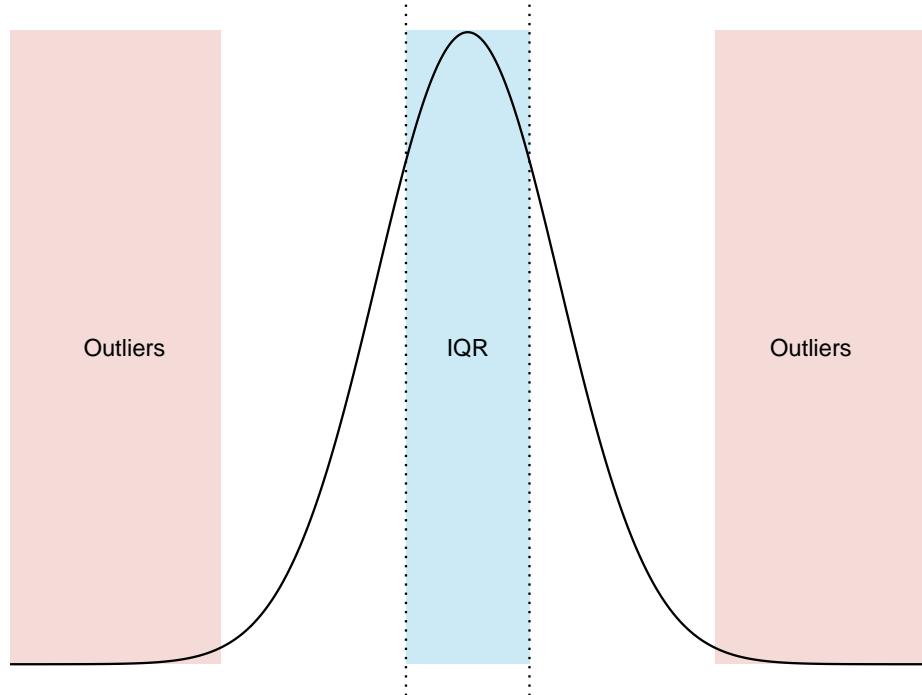
```
IQR(conteo_delitos$n)
```

```
## [1] 153
```

10. Valores atípicos (outliers) Dado un vector de datos numéricos $\vec{x} = (x_1, x_2, \dots, x_n)^T$ seguimos a Panaretos (2016) para definir los valores atípicos *outliers* como aquellas observaciones:

$$\text{Outliers}_{\vec{x}} = \left\{ x_i \in \vec{x} \mid x_i \notin \left[\text{Cuartil}_{\vec{x}}(0.25) - \frac{3}{2} \text{IQR}_{\vec{x}}, \text{Cuartil}_{\vec{x}}(0.75) + \frac{3}{2} \text{IQR}_{\vec{x}} \right] \right\}$$

Los *outliers* en esta definición son valores que serían verdaderamente improbables bajo una distribución normal.



Particularmente en el caso de la normal los *outliers* son valores que tienen una probabilidad de salir aproximadamente de 0.0069766 (por eso son atípicos porque no se esperaría que aparecieran nunca).

Para identificar los *outliers* calculamos el IQR primero y los cuartiles:

```
iqr      <- IQR(conteo_delitos$n)
cuartil1 <- quantile(conteo_delitos$n, 0.25)
cuartil3 <- quantile(conteo_delitos$n, 0.75)
```

después identificamos el límite inferior y superior del conjunto

```
lim.inf <- cuartil1 - 3/2*iqr
lim.sup <- cuartil3 + 3/2*iqr
```

finalmente preguntamos por cuáles están antes o después:

```
outliers <- conteo_delitos %>% filter(n < lim.inf | n > lim.sup)
```

En este caso no tenemos *outliers*.

NOTA Según la aplicación que tenemos la definición de *outlier* cambia. La actual es la que se utiliza para datos que pudieran ser descritos mediante una Normal; empero, no siempre esta definición de *outlier* es un buen modelo (por ejemplo en datos como ingreso que son cantidades positivas, con mucha asimetría y cola pesada). Un buen tratamiento sobre los *outliers* puedes encontrarlo en SURI, Murty, and Athithan (2019).

11. Rango El rango (Peck, Olsen, and Devore (2015)) se define como la diferencia entre el mínimo y el máximo de los valores de un vector numérico $\vec{x} = (x_1, x_2, \dots, x_n)^T$:

$$\text{Rango}_{\vec{x}} = \max\{x_1, x_2, \dots, x_n\} - \min\{x_1, x_2, \dots, x_n\}$$

En R puede calcularse con la resta:

```
#Obtenemos máximo y mínimo
maximo <- max(conteo_delitos$n)
minimo <- min(conteo_delitos$n)

#Rango
maximo - minimo
```

```
## [1] 393
```

Nota En algunos casos el rango se refiere al intervalo $[a, b]$ de valores donde $a = \min\{x_1, x_2, \dots, x_n\}$ y $b = \max\{x_1, x_2, \dots, x_n\}$. Éste es el caso de la función `range` en R:

```
range(conteo_delitos$n)
```

```
## [1] 397 790
```

12. Conteo asociado a un conjunto Sea $\vec{y} = (y_1, y_2, \dots, y_n)^T$ un vector de datos de cualquier tipo (numéricos, categóricos, lógicos, caracteres, etc). Para

un conjunto A definimos el conteo asociado al conjunto A como:

$$\text{Conteo}_{\bar{y}}(A) = \sum_{i=1}^n \mathbb{I}_A(y_i)$$

donde

$$\mathbb{I}_A(y) = \begin{cases} 1 & \text{si } y \in A, \\ 0 & \text{en otro caso ,} \end{cases}$$

es una variable indicadora.

Una forma rápida de obtener dicho conteo en R es mediante `table`:

```
table(datos$delito)
```

```
##          ABANDONO DE PERSONA           ABORTO   ABUSO DE AUTORIDAD   ABUSO DE CONFIANZA
##                53                      15                  102                   276
##          ABUSO SEXUAL           ACOSO SEXUAL
##                252                     30
```

O bien si se desean contar en la base de datos por ejemplo los delitos de **ABANDONO DE PERSONA** pueden hacerse mediante un filtro.

```
datos %>% filter(delito == "ABANDONO DE PERSONA") %>% tally()
```

```
## # ... with 11 more rows
```

Al filtro pueden agregárseles grupos por si se desea obtener por fecha:

```
datos %>% filter(delito == "ABANDONO DE PERSONA") %>%
  group_by(fecha) %>% tally()
```

```
## # A tibble: 21 x 2
##       fecha     n
##       <date>   <int>
## 1 2018-12-01     3
## 2 2018-12-02     3
## 3 2018-12-04     2
## 4 2018-12-05     8
## 5 2018-12-06     1
## 6 2018-12-07     1
## 7 2018-12-10     1
## 8 2018-12-12     2
## 9 2018-12-13     3
## 10 2018-12-14    2
## # ... with 11 more rows
```

El filtro funciona igual que un `if` pudiéndose usar `and` (`&`) u `or` (`|`):

```

datos %>%
  filter(delito == "ABANDONO DE PERSONA" | delito == "ABORTO") %>%
  group_by(fecha) %>% tally()

## # A tibble: 25 x 2
##   fecha      n
##   <date>    <int>
## 1 2018-12-01     3
## 2 2018-12-02     3
## 3 2018-12-04     5
## 4 2018-12-05     8
## 5 2018-12-06     1
## 6 2018-12-07     2
## 7 2018-12-10     1
## 8 2018-12-12     2
## 9 2018-12-13     4
## 10 2018-12-14    2
## # ... with 15 more rows

datos %>%
  filter(delito == "ABANDONO DE PERSONA" &
         fiscalía == "INVESTIGACIÓN EN IZTAPALAPA") %>%
  group_by(fecha) %>% tally()

## # A tibble: 3 x 2
##   fecha      n
##   <date>    <int>
## 1 2018-12-02     1
## 2 2018-12-13     1
## 3 2018-12-20     1

```

13. Moda En términos simples, la moda es el conjunto de los valores que más se repiten. Matemáticamente (ver Peck, Olsen, and Devore (2015)) la moda es el conjunto $\text{Moda}_{\vec{y}} = \{m_1, m_2, \dots, m_k\}$ tal que $m \in \text{Moda}$ sí y sólo si

$$\sum_{i=1}^n \mathbb{I}_{\{m\}}(y_i) \geq \sum_{i=1}^n \mathbb{I}_{\{\ell\}}(y_i) \quad \forall \ell \neq m \text{ donde } y_i \in \vec{y}.$$

Para calcularla en R no existe una función predefinida para calcular la moda. Nosotros podemos crearla con el comando `function`. El término `function` nos sirve para construir funciones; por ejemplo, una función que eleva al cuadrado:

```

elevar_cuadrado <- function(x){
  return(x^2)
}

```

Observa la estructura que siempre será de esta forma:

```
nombre de la función <- function(parámetro, otro parámetro){
  #Lo que sea que haga
  return(lo que devuelve)
}
```

Podemos llamar a la función con un número:

```
elevar.cuadrado(8)
```

```
## [1] 64
```

o bien con un vector:

```
elevar.cuadrado(12)
```

```
## [1] 144
```

En nuestro caso vamos a crear una función que se llame `moda` para estimar la moda:

```
#Función para estimar la moda de un vector x
moda <- function(x){

  #Contar cuántas veces aparecen las observaciones
  conteo <- table(x)

  #Obtengo el máximo que aparece
  max_aparece <- max(conteo)

  #Busco cuáles aparecen más y obtengo los nombres
  moda <- names(conteo)[which(conteo == max_aparece)]

  #Finalmente checo que si los datos eran numéricos moda debe
  #ser numérico
  if (is.numeric(x)){
    moda <- as.numeric(moda)
  }

  return(moda)
}
```

Podemos probar nuestra función con datos que ya sepamos su resultado nada más para asegurarnos que funciona:

```
#Creamos un vector numérico con dos modas
vector.ejemplo.1 <- c(1,6,6,1,2,7,8,10)
moda(vector.ejemplo.1)
```

```
## [1] 1 6
```

Podemos probarlo también con caracteres:

```
#Creamos un vector numérico con dos modas
vector.ejemplo.2 <- c("manzana", "pera", "guayaba", "perejil", "manzana")
moda(vector.ejemplo.2)
```

```
## [1] "manzana"
```

Una vez sabemos funciona podemos buscar el delito que ocurrió más:

```
moda(datos$delito)
```

```
## [1] "VIOLENCIA FAMILIAR"
```

3.7 Ejercicios

- Construye una función que tome de input dos variables: x un vector y k un entero de tal manera que calcule el k -ésimo **momento central** de los datos:

$$\text{Momento}_{\vec{x}}(k) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$$

La función debe tener la siguiente estructura:

```
kesimo.momento <- function(x, k){
  #Rellena aquí
}
```

- Sin usar la opción de `trim` ni `trimmed.mean` crea una función que calcule la media de los datos que están entre el cuantil $\alpha/2$ y el cuantil $1 - \alpha/2$ ($0 \leq \alpha \leq 1$). A esta media se le conoce como **media truncada al nivel $\alpha \times 100\%$** . Matemáticamente se define como:

$$\text{Media Truncada}_{\vec{x}}(\alpha) = \frac{1}{n_\alpha} \sum_{i=1}^n x_i \cdot \mathbb{I}_{[q_{\alpha/2}, q_{1-\alpha/2}]}(x_i)$$

donde $n_\alpha = \sum_{i=1}^n \mathbb{I}_{[q_{\alpha/2}, q_{1-\alpha/2}]}(x_i)$ es la cantidad de x_i que están en el intervalo $[q_{\alpha/2}, q_{1-\alpha/2}]$ donde $q_{\alpha/2} = \text{Cuantil}_{\vec{x}}(\alpha/2)$ y $q_{1-\alpha/2} = \text{Cuantil}_{\vec{x}}(1 - \alpha/2)$.

- Una función llamada `jesimo.dato` de dos argumentos que dado un vector de datos \vec{x} me devuelva el j -ésimo dato ordenado (es decir el $x_{(j)}$). **NOTA** No confundir con devolver el x_j que es la j -ésima entrada. Como sugerencia usar `arrange`, `order` ó `sort`. Un ejemplo de lo que debe hacer la función es:

```
x <- c(12, 8, 9, 7, 14, 21)
jesimo.dato(x, 4)
```

```
## [1] 12
```

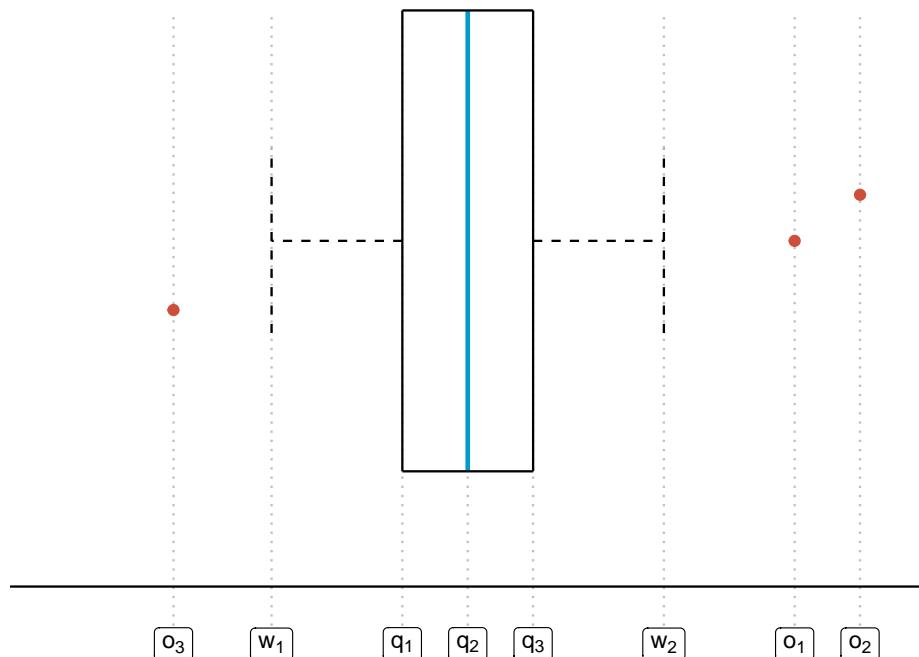
3.8 Gráficas univariadas

1. Gráfica de caja (boxplot) Una gráfica de caja pretende resumir los cuartiles, la mediana e identificar los *outliers* todo en una sola imagen (Panaretos (2016)). Para ello considera un vector numérico $\vec{x} = (x_1, x_2, \dots, x_n)^T$ tal que:

1. q_1 sea el primer cuartil ($\text{Cuantil}_{\vec{x}}(0.25)$), q_2 sea la mediana (que es lo mismo que el segundo cuartil o bien $\text{Cuantil}_{\vec{x}}(0.5)$) y q_3 corresponda al tercer cuartil ($\text{Cuantil}_{\vec{x}}(0.75)$).
2. $w_1 = \min\{x_j \in \vec{x} | x_j \geq q_1 - \frac{3}{2}\text{IQR}\}$ es el valor más pequeño de \vec{x} que *no es outlier* y $w_2 = \max\{x_j \in \vec{x} | x_j \leq q_3 + \frac{3}{2}\text{IQR}\}$ es el valor más grande de \vec{x} que *no es outlier*.
3. Sea $\text{Outliers}_{\vec{x}}$ el conjunto de outliers como lo definimos anteriormente:

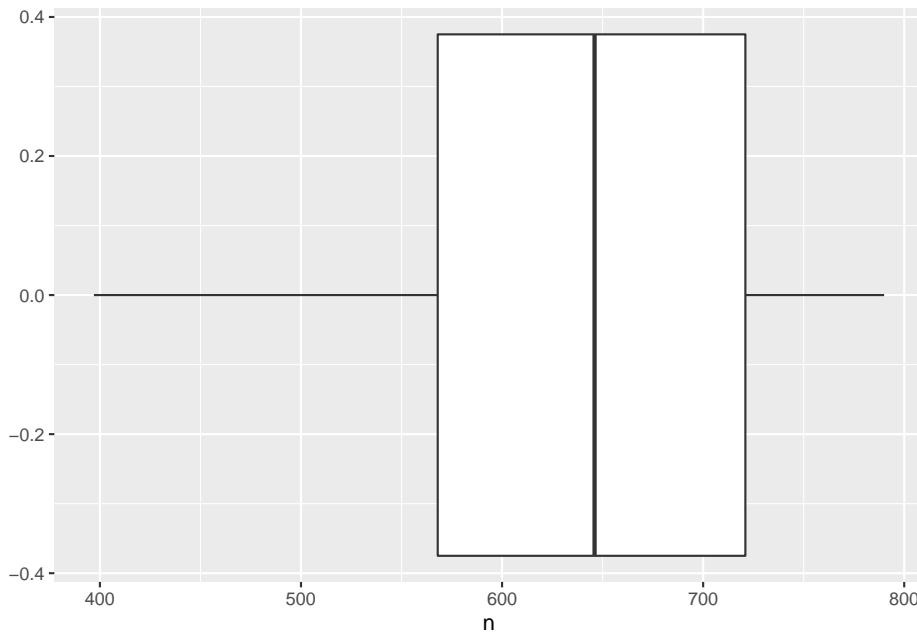
$$\text{Outliers}_{\vec{x}} = \left\{ x_i \in \vec{x} \mid x_i \notin [q_1 - \frac{3}{2}\text{IQR}_{\vec{x}}, q_3 + \frac{3}{2}\text{IQR}_{\vec{x}}] \right\}$$

donde $\text{Outliers}_{\vec{x}} = \{o_1, o_2, \dots, o_d\}$. Una gráfica de caja corresponde al siguiente diagrama:



La imagen anota la mediana, los cuartiles así como el rango de valores donde se sabe que no hay outliers. Finalmente la gráfica identifica los *outliers* si es que hay. Para armar una gráfica de boxplot usamos la librería de *ggplot2* especificando dentro de la función `ggplot` la base de datos de donde sale nuestra información:

```
ggplot(conteo_delitos) +
  geom_boxplot(aes(x = n))
```



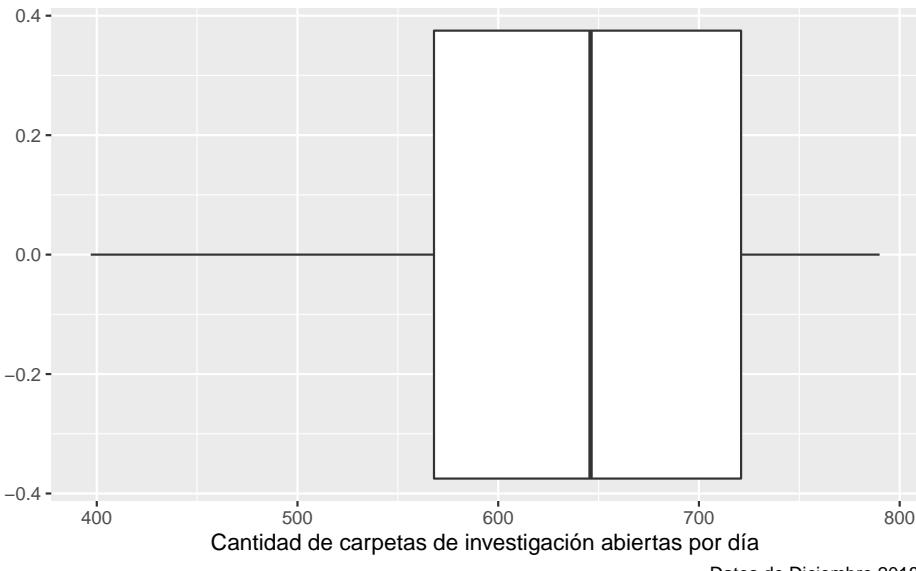
la cual pone la mediana en 646 como habíamos calculado, los cuartiles en 568 y 721 respectivamente. Finalmente no presenta *outliers* pues nuestro análisis previo nos mostraba que no había *outliers*.

Podemos personalizar nuestra gráfica agregando títulos con la función lab:

```
ggplot(conteo_delitos) +
  geom_boxplot(aes(x = n)) +
  labs(
    x = "Cantidad de carpetas de investigación abiertas por día",
    y = "",
    title = "Gráfica de cajas de los delitos en CDMX",
    subtitle = "Fuente: Carpetas de investigación FGJ de la Ciudad de México",
    caption = "Datos de Diciembre 2018"
  )
```

Gráfica de cajas de los delitos en CDMX

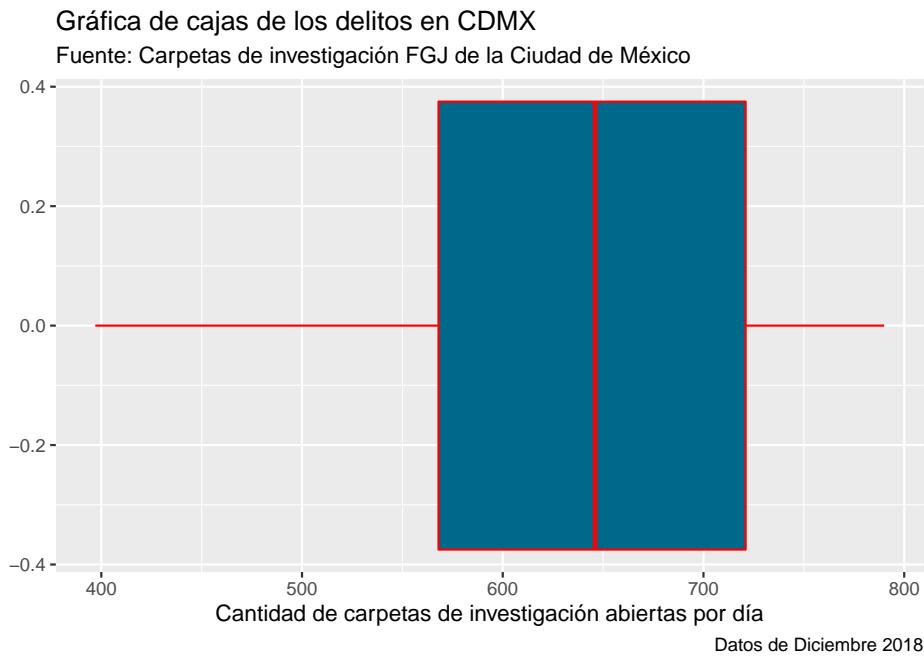
Fuente: Carpetas de investigación FGJ de la Ciudad de México



Datos de Diciembre 2018

Finalmente, podemos personalizar los colores de la gráfica editando directamente en el `geom_boxplot`:

```
ggplot(conteo_delitos) +
  geom_boxplot(aes(x = n), color = "red", fill = "deepskyblue4") +
  labs(
    x = "Cantidad de carpetas de investigación abiertas por día",
    y = "",
    title = "Gráfica de cajas de los delitos en CDMX",
    subtitle = "Fuente: Carpetas de investigación FGJ de la Ciudad de México",
    caption = "Datos de Diciembre 2018"
  )
```



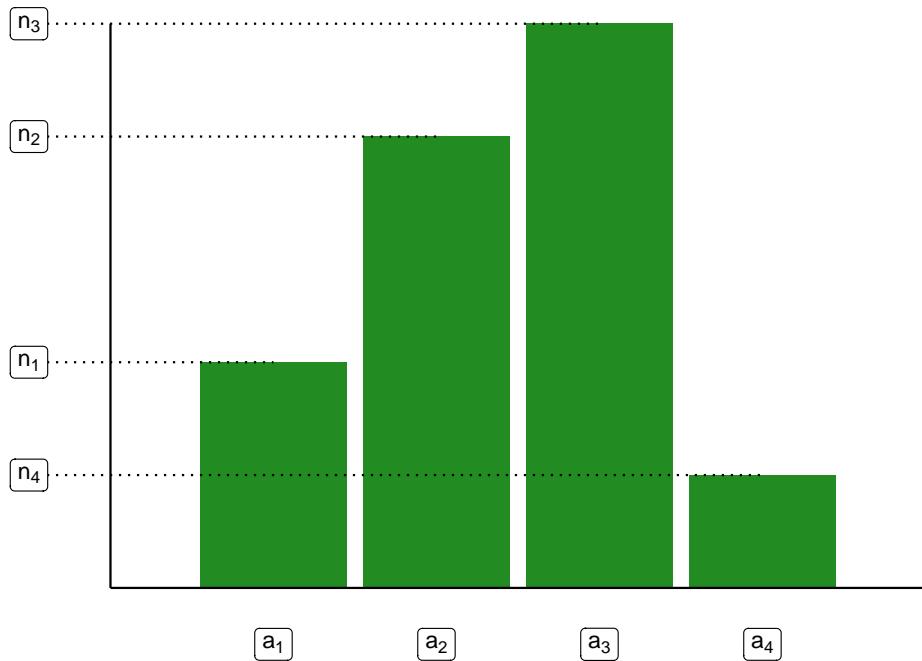
2. Gráfica de barras Sea $\vec{c} = (c_1, c_2, \dots, c_n)^T$ un vector de datos categóricos. Sea $C = \{a_i | a_i \in \vec{c}\}$ el conjunto de ℓ valores únicos que se tienen registrados en el vector \vec{c} . Denotamos la cantidad de veces que aparece a_i en \vec{c} como n_i ; es decir:

$$n_i = \sum_{k=1}^n \mathbb{I}_{\{a_i\}}(c_k)$$

Una gráfica de barras consiste en una representación gráfica del conjunto:

$$\text{Barras} = \{(a_i, n_i) | a_i \in C\}$$

Gráficamente:

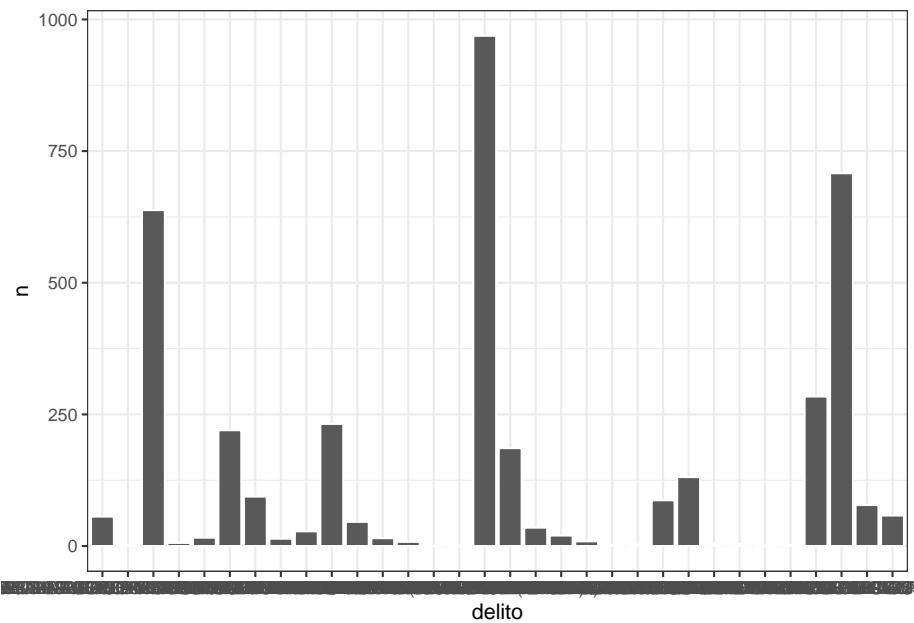


Podemos crear una gráfica de barras con el comando `geom_col` para ello creamos unas barras correspondientes al tipo de delito (sólo en delitos que `categoria_delito` dice ROBO) haciendo una nueva base que cuente por delito:

```
conteo_tipo <- datos %>% filter(str_detect(categoría_delito,"ROBO")) %>%
  group_by(delito) %>% tally()
```

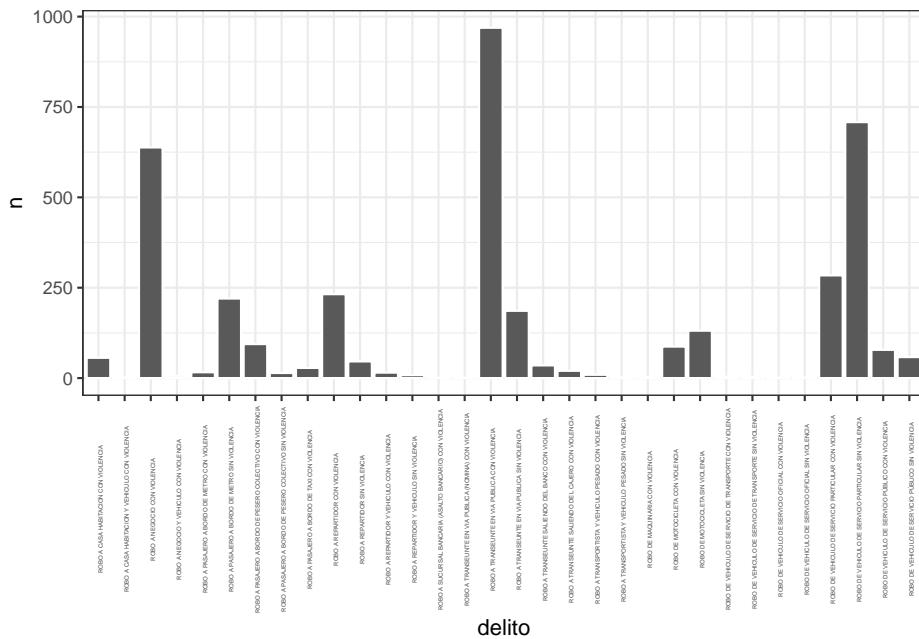
Y hagamos la gráfica:

```
ggplot(conteo_tipo) +
  geom_col(aes(x = delito, y = n), color = "white") +
  theme_bw()
```



Para evitar que se encime todo el texto podemos establecer un ángulo del mismo al usar `theme`:

```
ggplot(conteo_tipo) +  
  geom_col(aes(x = delito, y = n), color = "white") +  
  theme_bw() +  
  theme(axis.text.x = element_text(angle = 90, size = 3))
```

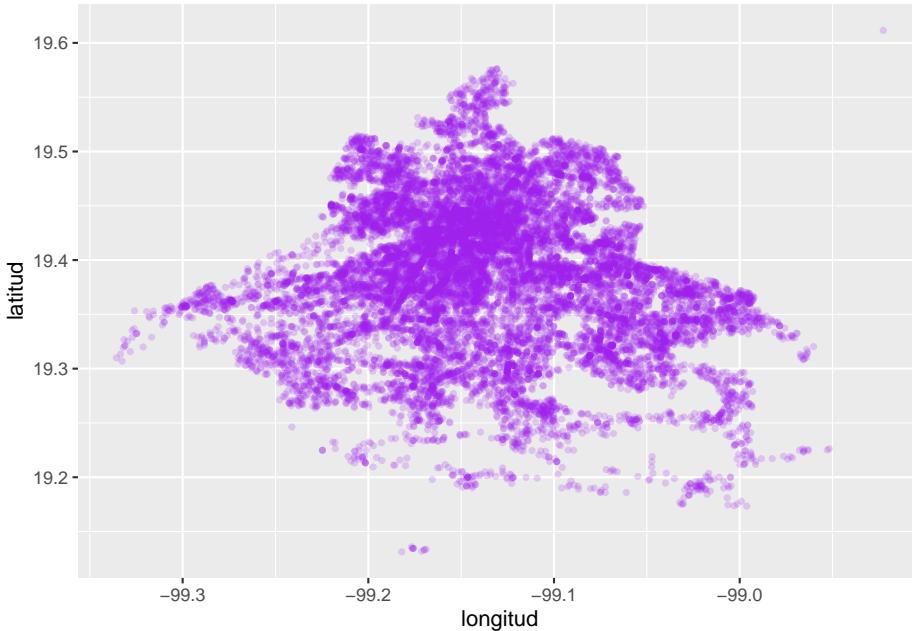


NOTA Una mala praxis es usar gráficas de pay pues es muy complicado contar una historia a partir de ellas. ¡No lo hagas!

3.9 Gráficas bivariadas

1. Gráfica de puntos (scatterplot) Dada una matriz de datos Z consideramos dos columnas numéricas z_i y z_j ($i \neq j$) de dicha matriz. Sea $\mathbb{X} = \{(z_{i,1}, z_{j,1}), (z_{i,2}, z_{j,2}), \dots, (z_{i,n}, z_{j,n})\}$ el conjunto de parejas ordenadas correspondientes a dichas columnas. Una gráfica de puntos consiste en la proyección de dichos puntos sobre \mathbb{R}^2 . Para generarla en R podemos usar `ggplot`:

```
ggplot(datos) +  
  geom_point(aes(x = longitud, y = latitud), size = 1, color = "purple",  
             alpha = 0.2)
```



donde los parámetros `size` establecen el tamaño del punto, `color` su color y `alpha` su nivel de transparencia ($0 \leq \alpha \leq 1$).

2. Gráfica de líneas (lineplot) Dada una matriz de datos Z consideramos dos columnas numéricas z_i y z_j ($i \neq j$) de dicha matriz. Sea $\mathbb{X} = \{(z_{i,1}, z_{j,1}), (z_{i,2}, z_{j,2}), \dots, (z_{i,n}, z_{j,n})\}$ el conjunto de parejas ordenadas correspondientes a dichas columnas. Para evitar confusión de subíndices escribiré a las z_i como x y a las z_j como y de tal forma que $\mathbb{X} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. Supongamos, sin pérdida de generalidad que los datos están ordenados según las x : $x_1 \leq x_2 \leq \dots \leq x_n$. Sea f la función de interpolación lineal dada por:

$$f(x) = \begin{cases} y_1 + \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) & \text{si } x_1 \leq x \leq x_2 \\ \vdots \\ y_{k-1} + \frac{y_k - y_{k-1}}{x_k - x_{k-1}}(x - x_{k-1}) & \text{si } x_{k-1} \leq x \leq x_k \\ \vdots \\ y_{n-1} + \frac{y_n - y_{n-1}}{x_n - x_{n-1}}(x - x_{n-1}) & \text{si } x_{n-1} \leq x \leq x_n \end{cases}$$

Una gráfica de líneas corresponde a la representación gráfica del conjunto

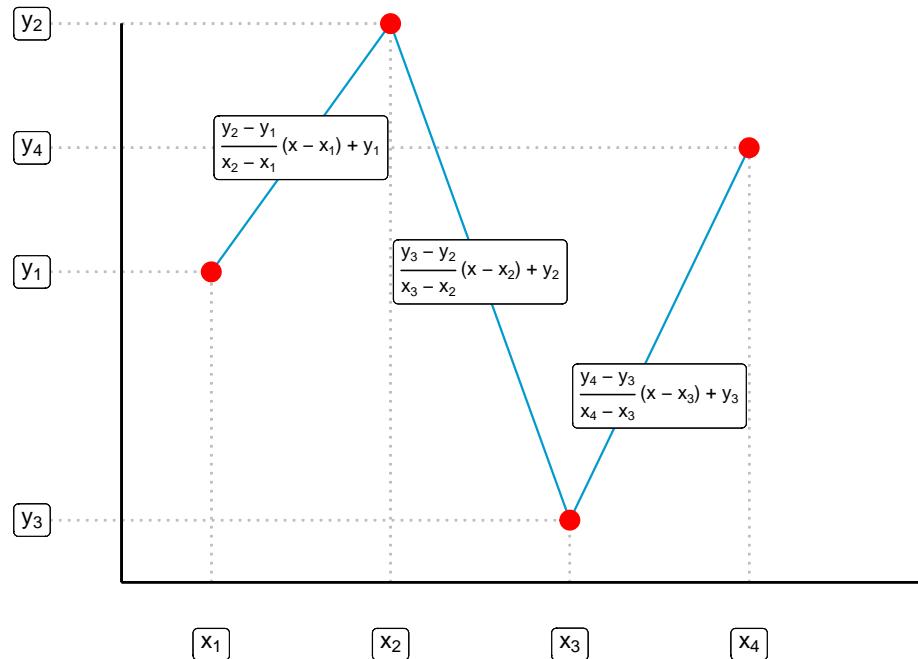
$$\text{Gr}_f = \{(x, f(x)) | x_1 \leq x \leq x_n\}$$

De manera un poco más intuitiva notamos que si tenemos, por ejemplo, $\mathbb{X} = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)\}$ una gráfica de líneas se construye interpolando una línea entre (x_1, y_1) y (x_2, y_2) , otra línea entre (x_2, y_2) y (x_3, y_3)

y, finalmente, otra recta entre (x_3, y_3) y (x_4, y_4) . Usando la ecuación de la línea

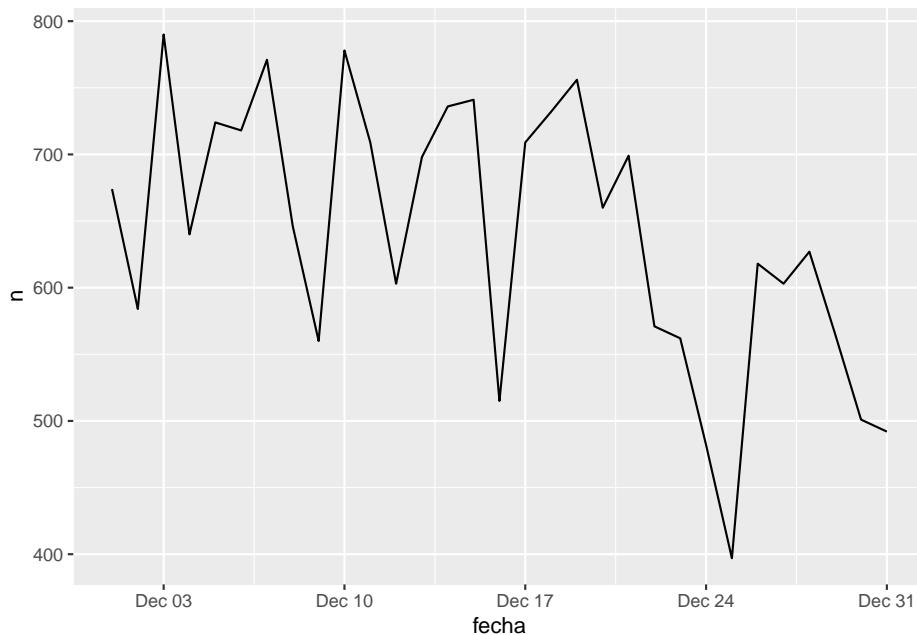
$$y = \frac{y_2 - y_1}{x_2 - x_1} (x - x_1) + y_1$$

interpolamos cada uno de los puntos como en la gráfica siguiente:



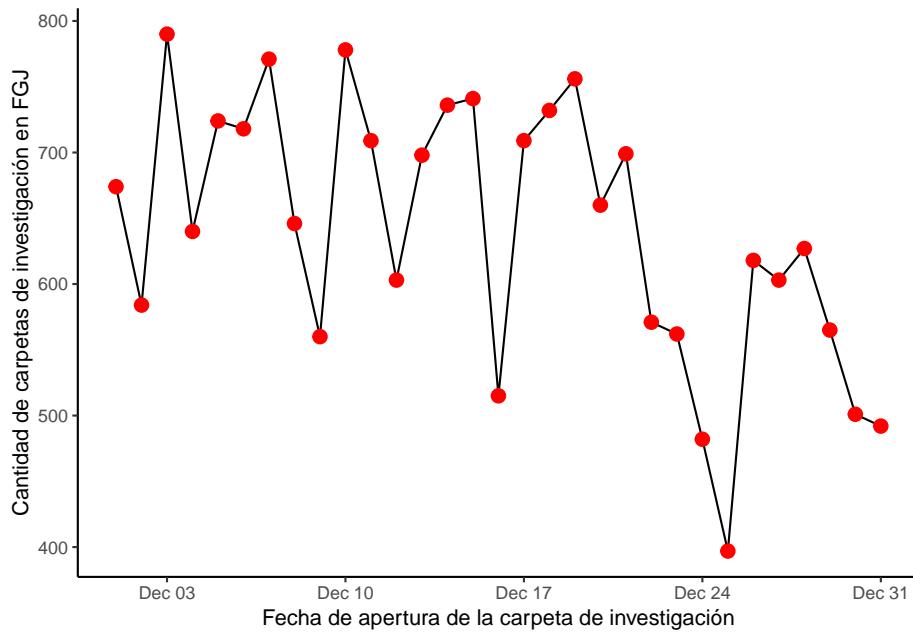
Para realizar una gráfica de líneas podemos usar de nuevo ggplot2 con la opción de `geom_line`:

```
ggplot(conteo_delitos) +
  geom_line(aes(x = fecha, y = n))
```



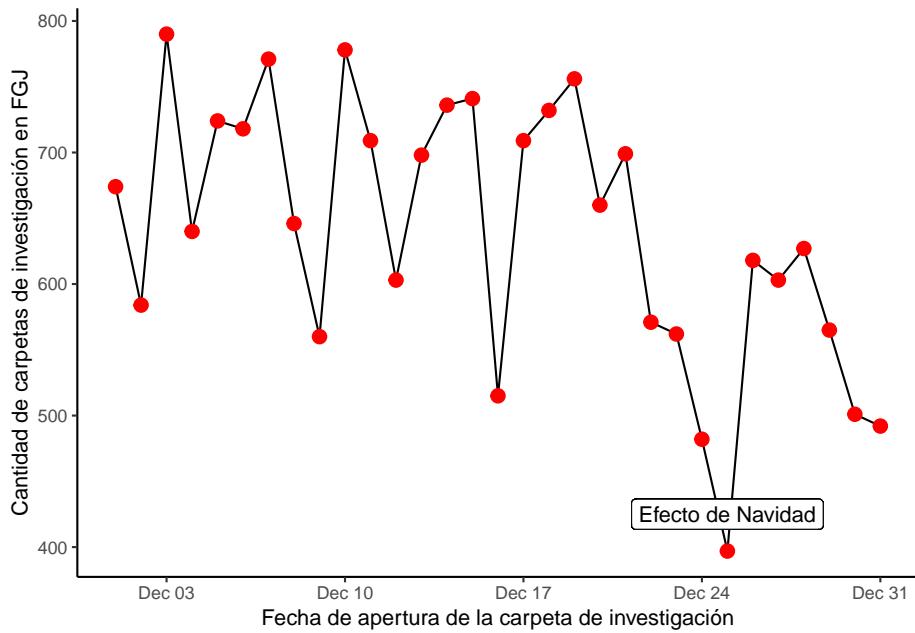
Podemos cambiar el tema y agregar puntos de otro color para que nuestra gráfica se vea más bonita:

```
ggplot(conteo_delitos) +
  geom_line(aes(x = fecha, y = n)) +
  geom_point(aes(x = fecha, y = n), color = "red", size = 3) +
  theme_classic() +
  labs(
    x = "Fecha de apertura de la carpeta de investigación",
    y = "Cantidad de carpetas de investigación en FGJ"
  )
```



Finalmente con `geom_label` podemos agregar anotaciones a nuestra gráfica:

```
ggplot(conteo_delitos) +
  geom_line(aes(x = fecha, y = n)) +
  geom_point(aes(x = fecha, y = n), color = "red", size = 3) +
  theme_classic() +
  labs(
    x = "Fecha de apertura de la carpeta de investigación",
    y = "Cantidad de carpetas de investigación en FGJ"
  ) +
  geom_label(aes(x = dmy("25/12/2018"), y = 425), label = "Efecto de Navidad")
```



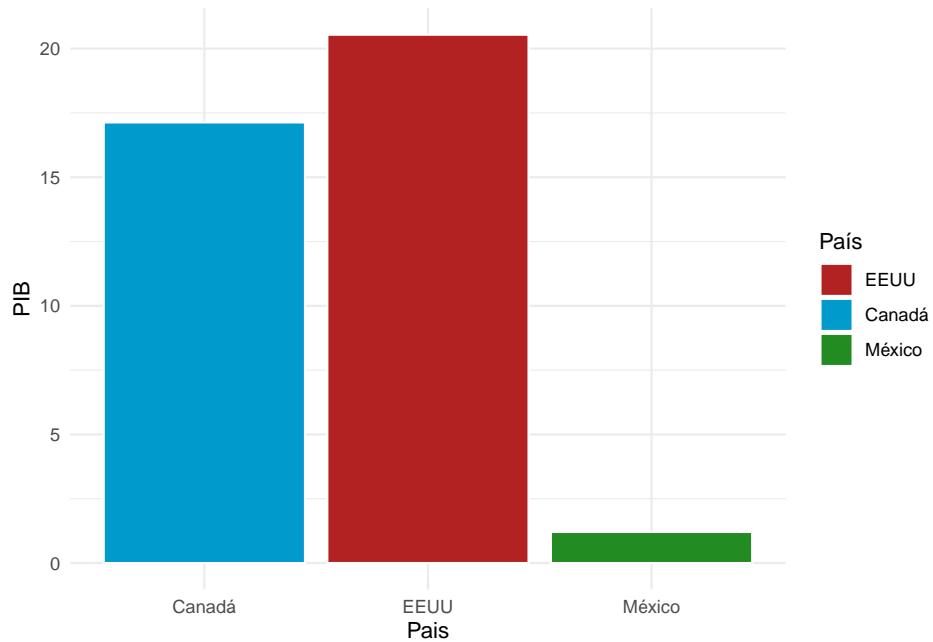
3.9.1 Ejercicio

Utiliza las siguientes bases de datos para replicar exactamente el formato de las gráficas que se muestran abajo de las bases. No todo viene en estas notas, la idea es que investigues y para ello te sugiero consultar este libro

Gráfica de barras

```
datos.barras <- data.frame(Pais = c("EEUU", "Canadá", "Méjico"),
                           PIB = c(20.54, 17.13, 1.21))
```

Los colores usados son firebrick, deepskyblue3 y forestgreen:

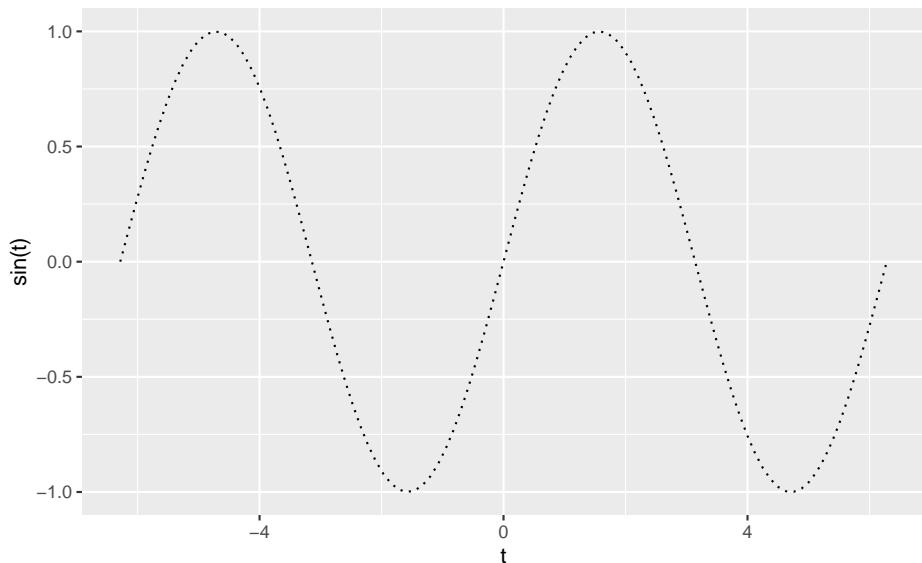


Línea

```
x <- seq(-2*pi, 2*pi, length.out = 100)
datos.linea <- data.frame(x = x, y = sin(x))
```

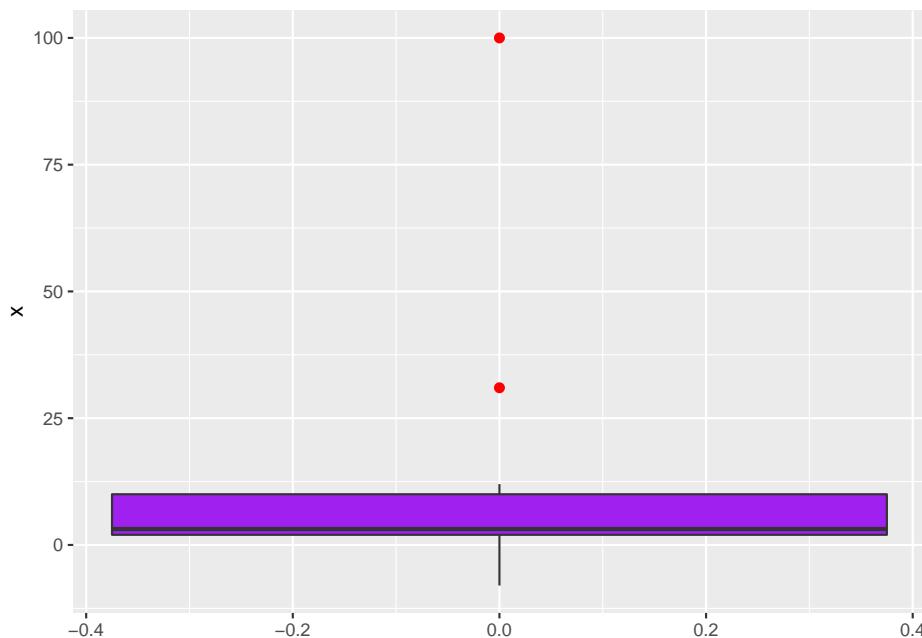
Función seno

Aproximación por computadora



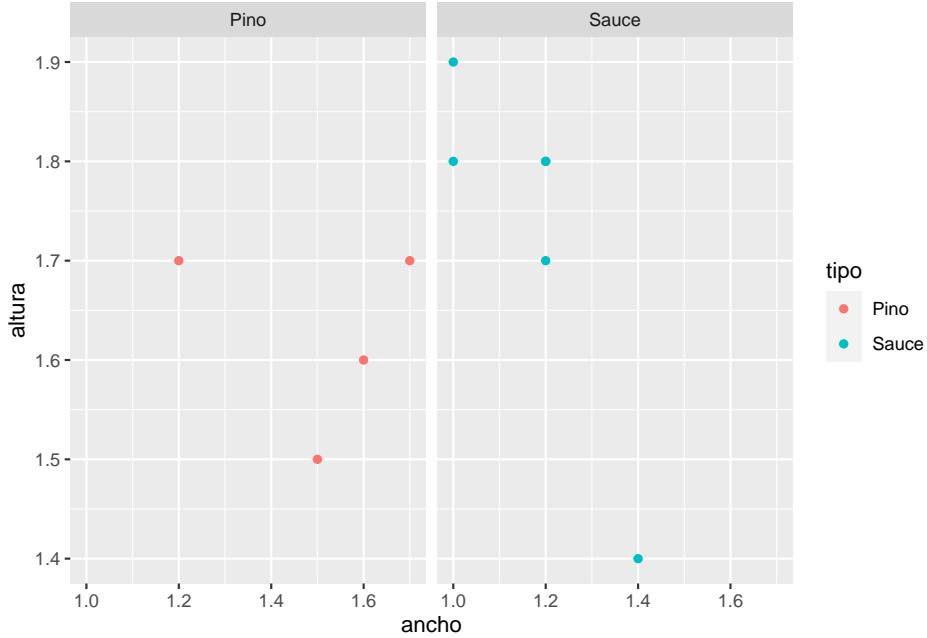
Boxplot

```
x <- c(1,10, 100, -2, 3, 5, 6, 12, -8, 31, 2, pi, 3)
datos.linea <- data.frame(Dientes = x)
```



Puntos

```
datos.arbol <- data.frame(altura = c(1.7, 1.4, 1.8, 1.9, 1.5, 1.7,
                                         1.6, 1.8, 1.7, 1.8),
                           ancho = c(1.2, 1.4, 1.2, 1, 1.5, 1.7, 1.6,
                                     1.2, 1.2, 1),
                           tipo = c("Pino","Sauce","Sauce","Sauce","Pino",
                                   "Pino","Pino","Sauce","Sauce","Sauce"))
```



3.10 Estadísticos bivariados

NOTACIÓN Para esta sección vamos a considerar dos (vectores) columnas de la matriz de datos Z y los denominaremos \vec{x} y \vec{y} (en lugar de z_i y z_j). En particular, denotaremos $\mathcal{X} = \{a_{i,x} | a_{i,x} \in \vec{x}\}$ el conjunto de valores únicos del vector \vec{x} y $\mathcal{Y} = \{a_{y,j} | a_{y,j} \in \vec{y}\}$ el conjunto de valores únicos de \vec{y} . La cardinalidad de dichos conjuntos es ℓ_x y ℓ_y respectivamente. Finalmente, definimos el conteo de cuántas veces aparece el valor $a_{i,x}$ (respectivamente el $a_{y,j}$) en los vectores \vec{x} (respectivamente \vec{y}) como:

$$\begin{aligned} n_{i,x} &= \sum_{k=1}^n \mathbb{I}_{\{a_{i,x}\}}(x_k) \\ n_{y,j} &= \sum_{k=1}^n \mathbb{I}_{\{a_{y,j}\}}(y_k) \end{aligned} \tag{3.1}$$

para $1 \leq i \leq \ell_x$ y $1 \leq j \leq \ell_y$.

Por poner un ejemplo, considera el siguiente conjunto de datos:

En este sentido el vector es $\vec{x} = (\text{Rojo}, \text{Azul}, \text{Verde}, \text{Rojo}, \text{Verde})^T$ mientras que el conjunto de valores únicos asociados está dado por $\mathcal{X} = \{\text{Rojo}, \text{Azul}, \text{Verde}\}$. En este sentido (siguiendo el conjunto) se tiene que $a_{1,x} = \text{Rojo}$, $a_{2,x} = \text{Azul}$ y $a_{3,x} = \text{Verde}$ mientras que (siguiendo el vector) se observa $x_1 = \text{Rojo}$, $x_2 = \text{Azul}$, $x_3 = \text{Verde}$,

x	y
Rojo	Coche
Azul	Taza
Verde	Árbol
Rojo	Taza
Verde	Libro

	Coche	Taza	Árbol	Libro	Total (fila)
Rojo	1	1	0	0	2
Azul	0	1	0	0	1
Verde	0	0	1	1	2
Total (columna)	1	2	1	1	5

$x_4 = \text{Rojo}$, $x_5 = \text{Verde}$. Finalmente notamos que el conteo de veces que aparece cada cosa es: $n_{1,x} = 2$ (aparece el $a_{1,x}$ que es rojo dos veces), $n_{2,x} = 1$ y $n_{3,x} = 2$ (el azul y verde dados por $a_{2,x}$ y $a_{3,x}$ respectivamente aparecen una vez para azul y dos veces para verde). Por otro lado, $\vec{y} = (\text{Coche}, \text{Taza}, \text{Árbol}, \text{Taza}, \text{Libro})^T$ con su conjunto de valores únicos $\mathcal{Y} = \{\text{Coche}, \text{Taza}, \text{Árbol}, \text{Libro}\}$. Para el caso de \vec{y} se tiene que $y_1 = \text{Coche}$, $y_2 = \text{Taza}$, $y_3 = \text{Árbol}$, $y_4 = \text{Taza}$, $y_5 = \text{Libro}$ mientras que en el caso de valores únicos $a_{y,1} = \text{Coche}$, $a_{y,2} = \text{Taza}$, $a_{y,3} = \text{Árbol}$, $a_{y,4} = \text{Libro}$. Los conteos asociados son: $n_{y,1} = n_{y,3} = n_{y,4} = 1$ (aparecen el coche, el árbol y el libro una vez) mientras que $n_{y,2} = 2$ representa que la taza está dos veces.

Por otro lado denotamos a la submatriz de Z compuesta solamente por las columnas \vec{x} y \vec{y} como:

$$Z_{(x,y)} = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}$$

Sea $\mathcal{X} \times \mathcal{Y} = \{a_{i,j} = (x_i, y_j) | x_i \in \mathcal{X} \quad \& \quad y_j \in \mathcal{Y}\}$ el conjunto de observaciones únicas posibles de las parejas (x, y) (todas las permutaciones). Finalmente, el conteo de cuántas veces aparece el vector bivariado $a_{i,j}$ en los datos está dado por:

$$n_{i,j} = \sum_{k=1}^n \mathbb{I}_{\{a_{i,j}\}}((x_k, y_k))$$

En el ejemplo anterior, la tabla se vería:

```
## Warning: Setting row names on a tibble is deprecated.
```

Una excelente referencia para esta sección es el capítulo 4 de Peck, Olsen, and Devore (2015).

1. Tabla de contingencia Para \vec{x} , \vec{y} definidas como al inicio de la sección (y siguiendo la notación anterior), definimos una tabla de contingencia como la matriz $N_{x,y}$ dada por:

$$N_{x,y} = \begin{pmatrix} n_{1,1} & n_{1,2} & \dots & n_{1,\ell_y} \\ n_{2,1} & n_{2,2} & \dots & n_{2,\ell_y} \\ \vdots & \vdots & \ddots & \vdots \\ n_{\ell_x,1} & n_{\ell_x,2} & \dots & n_{\ell_x,\ell_y} \end{pmatrix}$$

Al vector $n_x = (n_{1,x}, n_{2,x}, \dots, n_{\ell_x,x})^T$ se le conoce como **distribución frecuencial (observada) marginal** de \vec{x} mientras que $n_y = (n_{y,1}, n_{y,2}, \dots, n_{y,\ell_y})^T$ es la **distribución frecuencial (observada) marginal** de \vec{y} .

Una tabla de contingencia representa el conteo de observaciones de una variable ajustado por la otra. Para crear una tabla de contingencia en R podemos usar el mismo comando `table` que ya usamos antes pero esta vez introduciendo dos vectores como en el siguiente ejemplo donde notamos alcaldía contra año del registro:

```
table(datos$alcaldia_hechos, datos$ao_inicio)
```

```
##                                     2018 2019
##  VERACRUZ          0    1
##  VILLAGRAN        1    0
##  XALATLACO        2    0
##  XOCHIMILCO     465  115
##  XOCHITEPEC       1    0
##  ZACATECAS         0    2
```

Para agregar las distribuciones frecuenciales marginales a la tabla podemos usar el comando `addmargins`:

```
addmargins(table(datos$alcaldia_hechos, datos$ao_inicio))
```

```
##                                     2018 2019   Sum
##  VILLAGRAN        1    0    1
##  XALATLACO        2    0    2
##  XOCHIMILCO     465  115  580
##  XOCHITEPEC       1    0    1
##  ZACATECAS         0    2    2
##  Sum            15952 3896 19848
```

2. Tabla de frecuencias

Una tabla de frecuencia es la matriz $\text{Freq}_{x,y}$ dada por:

$$\text{Freq}_{x,y} = \begin{pmatrix} f_{1,1} & f_{1,2} & \dots & f_{1,\ell_y} \\ f_{2,1} & f_{2,2} & \dots & f_{2,\ell_y} \\ \vdots & \vdots & \ddots & \vdots \\ f_{\ell_x,1} & f_{\ell_x,2} & \dots & f_{\ell_x,\ell_y} \end{pmatrix}$$

donde $f_{i,j} = \frac{n_{i,j}}{n}$ representa la frecuencia relativa de la observación de $(a_{i,x}, a_{y,j})$ i.e. cuánto representa del total. Al vector $f_x = (f_{1,x}, f_{2,x}, \dots, f_{\ell_x,x})^T$ se le conoce como la **distribución frecuencial marginal relativa** de \vec{x} . Análogamente para y se tiene la **distribución frecuencial marginal relativa** de \vec{y} dada por: $f_y = (f_{y,1}, f_{y,2}, \dots, f_{y,\ell_y})^T$. Las entradas de dichos vectores son de la forma $f_{i,x} = n_{i,x}/n$ y $f_{y,j} = n_{y,j}/n$.

En R podemos obtener las frecuencias mediante `prop.table`:

```
prop.table(table(datos$alcaldia_hechos, datos$ao_inicio))
```

```
##                                     2018      2019
## VERACRUZ  0.000000e+00 5.038291e-05
## VILLAGRAN 5.038291e-05 0.000000e+00
## XALATLACO 1.007658e-04 0.000000e+00
## XOCHIMILCO 2.342805e-02 5.794035e-03
## XOCHITEPEC 5.038291e-05 0.000000e+00
## ZACATECAS  0.000000e+00 1.007658e-04
```

Así mismo, podemos agregar las marginales:

```
addmargins(prop.table(table(datos$alcaldia_hechos, datos$ao_inicio)))

##                                     2018      2019      Sum
## VILLAGRAN 5.038291e-05 0.000000e+00 5.038291e-05
## XALATLACO 1.007658e-04 0.000000e+00 1.007658e-04
## XOCHIMILCO 2.342805e-02 5.794035e-03 2.922209e-02
## XOCHITEPEC 5.038291e-05 0.000000e+00 5.038291e-05
## ZACATECAS  0.000000e+00 1.007658e-04 1.007658e-04
## Sum        8.037082e-01 1.962918e-01 1.000000e+00
```

3. Riesgo Relativo (discreto) Para definir Riesgo Relativo empezaremos por un ejemplo. Tomamos la tabla (??) donde se guardó un registro de personas según si fumaban o no así como si dichas personas desarrollaron o no enfisema pulmonar.

```
## Warning: Setting row names on a tibble is deprecated.
```

Si quisieramos analizar la hipótesis de que FUMAR está asociado con ENFISEMA tendríamos que ver, dentro de la población de fumadores (FUMAR = SÍ) cuántos

	FUMA	NO FUMA
Con enfisema	100	40
Sin enfisema	30	50
Resultado (x)	Expuesto (y)	NO expuesto (y)
Resultado (x)	a	b
Sin resultado (x)	c	d

hay (proporcionalmente) con ENFISEMA. La hipótesis es que si no hubiera relación, saldría que las proporciones de fumadores con y sin enfisema serían 50% cada una. La proporción de fumadores con enfisema está dada por 100/130 mientras que la de no fumadores con enfisema es 40/90. El riesgo relativo (intuitivamente). se define como la división entre ambas proporciones:

$$\text{Riesgo Relativo de Enfisema} = \frac{\frac{\text{Expuestos enfermos}}{\text{Total de expuestos}}}{\frac{\text{No Expuestos enfermos}}{\text{Total de no expuestos}}} = \frac{100/130}{40/90} \approx 1.73$$

Lo que se interpreta como que los fumadores tienen 1.73 veces más riesgo de desarrollar enfisema que los no fumadores ya que si despejamos de la fórmula anterior:

$$\frac{\text{Expuestos enfermos}}{\text{Total de expuestos}} \approx 1.73 \times \frac{\text{No Expuestos enfermos}}{\text{Total de no expuestos}}$$

De manera general, dadas dos vectores lógicos \vec{x} (interpretada como el resultado) y \vec{y} (interpretada como la exposición) con una tabla de contingencia y frecuencias marginales dadas por la tabla:

Warning: Setting row names on a tibble is deprecated.

definimos el riesgo relativo de \vec{x} dado \vec{y} como:

$$RR(\vec{x}|\vec{y}) = \frac{\frac{a}{a+c}}{\frac{b}{b+d}}$$

Mientras que el riesgo relativo de no \vec{x} dado \vec{y} está dado por:

$$RR(\neg\vec{x}|\vec{y}) = \frac{\frac{c}{a+c}}{\frac{d}{b+d}}$$

La base de datos de los delitos no contiene información suficiente para poder calcular un riesgo relativo pero podemos crear la base de datos correspondiente a la tabla ?? como sigue:

```
fumadores <- data.frame(SI_FUMA = c(100, 30), NO_FUMA = c(40, 50))
```

Podemos agregar nombres a las filas para tener la base de datos mejor:

```
rownames(fumadores) <- c("ENFISEMA", "NO_ENFISEMA")
```

La tabla se ve así:

```
fumadores
```

	SI_FUMA	NO_FUMA
## ENFISEMA	100	40
## NO_ENFISEMA	30	50

Luego el riesgo relativo de ENFISEMA está dado por:

```
numerador <- fumadores["ENFISEMA", "SI_FUMA"]/sum(fumadores$SI_FUMA)
denominador <- fumadores["ENFISEMA", "NO_FUMA"]/sum(fumadores$NO_FUMA)
```

```
rr <- numerador/denominador #El riesgo relativo
rr
```

```
## [1] 1.730769
```

Por otro lado, el riesgo relativo de no enfisema es:

```
numerador <- fumadores["NO_ENFISEMA", "SI_FUMA"]/sum(fumadores$SI_FUMA)
denominador <- fumadores["NO_ENFISEMA", "NO_FUMA"]/sum(fumadores$NO_FUMA)
```

```
rr_neg <- numerador/denominador #El riesgo relativo
rr_neg
```

```
## [1] 0.4153846
```

Éste último se interpreta como si la proporción de individuos sin enfisema es 0.41 veces más pequeña entre fumadores que no fumadores.

4. Razón de momios (discreto) Para dos vectores lógicos \vec{x} y \vec{y} definimos la razón de momios como:

$$\text{OR}(\vec{x}|\vec{y}) = \frac{RR(\vec{x}|\vec{y})}{RR(\neg\vec{x}|\vec{y})}$$

Podemos calcular en R la razón de momios a partir de los datos:

```
razon.momios <- rr/rr_neg
```

donde la razón de momios de 4.17 se interpreta como “si un individuo tiene enfisema, la factibilidad de que dicho individuo sea fumador es 4.17 veces más

alta.” Esta interpretación se obtiene a partir de un despeje y sustitución:

$$RR(\vec{x}|\vec{y}) = 4.16 \cdot RR(-\vec{x}|\vec{y})$$

$$\Leftrightarrow \frac{\frac{\text{Expuestos enfermos}}{\text{Total de expuestos}}}{\frac{\text{No Expuestos enfermos}}{\text{Total de no expuestos}}} = 4.16 \cdot \frac{\frac{\text{Expuestos no enfermos}}{\text{Total de expuestos}}}{\frac{\text{No Expuestos no enfermos}}{\text{Total de no expuestos}}}$$

$$\Leftrightarrow \frac{\text{Expuestos enfermos}}{\text{No Expuestos enfermos}} = 4.16 \cdot \frac{\text{Expuestos no enfermos}}{\text{No Expuestos no enfermos}}$$

$$\Leftrightarrow \frac{\text{Expuestos enfermos}}{\text{Expuestos no enfermos}} = 4.16 \cdot \frac{\text{No Expuestos enfermos}}{\text{No Expuestos no enfermos}}$$

5. Correlación (Bravais-Pearson) Sean \vec{x} y \vec{y} dos vectores columna numéricos de nuestra matriz de datos Z . Tomemos $\tilde{x} = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})$ la versión centrada de \vec{x} y $\tilde{y} = (y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y})$ la versión centrada de \vec{y} . Al coseno entre dichos vectores (bajo el producto punto) se le conoce como correlación de Bravais-Pearson y se le denota $\rho_{\vec{x}, \vec{y}}$. Es decir:

$$\rho_{\vec{x}, \vec{y}} = \cos(\tilde{x}, \tilde{y}) = \frac{\tilde{x} \cdot \tilde{y}}{\|\tilde{x}\| \cdot \|\tilde{y}\|}$$

donde $\tilde{x} \cdot \tilde{y} = \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$ representa el producto de los vectores \tilde{x} y \tilde{y} y se conoce como **covarianza entre \vec{x} y \vec{y}** . Por otro lado,

$$\|\tilde{x}\| = \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} = \sigma_{\vec{x}}$$

Por tanto la correlación también puede medirse como:

$$\rho_{\vec{x}, \vec{y}} = \cos(\tilde{x}, \tilde{y}) = \frac{1}{\sigma_{\vec{y}} \sigma_{\vec{x}}} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Para matriz de datos Z con ℓ columnas, definimos la matriz de correlaciones \mathcal{C} como la matriz dada por:

$$\mathcal{C} = \begin{pmatrix} \rho(z_1, z_1) & \rho(z_1, z_2) & \dots & \rho(z_1, z_\ell) \\ \rho(z_2, z_1) & \rho(z_2, z_2) & \dots & \rho(z_2, z_\ell) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(z_\ell, z_1) & \rho(z_\ell, z_2) & \dots & \rho(z_\ell, z_\ell) \end{pmatrix}$$

Donde notamos (**demuestra**) que $\rho(z_i, z_i) = 1$.

Podemos usar la base `mtcars` precargada en R para analizar las correlaciones:

```
data(mtcars)
datos.coches <- mtcars
```

La base está explicada en la ayuda de R:

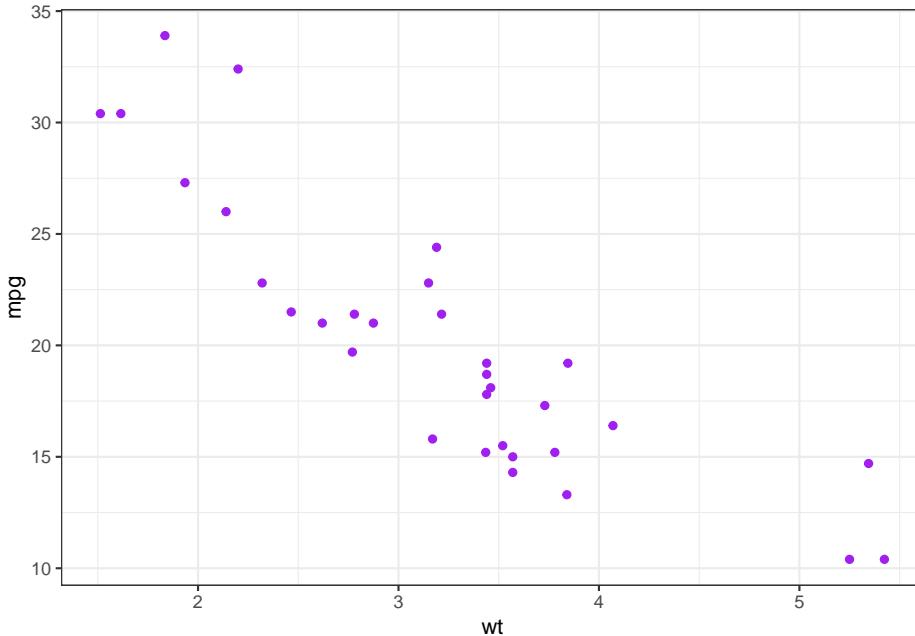
```
?mtcars
```

Podemos obtener la correlación entre el número de millas por galón `mpg` y el peso del automóvil `wt` haciendo:

```
cor(datos.coches$mpg, datos.coches$wt, method = "pearson")
## [1] -0.8676594
```

Esta correlación se interpreta como que por cada aumento en el peso corresponde una disminución en las millas por galón. Podemos ver gráficamente que esto es así:

```
ggplot(datos.coches) +
  geom_point(aes(x = wt, y = mpg), color = "purple") +
  theme_bw()
```



Para obtener toda la matriz de correlaciones de la base podemos tomar `cor` aplicado a toda la base de datos:

```
cor(datos.coches, method = "pearson")
```

```
##          mpg         cyl        disp         hp       drat        wt
## mpg  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
```

```

## cyl -0.8521620 1.0000000 0.9020329 0.8324475 -0.69993811 0.7824958
## disp -0.8475514 0.9020329 1.0000000 0.7909486 -0.71021393 0.8879799
## hp -0.7761684 0.8324475 0.7909486 1.0000000 -0.44875912 0.6587479
## drat 0.6811719 -0.6999381 -0.7102139 -0.4487591 1.00000000 -0.7124406
## wt -0.8676594 0.7824958 0.8879799 0.6587479 -0.71244065 1.0000000
## qsec 0.4186840 -0.5912421 -0.4336979 -0.7082234 0.09120476 -0.1747159
## vs 0.6640389 -0.8108118 -0.7104159 -0.7230967 0.44027846 -0.5549157
## am 0.5998324 -0.5226070 -0.5912270 -0.2432043 0.71271113 -0.6924953
## gear 0.4802848 -0.4926866 -0.5555692 -0.1257043 0.69961013 -0.5832870
## carb -0.5509251 0.5269883 0.3949769 0.7498125 -0.09078980 0.4276059
##           qsec          vs          am          gear          carb
## mpg   0.41868403 0.6640389 0.59983243 0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866 0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692 0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043 0.74981247
## drat  0.09120476 0.4402785 0.71271113 0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870 0.42760594
## qsec  1.00000000 0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544 1.0000000 0.16834512 0.2060233 -0.56960714
## am   -0.22986086 0.1683451 1.00000000 0.7940588 0.05753435
## gear -0.21268223 0.2060233 0.79405876 1.0000000 0.27407284
## carb -0.65624923 -0.5696071 0.05753435 0.2740728 1.00000000

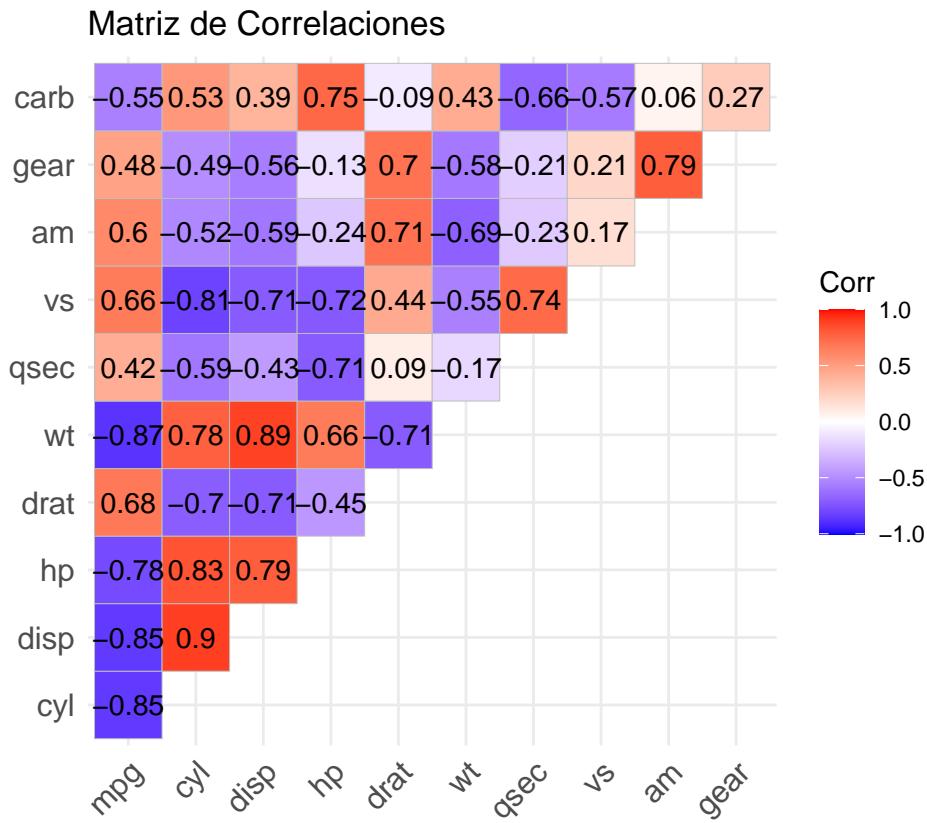
```

Finalmente, el paquete `ggcorrplot` puede ayudarnos a visualizar gráficamente dicha matriz:

```

ggcorrplot(cor(datos.coches, method = "pearson"),
           lab = TRUE,
           type = "upper") +
  labs(title = "Matriz de Correlaciones")

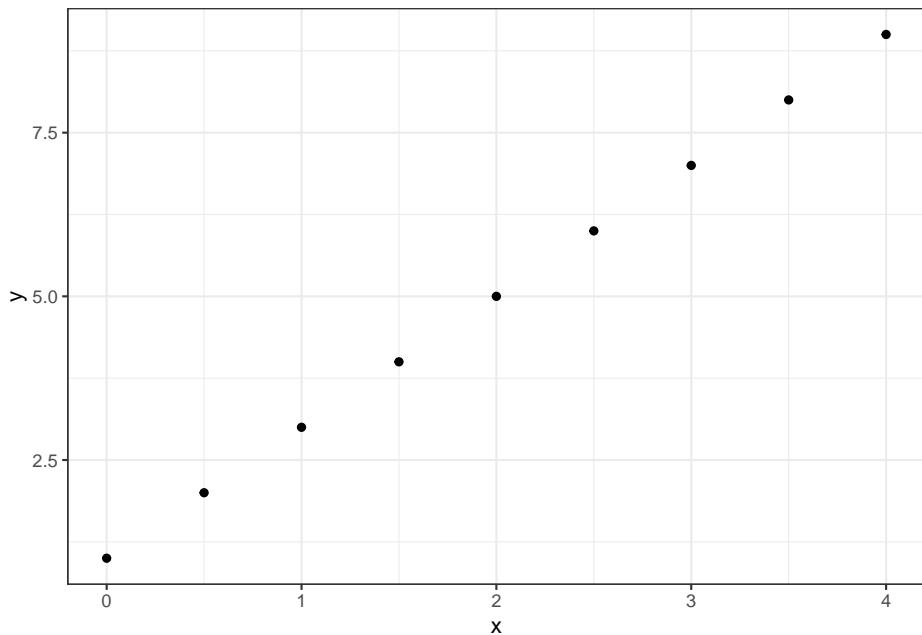
```



Una correlación de Pearson igual a 1 ó -1 se interpreta como que hay una relación lineal perfecta mientras que una correlación igual a 0 se interpreta como que no hay relación lineal (aunque puede existir de otro tipo)

```
#Ejemplo de correlación lineal perfecta
x <- seq(0,4, length.out = 9)
y <- 2*x + 1
```

Gráficamente:



El valor en este caso de la correlación es:

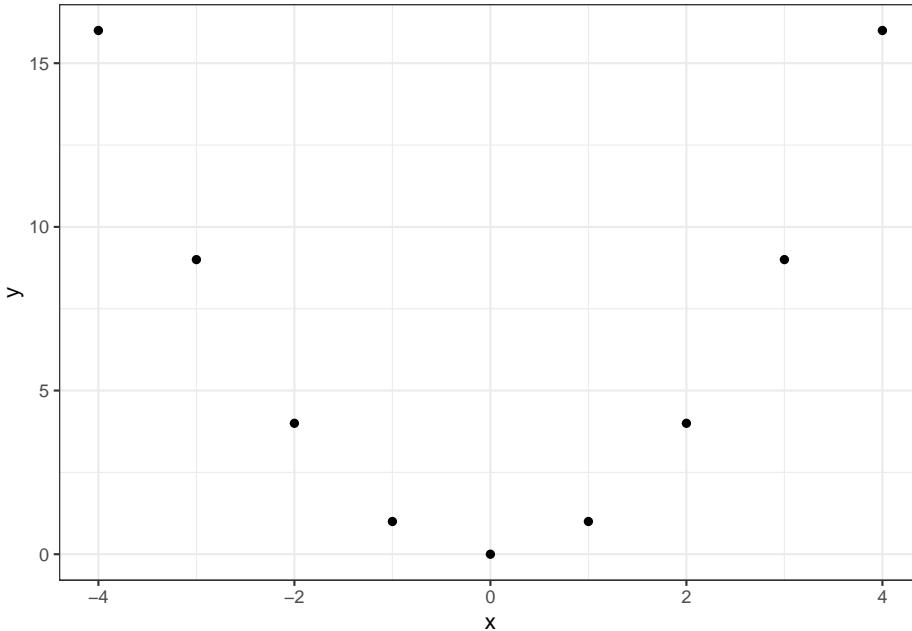
```
cor(x,y, method = "pearson")
```

```
## [1] 1
```

Mientras que por otro lado podemos tener variables relacionadas pero *sin correlación*:

```
#Ejemplo sin correlación lineal pero con variables relacionadas
x <- seq(-4, 4,length.out = 9)
y <- x^2
cor(x,y, method = "pearson")
```

```
## [1] 0
```



6. Correlación de rango de Spearman Para hablar de la correlación de rango de Spearman es necesario definir una variable como **ordinal**.

Un vector $\vec{x} = (x_1, x_2, \dots, x_n)^T$ de variables numéricas o categóricas es **ordinal** si existe una relación \leq de orden total sobre los elementos del vector tal que:

1. Es *antisimétrica*: si $x_i \leq x_j$ y $x_j \leq x_i$ entonces $x_i = x_j$.
2. Es *transitiva*: si $x_i \leq x_j$ y $x_j \leq x_k$ entonces $x_i \leq x_k$.
3. Es *conexa*: $x_i \leq x_j$ ó $x_j \leq x_i$.

De manera intuitiva un vector es **ordinal** si hay un orden para sus entradas. Por ejemplo, cuando calificas un servicio como **Malo \leq Regular \leq Bueno** o bien cuando se compara nivel educativo (en términos de años) **Primaria \leq Secundaria \leq Preparatoria \leq Educación superior**. **Toda variable numérica es ordinal**.

Para un vector ordinal definimos su ordenamiento como $x_{(1)} = \min\{x_1, x_2, \dots, x_n\}$ y $x_{(j)} = \min\{x_1, x_2, \dots, x_n\} \setminus \{x_{(1)}, x_{(2)}, \dots, x_{(j-1)}\}$ de tal forma que $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. El rango de $x_{(j)}$ denotado como $R(x_{(j)})$ es j (su posición en el ordenamiento). Es decir:

$$R(x_i) = j \Leftrightarrow x_i = x_{(j)}$$

Dado un vector \vec{x} definimos su **vector de rango** como:

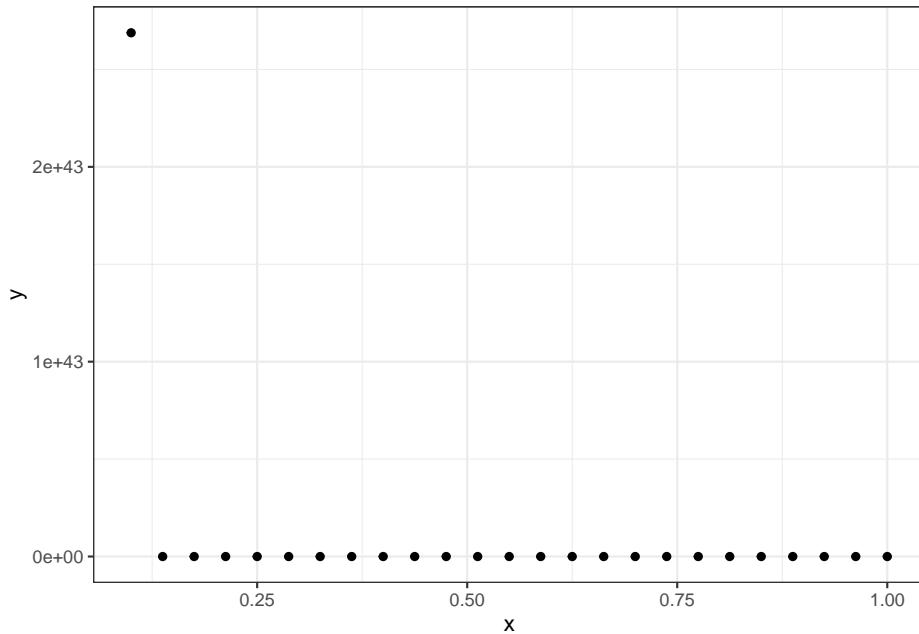
$$R(\vec{x}) = (R(x_1), R(x_2), \dots, R(x_n))^T$$

Para dos variables ordinales, \vec{x} y \vec{y} se define la **correlación de rango de Spearman** como la correlación de Pearson entre sus vectores de rangos:

$$\rho_{\text{Spearman}} = \rho(R(\vec{x}), R(\vec{y}))$$

Mientras que la correlación de Pearson mide linealidad; la de Spearman mide monotonicidad (que si una aumenta la otra también; que si una disminuye la otra también).

```
#Comparativo de correlaciones: la de Pearson no encuentra mucha linea
x <- seq(0.1, 1, length.out = 25)
y <- exp(1/x^2)
```



En este caso la correlación de Pearson es muy mala:

```
cor(x,y, method = "pearson")
```

```
## [1] -0.3396831
```

Mientras que la de Spearman sí muestra la relación:

```
cor(x,y, method = "spearman")
```

```
## [1] -1
```

7. τ de Kendall Consideremos \vec{x} y \vec{y} dos vectores columna ordinales de una matriz de datos Z . Para cualquier par de observaciones (x_i, y_i) y (x_j, y_j) con $i < j$ decimos que dos observaciones son **concordantes** (c) si los rangos de ambas x y y coinciden; es decir si se cumple una de las siguientes:

calidad_alimentos	calidad_servicio
Malo	1 estrella
Bueno	4 estrellas
Bueno	5 estrellas
Regular	2 estrellas
Bueno	5 estrellas
Bueno	4 estrellas

1. $R(x_i) < R(x_j)$ y $R(y_i) < R(y_j)$ o bien,
2. $R(x_i) > R(x_j)$ y $R(y_i) > R(y_j)$.

Observaciones **discordantes** (*d*) ocurren cuando los rangos de las x y las y son inversos el uno del otro; es decir, se cumple una de las siguientes:

1. $R(x_i) > R(x_j)$ y $R(y_i) < R(y_j)$ o bien,
2. $R(x_i) < R(x_j)$ y $R(y_i) > R(y_j)$.

En el caso que cualquiera de las dos, x ó y sean igualdades ($x_i = x_j$ ó $y_i = y_j$) no son discordantes ni concordantes.

Observa que existen $\binom{n}{2}$ distintos pares de (x_i, y_i) y (x_j, y_j) para comparar. Sea $c_{\vec{x}, \vec{y}}$ la cantidad de pares concordantes y $d_{\vec{x}, \vec{y}}$ la cantidad de pares discordantes. Luego la probabilidad de que dos pares seleccionados de manera uniforme sean concordantes es:

$$\frac{c_{\vec{x}, \vec{y}}}{\binom{n}{2}}$$

mientras que la probabilidad de que dos pares seleccionados uniformemente sean discordantes es:

$$\frac{d_{\vec{x}, \vec{y}}}{\binom{n}{2}}.$$

Definimos entonces la τ de Kendall como la diferencia entre ambas probabilidades empíricas:

$$\tau_{\vec{x}, \vec{y}} = \frac{c_{\vec{x}, \vec{y}} - d_{\vec{x}, \vec{y}}}{\binom{n}{2}}$$

La tau de Kendall cumple que:

$$-1 \leq \tau_{\vec{x}, \vec{y}} \leq 1$$

donde el -1 se alcanza sólo si son completamente discordantes (el rango de x es el inverso del rango de las y) y el 1 si son completamente concordantes (el rango de x y de y tienen el mismo orden). Una τ cercana a cero se interpreta como ausencia de relación en los rangos.

Podemos aplicar la tau de Kendall a la siguiente base de datos que contiene la calificación de dos servicios de un restaurante:

Para ello codificamos las variables como factor diciéndole que son variables ordinales `order = TRUE` e indicando el orden de los niveles:

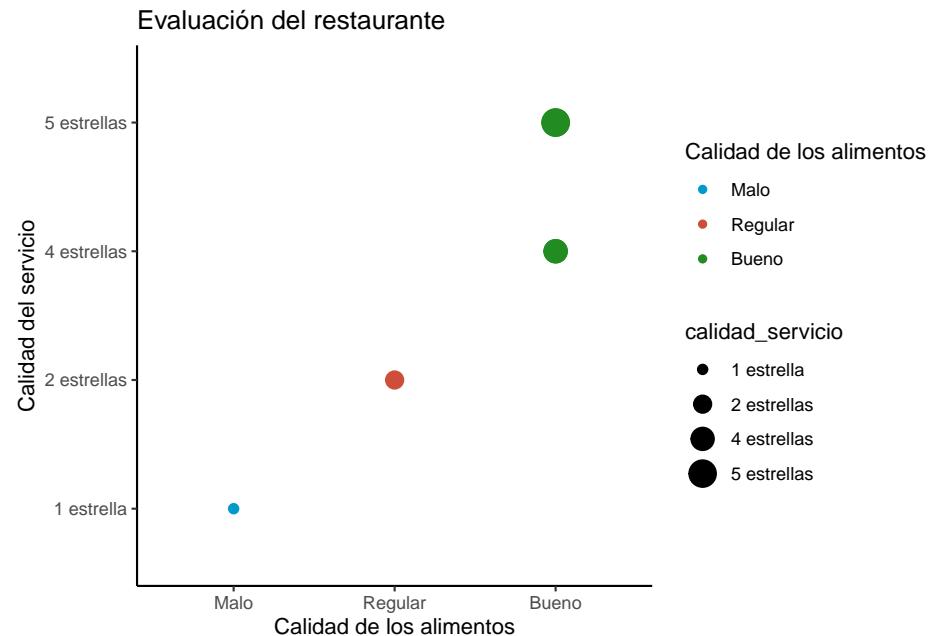
```
calidad_alimentos <- factor(c("Malo", "Bueno", "Bueno", "Regular", "Bueno", "Bueno"),
                             order = TRUE,
                             levels = c("Malo", "Regular", "Bueno"))
calidad_servicio <- factor(c("1 estrella", "4 estrellas", "5 estrellas",
                            "2 estrellas", "5 estrellas", "4 estrellas"),
                           order = TRUE,
                           levels = c("1 estrella", "2 estrellas", "3 estrellas",
                                     "4 estrellas", "5 estrellas"))
```

Esto de las variables ordinales permite hacer comparaciones ordinales, por ejemplo:

```
calidad_alimentos[2] > calidad_alimentos[4]
```

```
## [1] TRUE
```

Los datos se ven así:



Finalmente, calculamos la τ de Kendall, para ello es necesario obtener el rango de nuestras variables ordinales:

```
rango_alimentos <- as.numeric(calidad_alimentos)
rango_servicio <- as.numeric(calidad_servicio)
cor(rango_alimentos, rango_servicio, method = "kendall")
```

```
## [1] 0.8320503
```

Lo cual indica que hay una relación entre la calificación de calidad de alimentos y la del servicio.

8. Ajuste de modelo lineal

Sean \vec{x} y \vec{y} dos vectores columna de una matriz de datos Z . Supongamos, además, se tiene la hipótesis de que existe una relación afín entre los vectores; es decir que:

$$\vec{y} \approx \beta_1 \vec{x} + \beta_0 \vec{1}$$

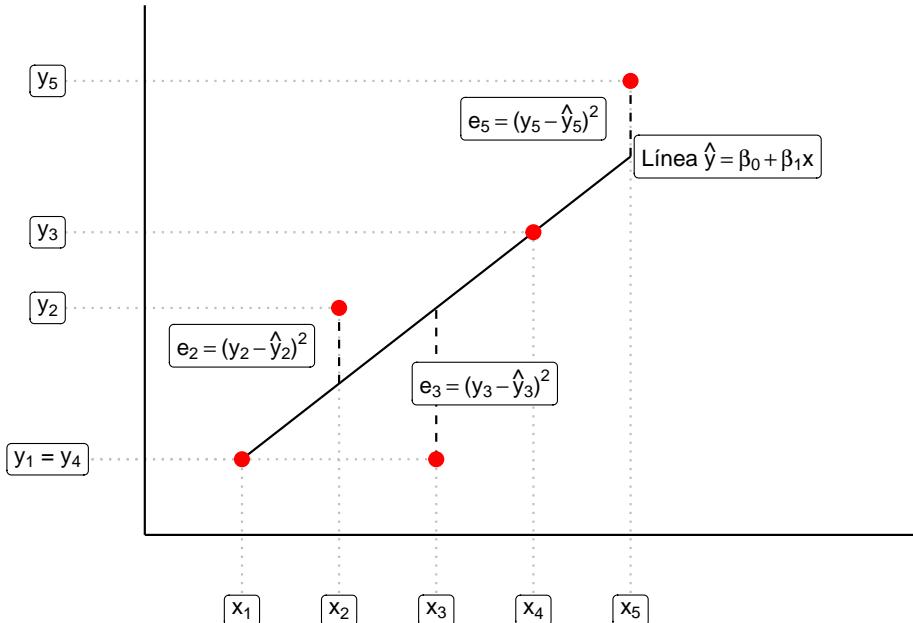
donde $\vec{1} = (1, 1, \dots, 1)^T$ es un vector con todas las entradas idénticas a 1 y $\beta_0, \beta_1 \in \mathbb{R}$. Algunas razones para tener esta hipótesis podría ser una correlación de Pearson cercana a ± 1 o por inspección gráfica. Esta hipótesis implica que:

$$y \approx \underbrace{\beta_1 x + \beta_0}_{\hat{y}}$$

Podemos entonces trazar la línea $y = \beta_0 + \beta_1 x$ y graficar contra los puntos $\{(x_i, y_i)\}_{i=1}^n$. Si la línea no ajusta perfecto tendremos errores $e_i = (y_i - \hat{y}_i)^2$ de predicción las cuales representan la diferencia entre la y observada (y_i) y la y predicha por la línea $\hat{y}_i = \beta_1 x_i + \beta_0$. La suma de estos errores es:

$$\text{SSR}(\beta_0, \beta_1) = \sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (\beta_1 x_i + \beta_0))^2$$

El nombre de *SSR* es por (*Sum of Squared Residuals*) dado que en estadística se define un residual como $r_i = (y_i - \hat{y}_i)$. Gráficamente:



Lo que se busca entonces es minimizar el error respecto a las constantes a determinar: β_0 y β_1 . Para ello buscamos un punto de inflexión derivando:

$$\frac{\partial \text{SSR}}{\partial \beta_0} = \sum_{i=1}^n 2(y_i - (\beta_1 x_i + \beta_0)) = 0$$

De donde se sigue que:

$$\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i - n\beta_0 = 0 \Rightarrow \beta_0 = \frac{1}{n} \sum_{i=1}^n y_i - \beta_1 \frac{1}{n} \sum_{i=1}^n x_i \Rightarrow \bar{y} - \beta_1 \bar{x},$$

de donde concluimos que de cumplirse la relación lineal se tiene que:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}.$$

Por otro lado, la derivada respecto a β_1 es:

$$\frac{\partial \text{SSR}}{\partial \beta_1} = - \sum_{i=1}^n 2(y_i - (\beta_1 x_i + \beta_0)) \cdot x_i = 0$$

De donde se sigue (si suponemos que existe al menos un $x_i \neq 0$):

$$\begin{aligned} 0 &= - \sum_{i=1}^n \left(x_i y_i - \beta_1 x_i^2 - \underbrace{\beta_0}_{\bar{y} - \beta_1 \bar{x}} x_i \right) \\ &= \sum_{i=1}^n \left(x_i y_i - \beta_1 x_i^2 - \bar{y} x_i + \beta_1 \bar{x} x_i \right) \\ &= \sum_{i=1}^n \left(y_i + \beta_1 x_i - \bar{y} - \beta_1 \bar{x} \right) x_i \\ &= \sum_{i=1}^n \left(y_i - \bar{y} \right) x_i - \beta_1 \sum_{i=1}^n \left(x_i - \bar{x} \right) x_i \end{aligned}$$

de donde se sigue (suponiendo que existen i, j tales que $x_i \neq x_j$ que:

$$\beta_1 = \frac{\sum_{i=1}^n (y_i - \bar{y}) x_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma_x \cdot \sigma_y \cdot \rho_{x,y}}{n \sigma_x^2}$$

por lo cual:

$$\beta_1 = \frac{\sigma_y}{\sigma_x} \cdot \frac{\rho_{x,y}}{n}$$

De donde se tienen las fórmulas para el β_0 y β_1 .

3.10.1 Ejercicio

Demuestra la igualdad que usamos anteriormente:

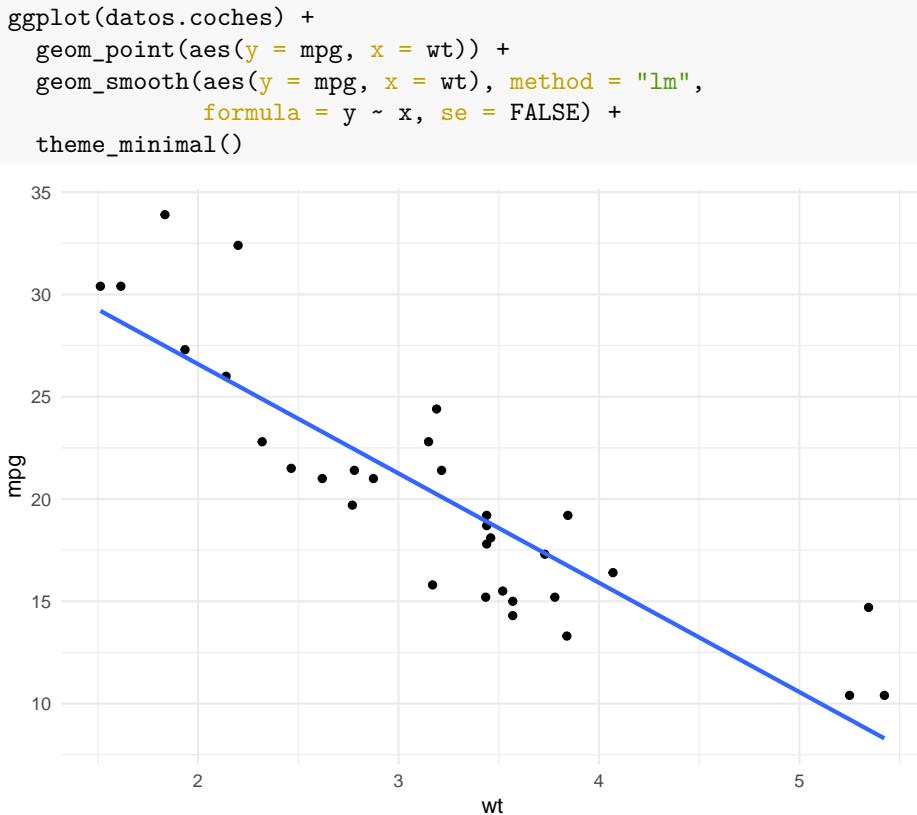
$$\frac{\sum_{i=1}^n (y_i - \bar{y})x_i}{\sum_{i=1}^n (x_i - \bar{x})x_i} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

En R podemos ajustar un modelo lineal para dos variables de una base de datos con `lm`:

```
modelo.lineal <- lm(mpg ~ wt, data = datos.coches)
coef(modelo.lineal)

## (Intercept)          wt
##  37.285126   -5.344472
```

Gráficamente podemos ver el modelo:



Para predecir, dada una nueva observación, cuál debe haber sido el valor de \hat{y} para una nueva observación x_* (o varias nuevas observaciones) puede usarse la

```
función predict
datos_a_predecir <- data.frame(wt = c(5.5, 6, 6.5))
predict(modelo.lineal, datos_a_predecir)

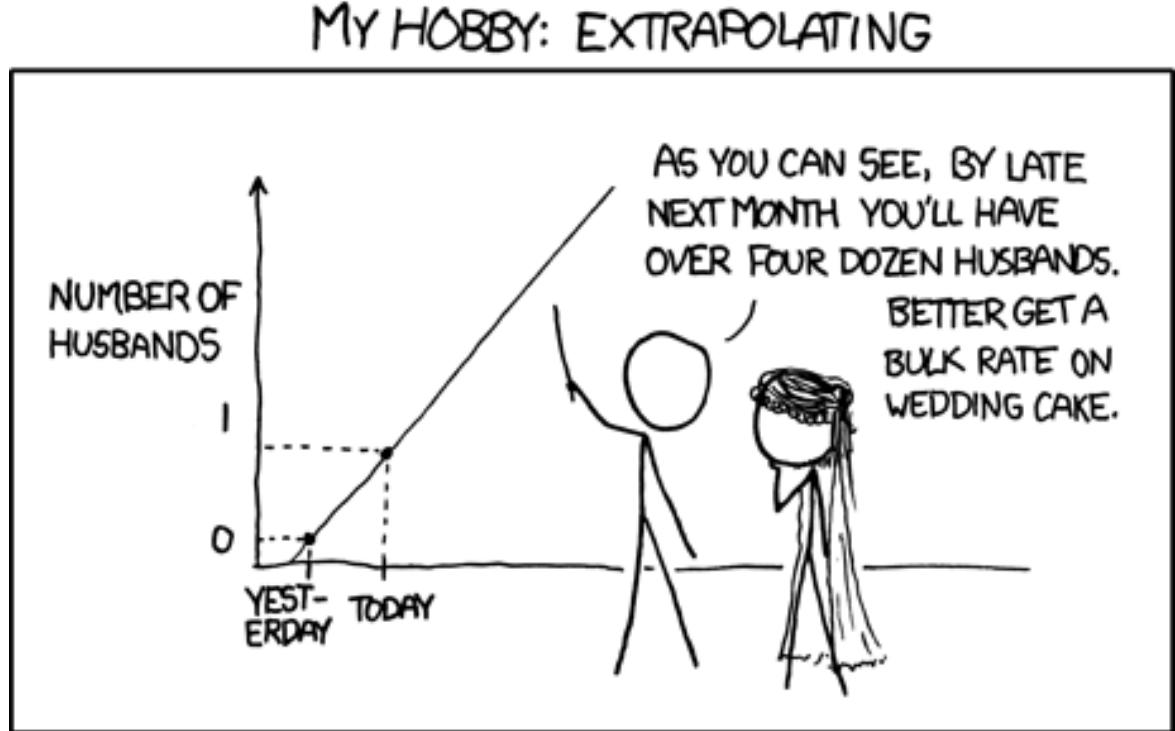
##           1          2          3
## 7.890533 5.218297 2.546061
```

Hay que tener mucho cuidado con la generalización de un modelo lineal como los siguientes valores muestran:

```
datos_a_predecir <- data.frame(wt = c(7,8,9))
predict(modelo.lineal, datos_a_predecir)

##           1          2          3
## -0.1261748 -5.4706464 -10.8151180
```

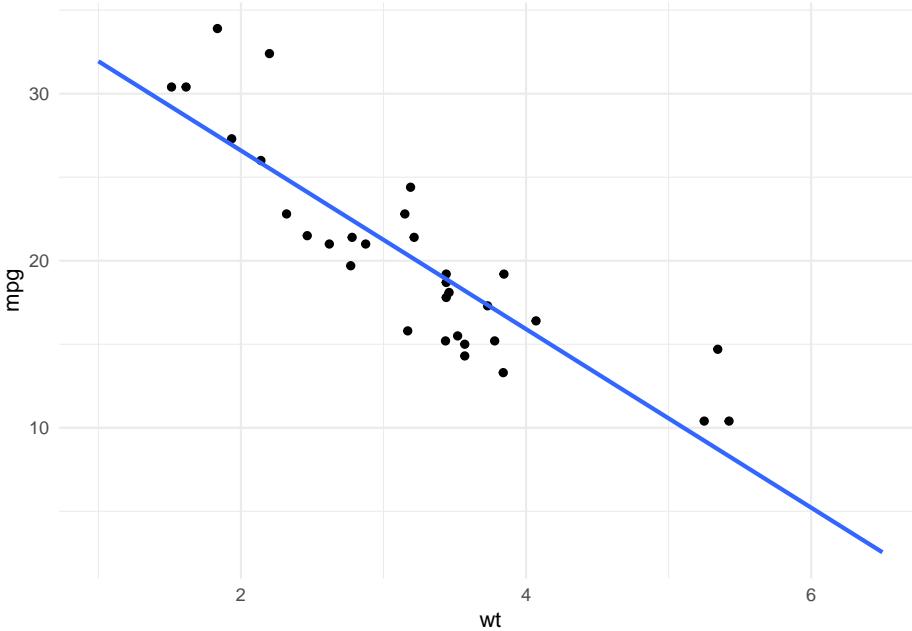
O bien el siguiente comic de xkcd:



Para hacer la extrapolación gráfica podemos agregar un `fullrange = TRUE` combinado con un `xlim`

```
ggplot(datos.coches) +
  geom_point(aes(y = mpg, x = wt)) +
  geom_smooth(aes(y = mpg, x = wt), method = "lm",
```

```
formula = y ~ x, se = FALSE, fullrange=TRUE) +
theme_minimal() +
xlim(c(1,6.5))
```



3.11 Ejercicio

- Generaliza el proceso de estimación para cuando se tiene un polinomio $y = \beta_0 + \beta_1 x + \beta_2 x^2$
- Utiliza los datos confirmados de COVID-19 a nivel nacional (sólo los confirmados) disponibles en este link. Ajusta un modelo cuadrático (en el `lm` la fórmula ahora es del estilo de `y ~ poly(x, 2)`) y predice cuántos casos confirmados habrá el 29 de junio. Grafica tu ajuste así como tu predicción en la misma gráfica.
- Ajuste general** Podemos generalizar el ajuste de mínimos cuadrados planteando el modelo $y = f(x, \vec{\beta})$ donde x puede ser una matriz y $\vec{\beta}$ es un vector de parámetros. Supondremos que f es diferenciable en $\vec{\beta}$.

Como ejemplo, en el caso del ajuste lineal:

$$y = f(x, \vec{\beta}) = \beta_0 + \beta_1 x \quad \text{con} \quad \vec{\beta} = (\beta_0, \beta_1)^T.$$

o bien podríamos pensar en un ajuste polinomial:

$$y = f(x, \vec{\beta}) = \sum_{i=0}^n \beta_i x^i \quad \text{con} \quad \vec{\beta} = (\beta_0, \beta_1, \dots, \beta_n)^T.$$

No tiene que ser un polinomio, f puede ser lo que ella quiera ser siempre y cuando sea diferenciable en los parámetros:

$$y = f(x, \vec{\beta}) = \left[\cos(\beta_0 + x) + \int_0^{\beta_1 x} e^{-t^2} dt \right] \cdot \beta_2 \ln(x) \quad \text{con} \quad \vec{\beta} = (\beta_0, \beta_1, \beta_2)^T.$$

3.12 Ajuste funcional

Hacemos una apuesta por teléfono. Yo voy a tirar una moneda 10 veces y si salen más **Soles** que **Águilas** yo gano 50 pesos. Si salen más **Águilas** que **Soles** tú ganas la misma cantidad. Al realizar el ejercicio yo te comunico que salieron en total 10 **Soles** y por tanto me debes el dinero. *¿Sospecharías algo de mí?*

Si no hablamos de probabilidad no hay forma en la que se pueda justificar que *aparentemente* hay algo raro con la moneda. Claro, siempre puede ser un caso improbable (hay gente que lo ha hecho) pero es *raro* que me hayan salido tantos **Soles**. Para cuantificar qué tan raro es el evento podemos suponer que las monedas siguen un modelo Binomial con parámetro $p = 1/2$ y en este caso $n = 10$ (fueron 10 tiros). La probabilidad de que haya obtenido 10 soles bajo este modelo es de:

```
dbinom(10, 10, 1/2)
```

```
## [1] 0.0009765625
```

¡Rarísimo! Este resultado te haría sospechar que quizás mi moneda no es *justa* y no se obtienen la misma cantidad de **Águilas** que **Soles** cuando la tiro. Esto porque, aparentemente, en mi moneda la probabilidad de **Sol** debería de ser $p = 1$ (por tu triste experiencia). Si por ejemplo en el onceavo tiro saliera un **Águila**, concluirías que, en mi moneda, aparentemente, la probabilidad de **Sol** es $p = \frac{10}{11}$. Por supuesto, entre más tiros y más información obtienes, mejor podrás caracterizar la moneda y con mayor sustento tendrás sospechas (o no) de que mi moneda es trampa.

Formalmente, en el ejemplo anterior, lo que se hace es suponer que existe una variable aleatoria $X \in \{\text{Águila}, \text{Sol}\}$ (el resultado de la moneda) de la cual observamos $n = 11$ realizaciones codificadas en el siguiente vector:

$$\vec{x} = (\text{Sol}, \text{Sol}, \text{Sol}, \text{Sol}, \text{Sol}, \text{Sol}, \text{Sol}, \text{Sol}, \text{Sol}, \text{Sol}, \text{Águila})^T$$

Aproximamos entonces la probabilidad $\mathbb{P}(X = \text{Sol})$ mediante:

$$\mathbb{P}(X = \text{Sol}) \approx \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{\text{Sol}\}}(x_i) = \frac{10}{11}$$

Mientras que la de **Águila** se aproxima mediante:

$$\mathbb{P}(X = \text{Águila}) \approx \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{\text{Águila}\}}(x_i) = \frac{1}{11}$$

Para ver que éstas son buenas aproximaciones, podemos considerar un vector aleatorio de los *posibles* datos observados:

$$\vec{X} = (X_1, X_2, \dots, X_{11})^T$$

Donde X_1 es una variable aleatoria que representa lo que *pudo* haber salido en el primer tiro, X_2 es una v.a. que representa lo que *pudo* haber salido en el segundo tiro y en general X_k es una v.a. que representa lo que *pudo* haber salido en el k -ésimo tiro.

Suponiendo que la moneda tiene una probabilidad p de arrojar **Sol** y $1 - p$ de arrojar **Águila**, notamos que las variables indicadoras evaluadas en las X_i (aleatorias) son variables aleatorias

$$\mathbb{I}_{\{\text{Sol}\}}(X_i) \sim \text{Beroulli}(p)$$

y que por tanto

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{\text{Sol}\}}(X_i)$$

es una variable aleatoria (al ser suma de variables aleatorias). Podemos entonces calcular su valor esperado:

$$\mathbb{E}[\hat{p}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{\text{Sol}\}}(X_i)\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\mathbb{I}_{\{\text{Sol}\}}(X_i)] = \frac{1}{n} \sum_{i=1}^n p = \frac{1}{n} \cdot np = p$$

Es decir, que en promedio el estimador \hat{p} va a atinarle al verdadero valor p . Esto lo podemos ver si hacemos `nsim = 1000` simulaciones de 100 `tiros` de una moneda con probabilidad `p = 8/10` de sol.

```

nsim <- 1000
tiros <- 100
p.val <- 8/10

#Creamos un vector para guardar los valores de p gorro
p.gorro <- rep(NA, nsim)

#Loop recorriendo cada una de las nsim simulaciones
for (i in 1:nsim){
  experimento <- sample(c("Sol", "Águila"), tiros, replace = TRUE,
                        prob = c(p.val, 1 - p.val))
  soles      <- table(experimento)["Sol"]
  p.gorro[i] <- soles/tiros
}

```

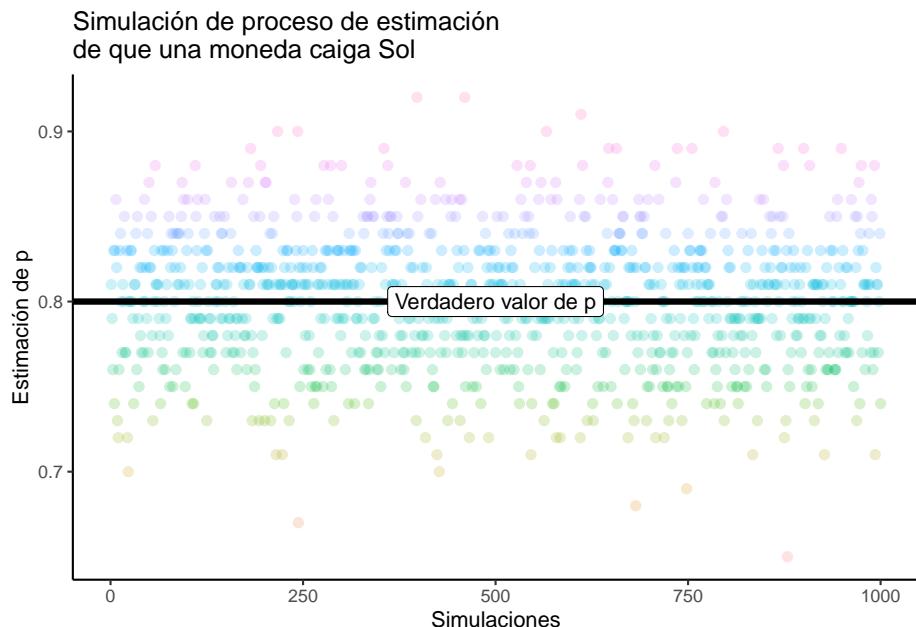
Podemos ver que en promedio le atinamos al valor verdadero:

```
#Vemos que en promedio le atina:  
mean(p.gorro)
```

```
## [1] 0.80131
```

Lo mismo podemos verlo gráficamente:

```
#Graficamos  
ggplot() +  
  geom_point(aes(x = 1:n sim, y = p.gorro, color = as.character(p.gorro)),  
             size = 2, alpha = 0.2) +  
  geom_hline(aes(yintercept = p.val), size = 1.5, linetype = "solid") +  
  theme_classic() +  
  theme(legend.position = "none") +  
  labs(  
    x = "Simulaciones",  
    y = "Estimación de p",  
    title = "Simulación de proceso de estimaciónnde que una moneda caiga Sol"  
) +  
  geom_label(aes(x = nsim/2, y = p.val), label = "Verdadero valor de p")
```



¿Qué significa esto? El que en promedio \hat{p} sea p (formalmente, que $E[\hat{p}] = p$) significa que, si yo hago muchísimos experimentos (o procesos de muestreo) de la misma cosa, mi \hat{p} es un buen estimador porque en promedio le va a atinar. Empero, esto no dice nada de qué tan bueno es mi estimador \hat{p} para mi caso (mi muestra o mi experimento) específico. Puedes pensarlo con los exámenes:

que alguien tenga un promedio de 8 dice que en general le ha ido bien en los exámenes, pero no dice nada respecto al primer examen de cálculo que hizo (donde pudo tener 10 ó 5 para llegar a ese promedio de 8 *pero no podemos saber de manera específica cuánto fue*). Esto es igual: en promedio el estimador \hat{p} será p pero para un análisis específico *no sabemos*.

OJO Los datos observados no son variables aleatorias: esos ya son fijos, ya los viste. Los *posibles* datos observados sí son variables aleatorias ya que ellos, consisten en las variables que se *pudieron* haber observado y te permiten calcular las probabilidades de tus datos observados bajo algún modelo. En el caso de la moneda, los datos observados son $\vec{x} = (\text{Sol}, \text{Sol}, \text{Sol}, \text{Sol}, \text{Sol}, \text{Sol}, \text{Sol}, \text{Sol}, \text{Sol}, \text{Águila})^T$ pero los que se *pudieron* haber observado son todas las $\binom{n}{2}$ formas en las que la moneda pudo haber salido.

1. Estimación de una función de masa de probabilidad Formalmente, para una variable aleatoria discreta X que puede tomar los valores $\{a_1, a_2, \dots, a_\ell\}$ de la cual se observaron n realizaciones descritas mediante $\vec{x} = (x_1, x_2, \dots, x_n)^T$ (observados, fijos, constantes). Definimos la **función de masa de probabilidad empírica** como:

$$\hat{p}(x) = \begin{cases} \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{a_1\}}(x_i) & \text{si } x = a_1 \\ \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{a_2\}}(x_i) & \text{si } x = a_2 \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{a_\ell\}}(x_i) & \text{si } x = a_\ell \\ 0 & \text{en otro caso} \end{cases}$$

donde se supone que $\mathbb{P}(X = x) \approx \hat{p}(x)$. Notamos que lo anterior puede resumirse en:

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{x\}}(x_i)$$

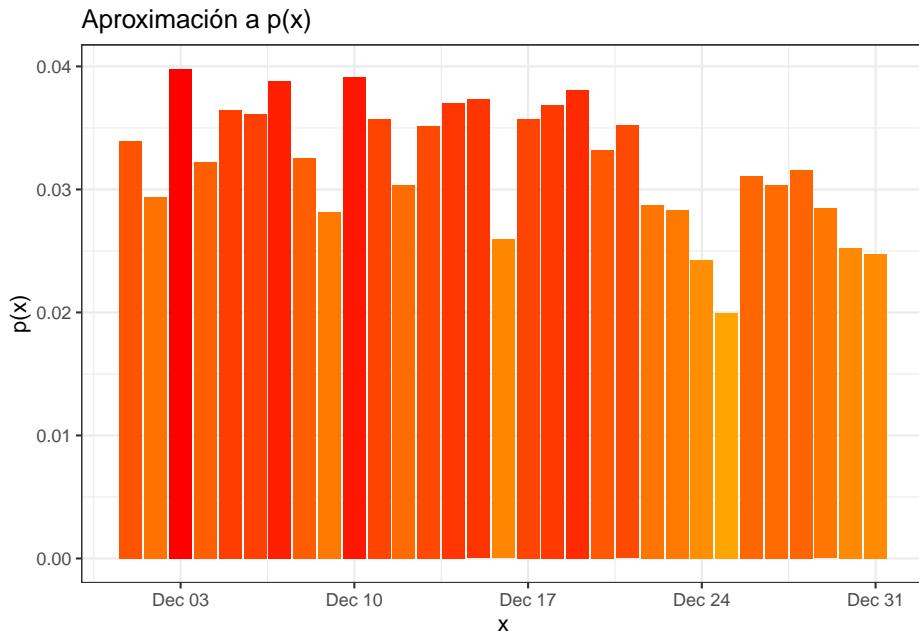
Análogamente, nota que para un conjunto (medible) A , la aproximación para $\mathbb{P}(X \in A)$ está dada por:

$$\hat{p}(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_A(x_i).$$

Podemos graficar para la base de datos `conteo_delitos` la probabilidad de que,

dado que se cometió un delito, éste haya ocurrido en el dia d_i de diciembre. Para ello usamos un `geom_col`:

```
ggplot(conteo_delitos) +
  geom_col(aes(x = fecha, y = n/sum(n), fill = n)) +
  scale_fill_gradient("Delito", low = "orange", high = "red") +
  theme_bw() +
  theme(legend.position = "none") +
  labs(
    y = "p(x)",
    x = "x",
    title = "Aproximación a p(x)"
  )
```



Una propiedad interesante de la función de masa de probabilidad es que, en promedio, le atina al verdadero valor (lo que comentábamos antes de que $\hat{p} = p$). Es decir, suponiendo que X tiene una función de masa dada por:

$$p(x) = \begin{cases} p_1 & \text{si } x = a_1 \\ p_2 & \text{si } x = a_2 \\ \vdots & \\ p_\ell & \text{si } x = a_\ell \end{cases}$$

y suponiendo un vector de muestras posibles $\vec{X} = (X_1, X_2, \dots, X_n)^T$ notamos

que

$$\mathbb{I}_{\{a_j\}}(X_i) \sim \text{Bernoulli}(p_j)$$

Luego para cualquier x se tiene que:

$$\mathbb{E}[\hat{p}(x)] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{x\}}(X_i)\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\mathbb{I}_{\{x\}}(X_i)] = \frac{1}{n} n \cdot p_j = p_j.$$

2. Función de distribución empírica

Recuerda que para cualquier variable aleatoria $X : \mathbb{R} \rightarrow \mathbb{R}$ existe su función de distribución F_X dada por:

$$F_X(x) = \mathbb{P}(X \leq x)$$

La idea de la función de distribución empírica es reconstruir (a partir de los datos observados) a F_X . Para ello, notamos que queremos estimar

$$\mathbb{P}(X \leq x) \quad \forall x \in \mathbb{R}$$

esto es equivalente a estimar:

$$\mathbb{P}(X \in (-\infty, x])$$

y podemos aplicar la aproximación que usamos arriba para un conjunto A :

$$\mathbb{P}(X \in (-\infty, x]) \approx \sum_{i=1}^n \mathbb{I}_{(-\infty, x]}(x_i)$$

La función de distribución empírica está definida para un vector numérico $\vec{x} = (x_1, x_2, \dots, x_n)^T$ por:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(-\infty, x]}(x_i)$$

La función de distribución empírica es una función de distribución pues cumple las siguientes propiedades (demuéstralos):

1. $\lim_{x \rightarrow -\infty} \hat{F}(x) = 0$
2. $\lim_{x \rightarrow \infty} \hat{F}(x) = 1$
3. Si $x < y$ entonces $\hat{F}(x) \leq \hat{F}(y)$ (no decreciente)
4. \hat{F} es continua por la derecha con límites por la izquierda (càdlàg).

Para demostrar 4. basta con demostrar que para x_i fija, la función $i(x) = \mathbb{I}_{(-\infty, x]}(x_i)$ es continua por la derecha con límites por la izquierda pues $\hat{F}(x)$ es una suma de dichas funciones.

En particular, podemos notar que la función de distribución empírica $\hat{F}(x)$ le atina a la función de distribución; es decir:

$$\mathbb{E}[\hat{F}(x)] = F(x)$$

Para ello consideramos un vector de valores posibles $\vec{X} = (X_1, X_2, \dots, X_n)^T$ donde las X_i tienen la misma distribución que X . Y notamos que:

$$\mathbb{I}_{(-\infty, x]}(X_i) \sim \text{Bernoulli}(F(x))$$

pues $\mathbb{I}_{(-\infty, x]}(X_i) = 1$ si $X_i \leq x$ y $\mathbb{I}_{(-\infty, x]}(X_i) = 0$ si $X_i > x$. Luego:

$$\mathbb{P}(\mathbb{I}_{(-\infty, x]}(X_i) = 1) = \mathbb{P}(X_i \leq x) = \mathbb{P}(X \leq x) = F(x)$$

donde la igualdad del medio se sigue de que X_i y X tienen la misma distribución. Entonces:

$$\mathbb{E}[\hat{F}(x)] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(-\infty, x]}(X_i)\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\mathbb{I}_{(-\infty, x]}(X_i)] = \frac{1}{n} n \cdot F(x) = F(x)$$

En R podemos calcular la función de distribución empírica con el comando `ecdf` el cual cuenta la cantidad de observaciones y regresa una función. Así, para la base de datos `conteo_delitos` podemos calcular la función de distribución empírica `ecdf` asociada a la cantidad de delitos que se cometan en un día mediante:

```
Fgorro <- ecdf(conteo_delitos$n)
```

De esta forma podemos calcular la probabilidad de que en un día se cometan menos de 500 delitos:

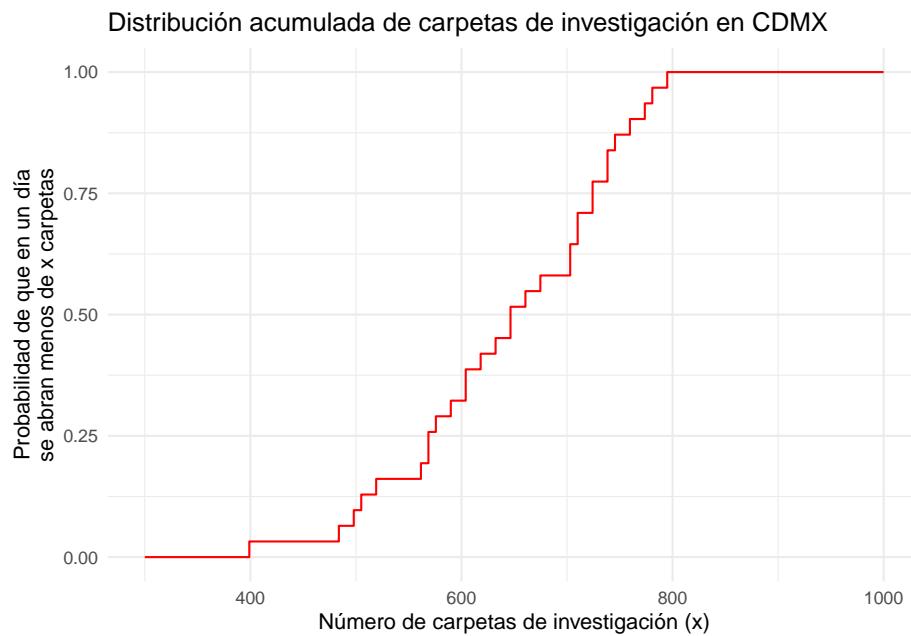
```
Fgorro(500)
```

```
## [1] 0.09677419
```

O bien podemos graficar la función:

```
x <- seq(300, 1000, length.out = 100)
y <- Fgorro(x)
ggplot() +
  geom_step(aes(x = x, y = y), color = "red") +
  labs(
    x = "Número de carpetas de investigación (x)",
    y = "Probabilidad de que en un día abran menos de x carpetas",
    title = "Distribución acumulada de carpetas de investigación en CDMX"
```

```
) +
theme_minimal()
```



Mediante simulaciones, podemos observar que \hat{F} realmente le atina a F como sigue:

```
#Cantidad de simulaciones
nsim      <- 100

#Tamaño de la muestra en cada simulacion
n_muestra <- 100

#Valores a evaluar la función
x          <- seq(-5, 5, length.out = 200)

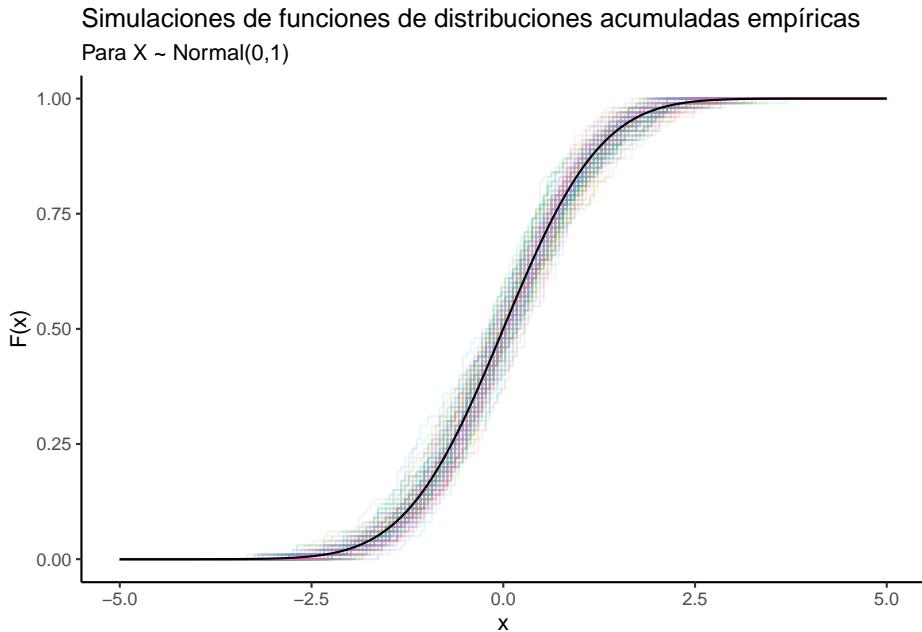
#Base de datos para guardar resultados de simulaciones
F_simulado <- data.frame(matrix(NA, ncol = nsim, nrow = length(x)))

for (i in 1:nsim){
  valores_simulados <- rnorm(n_muestra)
  F_empirica        <- ecdf(valores_simulados)
  F_simulado[,i]    <- F_empirica(x)
}

F_simulado$Valor_x <- x
```

```
#Cambiamos el formato de la base para graficar
F_simulado <- F_simulado %>% pivot_longer(cols = -Valor_x)

ggplot(F_simulado) +
  geom_step(aes(x = Valor_x, y = value, color = name), alpha = 0.1) +
  geom_line(aes(x = Valor_x, y = pnorm(Valor_x)), color = "black") +
  theme_classic() +
  theme(legend.position = "none") +
  labs(
    x = "x",
    y = "F(x)",
    title = "Simulaciones de funciones de distribuciones acumuladas empíricas",
    subtitle = "Para X ~ Normal(0,1)"
)
```



2. Histograma Para una variable aleatoria continua, la aproximación \hat{p} que hicimos no funciona (la masa siempre es 0). Por lo que es necesario analizar alternativas para estudiar la densidad si suponemos que los datos pueden modelarse mediante algo continuo. Para construir un histograma consideraremos $\vec{x} = (x_1, x_2, \dots, x_n)^T$ y una constante $h > 0$ llamada el **ancho de banda (bin-width)**. Sea $\{I_j\}$ una colección de intervalos no vacíos de \mathbb{R} tal que $\cup_{j=1}^n I_j = \mathbb{R}$ e $I_j \cap I_k = \emptyset$ (*i.e.* los $\{I_j\}$ forman una partición de \mathbb{R}). Supongamos, además, que los I_j son de la forma:

$$I_j = [\kappa + (j-1)h, \kappa + jh)$$

para algún $\kappa \in \mathbb{R}$ fijo. Sea

$$n_j(\vec{x}) = \sum_{i=1}^n \mathbb{I}_{I_j}(x_i)$$

la cantidad de x_i en el intervalo I_j .

Un histograma es la gráfica de la función (ver Panaretos (2016)):

$$\text{hist}_{\vec{x}}(x) = \frac{1}{n \cdot h} \sum_j n_j(\vec{x}) \cdot \mathbb{I}_{I_j}(x)$$

Una propiedad interesante de un histograma es que éste aproxima correctamente las probabilidades $\mathbb{P}(X \in I_j)$. Para ver esto, consideramos un vector de valores posibles $\vec{X} = (X_1, X_2, \dots, X_n)^T$ y que $x \in I_j$, luego:

$$\mathbb{E}\left[\int_{I_j} \text{hist}_{\vec{X}}(x) dx\right] = \mathbb{E}\left[\frac{1}{n \cdot h} \sum_j n_j(\vec{X}) \cdot \int_{I_j} \mathbb{I}_{I_j}(x) dx\right] = \frac{1}{n \cdot h} \sum_j \mathbb{E}[n_j(\vec{X})] \cdot h = \frac{1}{n} \sum_j \mathbb{E}[n_j(\vec{X})]$$

donde las $n_j(\vec{X})$ son variables aleatorias en este caso y:

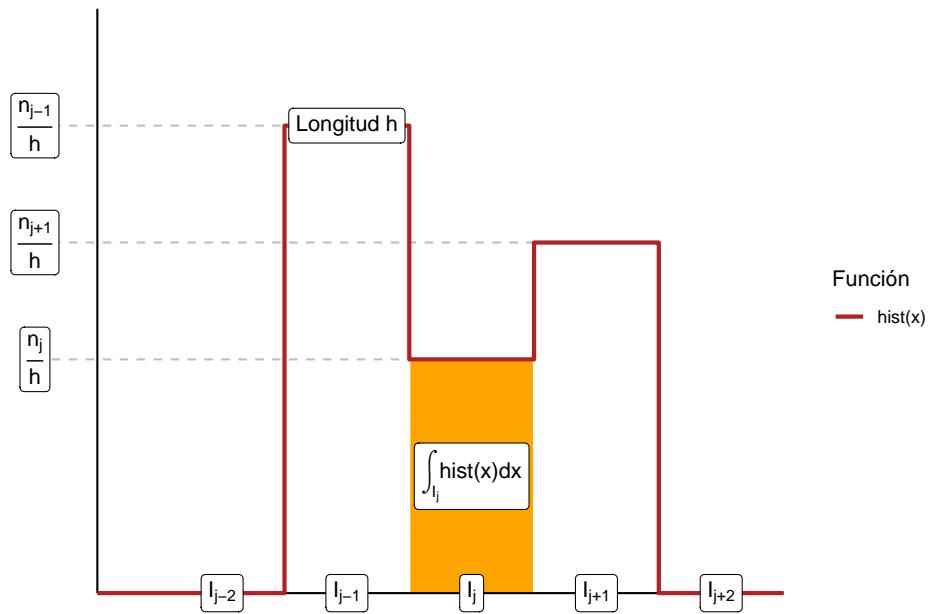
$$\mathbb{E}[n_j(\vec{X})] = \sum_{i=1}^n \mathbb{E}[\mathbb{I}_{I_j}(X_i)] = \sum_{i=1}^n \mathbb{P}(X_i \in I_j) = n \mathbb{P}(X \in I_j)$$

donde la última igualdad se da pues las X_i tienen la misma distribución que X . Luego:

$$\mathbb{E}\left[\int_{I_j} \text{hist}_{\vec{X}}(x) dx\right] = \mathbb{P}(X \in I_j)$$

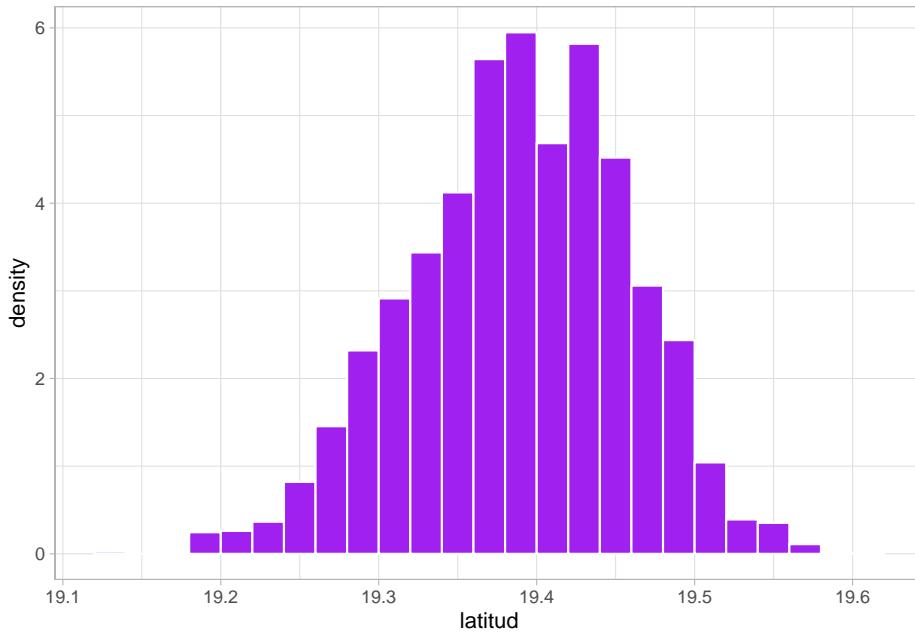
Es decir, el valor esperado del área bajo un histograma en un intervalo I_j coincide con la probabilidad de que X pertenezca a dicho intervalo.

Gráficamente:



En R podemos hacer un histograma a través de `geom_histogram`. En este caso lo haremos de la latitud:

```
#En este caso binwidth = h y kappa = boundary
ggplot(datos) +
  geom_histogram(aes(x = latitud, y = ..density..),
                 binwidth = 0.02, boundary = -99,
                 color = "white", fill = "purple") +
  theme_light()
```



3.12.1 Ejercicio

Considera la siguiente base de datos (obtenida de Cross Validated):

```
mis.datos <- data.frame(
  A = c(3.15, 5.46, 3.28, 4.20, 1.98, 2.28, 3.12, 4.10, 3.42, 3.91,
        2.06, 5.53, 5.19, 2.39, 1.88, 3.43, 5.51, 2.54, 3.64, 4.33,
        4.85, 5.56, 1.89, 4.84, 5.74, 3.22, 5.52, 1.84, 4.31, 2.01,
        4.01, 5.31, 2.56, 5.11, 2.58, 4.43, 4.96, 1.90, 5.60, 1.92),
  B = c(2.90, 5.21, 3.03, 3.95, 1.73, 2.03, 2.87, 3.85, 3.17, 3.66,
        1.81, 5.28, 4.94, 2.14, 1.63, 3.18, 5.26, 2.29, 3.39, 4.08,
        4.60, 5.31, 1.64, 4.59, 5.49, 2.97, 5.27, 1.59, 4.06, 1.76,
        3.76, 5.06, 2.31, 4.86, 2.33, 4.18, 4.71, 1.65, 5.35, 1.67),
  C = c(2.65, 4.96, 2.78, 3.70, 1.48, 1.78, 2.62, 3.60, 2.92, 3.41,
        1.56, 5.03, 4.69, 1.89, 1.38, 2.93, 5.01, 2.04, 3.14, 3.83,
        4.35, 5.06, 1.39, 4.34, 5.24, 2.72, 5.02, 1.34, 3.81, 1.51,
        3.51, 4.81, 2.06, 4.61, 2.08, 3.93, 4.46, 1.4, 5.1, 1.42),
  D = c(2.40, 4.71, 2.53, 3.45, 1.23, 1.53, 2.37, 3.35, 2.67, 3.16,
        1.31, 4.78, 4.44, 1.64, 1.13, 2.68, 4.76, 1.79, 2.89, 3.58,
        4.10, 4.81, 1.14, 4.09, 4.99, 2.47, 4.77, 1.09, 3.56, 1.26,
        3.26, 4.56, 1.81, 4.36, 1.83, 3.68, 4.21, 1.15, 4.85, 1.17)
)
```

Grafica un histograma de las variables A, B, C y D de dicha base con un ancho de banda (binwidth) igual a 1.

- ¿Podemos concluir la forma de la distribución a partir del histograma? Es decir ¿hay distribuciones sesgadas a la izquierda, a la derecha, uniformes, centradas o con colas pesadas?
- Realiza el mismo histograma pero ahora con un ancho de banda de 0.25 ¿por qué hubo cambios?
- Analiza la base de datos (los valores en función de la columna A) y concluye.

3. Densidad kernel Un histograma tiene muchos bemoles: en particular, es necesario decidir quién es h y quién κ y no hay una regla clara de cómo hacerlo. La densidad **kernel** es un intento de mejorar esta situación. Para ello recordamos que si X es una variable aleatoria continua con densidad F entonces:

$$f(x) = F'(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x-h)}{2h}$$

Por lo que para un h positiva con $h \approx 0$ tenemos que:

$$f(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

En el caso de un vector de observaciones $\vec{x} = (x_1, x_2, \dots, x_n)^T$ recordamos que podemos asociar una función de distribución empírica \hat{F} y por tanto obtener el *estimador de Rosenblatt* de la densidad f mediante:

$$\hat{f}(x) = \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h}$$

Podemos reescribir esto como:

$$\hat{f}(x) = \frac{1}{2nh} \sum_{i=1}^n \mathbb{I}_{(x-h, x+h]}(x_i) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)$$

donde:

$$K(u) = \frac{1}{2} \mathbb{I}_{(-1,1]}(u)$$

se conoce como el *kernel rectangular*. Una vez que llegamos hasta este punto notamos que para cualquier K que cumple:

$$1. \int_{-\infty}^{\infty} K(u) du = 1$$

$$2. K(u) \geq 0$$

la función \hat{f} es una función de densidad. La función \hat{f} se conoce como el *estimador de densidad del kernel K*. Algunos ejemplos de kernels K son:

- Rectangular:** $K(u) = \frac{1}{2} \mathbb{I}_{(-1,1]}(u)$
- Triangular:** $K(u) = (1 - |u|) \mathbb{I}_{(-1,1]}(u)$

3. **Epanechnikov:** $K(u) = \frac{3}{4}(1 - u^2)\mathbb{I}_{(-1,1]}(u)$

4. **Gaussiano:** $K(u) = \frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$

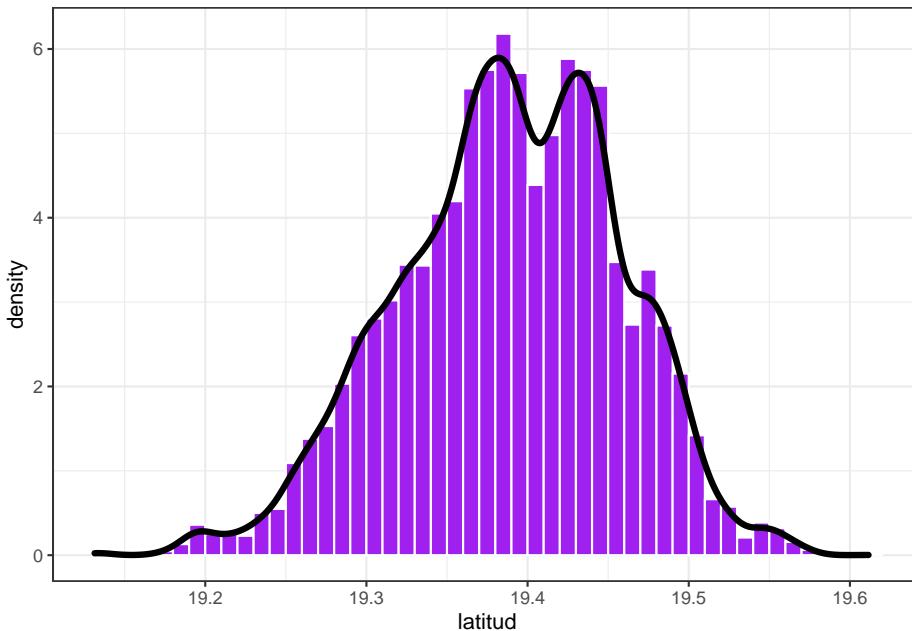
OJO No confundir el Kernel K (que es una función que integra a 1) con *función de densidad kernel* que es una función de los datos que utiliza un kernel y es una densidad por sí misma.

En R podemos calcular la densidad `kernel` en `n` puntos con relativa facilidad mediante `density`:

```
densidad_kernel <- density(datos$latitud, kernel = "gaussian", n = 700,
                             na.rm = TRUE)
```

Nota que R en automático preselecciona los valores de `h` mediante un criterio preprogramado de optimización. Podemos ver dicha densidad gráficamente (y compararla con un histograma):

```
ggplot(datos) +
  geom_histogram(aes(x = latitud, y = ..density..),
                 binwidth = 0.01, boundary = 19,
                 fill = "purple", color = "white") +
  geom_density(aes(x = latitud), kernel = "gaussian", size = 1.5) +
  theme_bw()
```

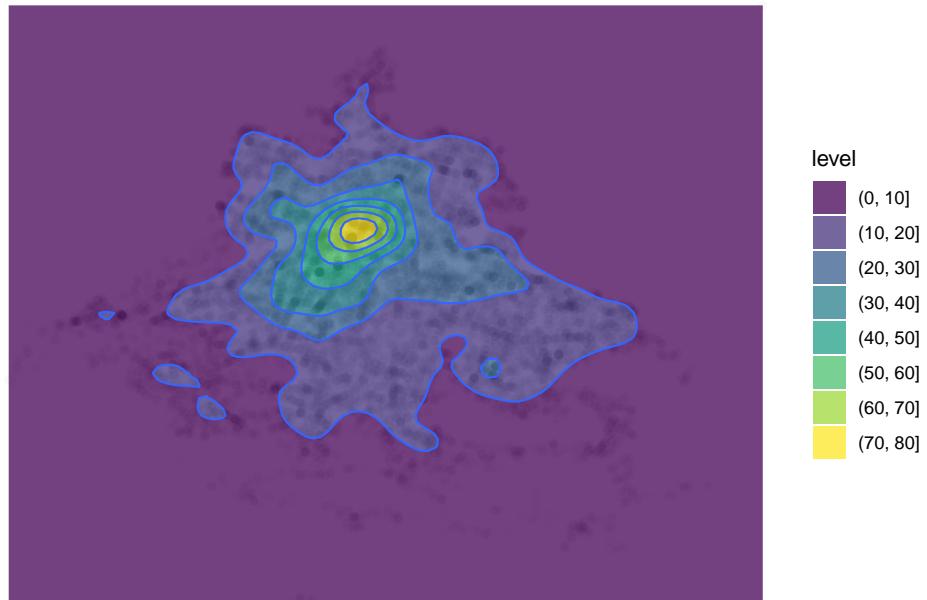


Esto no se queda ahí, podemos generalizar el concepto de kernel a dos dimensiones para aproximar una función de densidad $f(x, y)$ de dos variables aleatorias si tenemos dos vectores \vec{x} y \vec{y} y calculamos:

$$\hat{f}(x, y) = \frac{1}{nh^2} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)K\left(\frac{y_i - y}{h}\right)$$

En particular esto nos permite generar una densidad en R para saber en qué coordenadas de latitud y longitud ocurren más los delitos:

```
ggplot(datos) +
  geom_point(aes(x = longitud, y = latitud), alpha = 0.025) +
  geom_density_2d_filled(aes(x = longitud, y = latitud), alpha = 0.75) +
  geom_density2d(aes(x = longitud, y = latitud)) +
  theme_void()
```



3.12.2 Ejercicio sugerido

Este ejercicio es para que tengas la seguridad de que comprendiste los conceptos previos y sabes calcularlos. Es tedioso pero bueno para aclarar dudas.

Considera la siguiente base de datos:

Calcula **a mano** (es decir puedes usar calculadora pero no lo calcules en R) y luego **verifica tus cálculos haciéndolo en R**:

1. El total de \vec{x}
2. La media y varianza de \vec{y}
3. La curtosis y la asimetría de \vec{x} (su media es 2 y su varianza 0.8). Determina si tiene un sesgo a la derecha, a la izquierda o ninguno.

x	y	z	w
1	-100	Rojo	Bueno
2	-2	Azul	Malo
3	2	Azul	Regular
2	3	Rojo	Bueno
1	1	Verde	Bueno
3	4	Amarillo	Malo

	CDMX	MTY	Total
Chocó	1100	4000	5100
No Chocó	120	5080	5200
Total	1220	9080	10300

4. Determina mediante la curtosis si \vec{x} tiene colas más pesadas que \vec{y} .
5. Calcula el cuantil 0.25 y el 0.75 de \vec{y} así como su rango intercuartílico (IQR).
6. ¿Hay valores atípicos (*outliers*) en \vec{y} ? En caso afirmativo, determina cuáles son.
7. ¿Cuál es el rango de \vec{y} ? (no confundir con el IQR).
8. Determina la moda de \vec{z} .
9. Determina la mediana de \vec{x} .
10. Determina la MAD de \vec{x} .
11. Realiza el conteo de cuáles \vec{z} pertenecen al conjunto $A = \{\text{Rojo, Amarillo}\}$
12. Realiza una tabla de contingencia de \vec{w} y \vec{z} .
13. Determina la distribución frecuencial (observada) marginal de \vec{w} .
14. Realiza una tabla de frecuencias de \vec{w} y \vec{z} .
15. Calcula el riesgo relativo de estar en un choque dado que manejas en CDMX a partir de los datos en la tabla:

Interpreta tu resultado.

16. De la tabla anterior calcula la razón de momios asociada a chocar dado que manejas en CDMX. Interprétala.
17. Calcula la correlación de Bravais Pearson de \vec{x} y \vec{y} . Interpreta.
18. Obtén la correlación de Spearman de \vec{x} y \vec{y}
19. Para \vec{w} y \vec{x} obtén la τ de Kendall (son 15 comparaciones para generarla)
20. Descartando el *outlier* de \vec{y} (y su \vec{x} asociada), ajusta un modelo lineal $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$ y graficalo para ver qué tan buen modelo es.

21. Realiza una gráfica de caja (boxplot) para \vec{y}
22. Realiza un scatterplot para la submatriz $Z_{(x,y)}$.
23. Realiza una gráfica de líneas para $Z_{(x,y)}$ identificando la función de interpolación lineal $f(x)$ asociada.
24. Realiza una gráfica de barras de \vec{w} especificando quiénes son los a_i y los n_j .
25. Estima mediante \hat{p} la función de probabilidad de \vec{w} .
26. Identifica la función de distribución empírica para \vec{x} , \hat{F} y graficala.
27. Realiza un histograma con $h = 2$ para x . Toma $\kappa = 4$.
28. Ajusta una densidad kernel a \vec{x} con $h = 1$ y usando un kernel K triangular. Calcula $\hat{f}(x)$ para $x = 0, 1, 2, 3, 4$.

3.13 Ejercicios del capítulo

1. Dado \vec{x} vector de variables ordinales, obtén una expresión matemática para los siguientes estadísticos:
 - a. La media de las diferencias entre las x_i quitando la de x_k consigo misma.
 - b. El valor numérico o categoría menos común en \vec{x} .
 - c. Si ordenamos todos los valores, la diferencia más alta entre algún $x_{(i)}$ y su sucesor: $x_{(i+1)}$.
 - d. Este cálculo de R para una S dada como vector numérico:

```
#x es la muestra; x <- c(x1,x2, ..., xn)
datos_nuevos <- c()
for (i in 1:length(x)){
  datos_nuevos <- c(datos_nuevos, x[i]^i)
}
mean(datos_nuevos) #Este valor es el que me interesa
```

2. Demuestra que:

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

3. Para unos datos observados numéricos $\vec{x} = (x_1, x_2, \dots, x_n)^T$ se tiene que las x toman el valor $a_{x,1}$ $n_{x,1}$ veces, el valor $a_{x,2}$, $n_{x,2}$ veces y el valor $a_{x,\ell}$, $n_{x,\ell}$ veces ($n_j \geq 0$, $0 < \ell \leq n$ y $\sum_{j=1}^{\ell} n_j = n$). Demuestra que:

$$\bar{x} = \frac{1}{\sum_{j=1}^{\ell} n_j} \cdot \sum_{j=1}^{\ell} n_j a_j$$

4. Sea n impar y f una función estrictamente decreciente.
- Demuestra que si x_* es la mediana de $\vec{x} = (x_1, x_2, \dots, x_n)^T$ entonces $f(x_*)$ es la mediana de $\tilde{\vec{x}} = (f(x_1), f(x_2), \dots, f(x_n))^T$.
 - Demuestra que si \bar{x} es la media observada de $\vec{x} = (x_1, x_2, \dots, x_n)^T$ y $f(\bar{x})$ es la media observada de $\tilde{\vec{x}} = (f(x_1), f(x_2), \dots, f(x_n))^T$ y además f es diferenciable, entonces $f(x) = a \cdot x + b$ (es decir es una transformación afín). **Hint** Deriva.
5. Sea ϕ una función convexa. Demuestra que:

$$\phi(\bar{x}) \leq \frac{1}{n} \sum_{i=1}^n \phi(x_i)$$

Hint Deriva. Recuerda que si ϕ es convexa, para $0 \leq \alpha \leq 1$ se tiene que:

$$\phi(\alpha x + (1 - \alpha)y) \leq \alpha\phi(x) + (1 - \alpha)\phi(y)$$

6. Sea $\hat{p}_R(x)$ la densidad kernel asociada a $\vec{x} = (x_1, x_2, \dots, x_n)$ con un núcleo (kernel) $K(u) \geq 0$ y $h > 0$. Demuestra:
- $\hat{p}_h(x)$ es una función de densidad de probabilidad (*i.e.* integra a 1).
 - Determina la media de una variable aleatoria X que se distribuye con densidad $\hat{p}_h(x)$ bajo: Kernel triangular.
 - Determina la varianza de una variable aleatoria X que se distribuye con densidad $\hat{p}_h(x)$ bajo: Kernel Epanechnikov
7. Sea $\text{hist}_{\vec{x}}$ la función de histograma para un vector numérico \vec{x} con $h > 0$, $\kappa \in \mathbb{R}$ fijos y una partición $\{I_j\}_{j \in \mathbb{Z}}$. Demuestra que $\text{hist}_{\vec{x}}$ es una función de densidad.
8. Demuestra que para $\vec{x} = (x_1, x_2, \dots, x_n)^T$ la función de distribución empírica $\hat{F}(x)$ es una función de distribución acumulada; es decir:
- $\lim_{x \rightarrow -\infty} \hat{F}(x) = 0$
 - $\lim_{x \rightarrow \infty} \hat{F}(x) = 1$
 - $\lim_{x \rightarrow x_0^+} \hat{F}(x) = \hat{F}(x_0)$ (continua por la derecha)
 - $\lim_{x \rightarrow x_0^-} \hat{F}(x)$ existe
 - $F_n(x)$ es no decreciente.
9. La tabla ?? muestra datos observados del PIB de un país en billones de dólares:

Ajusta una parábola $q(x) = a \cdot x^2 + b \cdot x + c$ para obtener la mejor parábola que ajuste esos puntos. ¿Qué valor de PIB se espera para el 2020 bajo este modelo?

10. Demuestra que si $\vec{x} = -\vec{y}$ (dos vectores numéricos) la correlación de Spearman entre ambos es -1 .

Año	PIB
2000	0.5
2005	1.2
2010	1.5
2015	2.1

11. Para una variable aleatoria T que representa un tiempo, se define una función de supervivencia como la probabilidad de que T dure más que un cierto tiempo t ; es decir:

$$S(t) = \mathbb{P}(T > t)$$

Construye \hat{S} una aproximación empírica a la función de supervivencia S tal que $\mathbb{E}[\hat{S}(t)] = S(t)$. Demuestra este último resultado.

12. Demuestra que $\text{Var}[\hat{F}(x)] = \frac{1}{n}F(x)(1 - F(x))$.

13. Recuerda que para dos variables aleatorias X y Y se define la covarianza $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. Calcula entonces la covarianza dada por $\text{Cov}(\hat{p}(u), \hat{p}(v))$.

14. Demuestra que si X es independiente de Y entonces $\mathbb{I}_A(X)$ es independiente de $\mathbb{I}_A(Y)$. Recuerda que dos variables aleatorias X, Y son independientes si y sólo si $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$ para conjuntos A, B medibles.

15. Sean ρ la ρ de Spearman y τ la τ de Kendall para dos vectores ordinales \vec{x} y \vec{y} . Demuestra que:

$$\frac{1 + \rho}{2} \geq \left(\frac{1 + \tau}{2} \right)^2$$

16. Da un ejemplo de vector \vec{x} de al menos dos entradas tal que $\text{MAD}_{\vec{x}} \geq \sigma_{\vec{x}}$

17. Demuestra que $\text{MAD}_{\vec{x}} = 0 \Leftrightarrow \sigma_{\vec{x}} = 0$ para el mismo vector \vec{x} .

18. Si un vector \vec{x} tiene 3 entradas, media $\bar{x} = 1$ y varianza $\sigma_{\vec{x}} = 1$ y además se sabe que su curtosis es 1, ¿quién es \vec{x} ?

19. Bajo la correlación de Pearson demuestra que para vectores \vec{x}, \vec{y} y \vec{z} si $\rho_{\vec{x}, \vec{y}} = 1$ y $\rho_{\vec{x}, \vec{z}} = 1$ entonces $\rho_{\vec{y}, \vec{z}} = 1$

20. Sean $\vec{w}, \vec{x}, \vec{y}, \vec{z} \in \mathbb{R}^n$ y $a, b, c, d \in \mathbb{R}$ demuestra que:

$$\rho(a\vec{x} + b\vec{w}, c\vec{y} + d\vec{z}) = K_1 \cdot \rho(\vec{x}, \vec{y}) + K_2 \cdot \rho(\vec{w}, \vec{y}) + K_3 \cdot \rho(\vec{x}, \vec{z}) + K_4 \cdot \rho(\vec{w}, \vec{z})$$

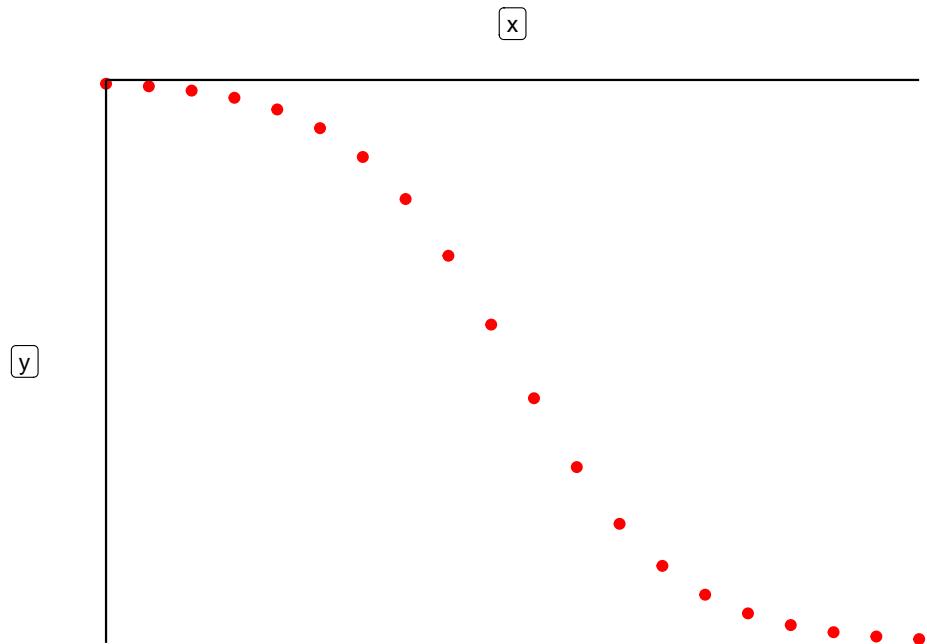
para algunas constantes K_1, K_2, K_3, K_4 ; donde, además, ρ es la correlación de Pearson.

21. Sean $\vec{x} = (x_1, x_2, \dots, x_n)^T$ el vector de los datos observados (fijo) y $\vec{X} = (X_1, X_2, \dots, X_n)^T$ el vector de los datos posibles (aleatorio). Supongamos que las entradas de \vec{X} son independientes e idénticamente distribuidas con la misma distribución de X con media μ y varianza σ^2 .

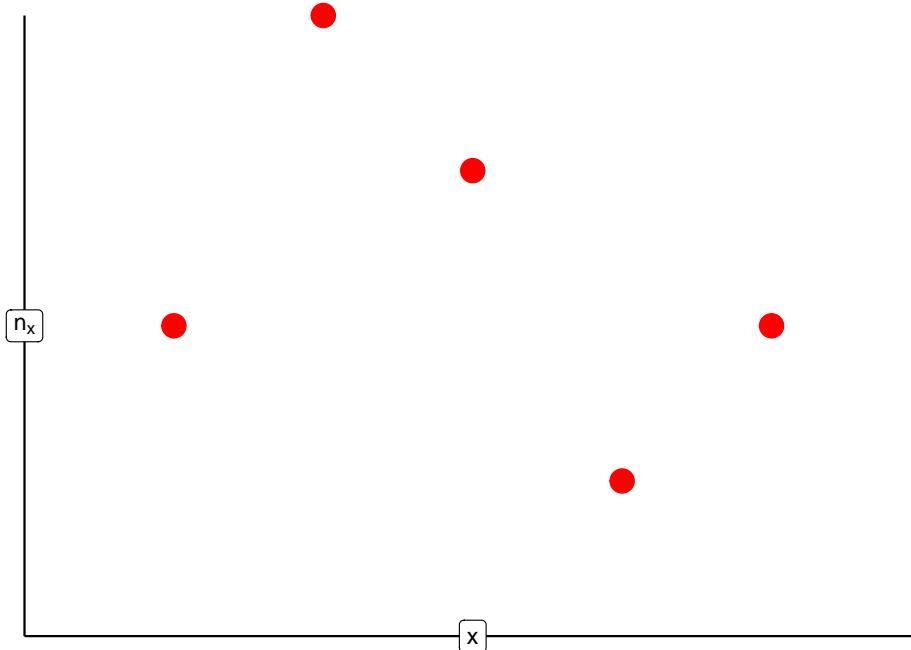
	Consomé	Taco	Postre
Tacos el Güero	1	3	2
Tacos la Güera	4	2	1

Calificación	Cantidad de alumnos
10	3
9	4
8	12
7	11
6	1
5	4

- a. Demuestra que $\mathbb{E}[\bar{X}] = \mu$ ¿Qué te dice esto de \bar{x} ?
- b. Demuestra que $\mathbb{E}[\sigma_X^2] \neq \sigma^2$ ¿Qué te dice esto de $\sigma_{\bar{x}}^2$? Donde $\sigma_X^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$. Hint: usa el ejercicio 2 de esta sección.
22. Quiénes son \vec{x}, \vec{y} si su tabla de contingencia es la dada por Cuadro ??:
23. Da un ejemplo de vectores \vec{x} y \vec{y} de tamaño 10 cuya τ de Kendall sea 0 pero estén completamente relacionados; es decir exista una función g tal que $\vec{y} = g(\vec{x})$.
24. Construye una función **function** en R que dado un vector $x <- c(x_1, x_2, \dots, x_n)$ regrese la densidad kernel (bajo kernel gaussiano) asociada a x y evaluada en los puntos dados por el vector $t <- c(t_1, t_2, \dots, t_m)$.
25. Considera la siguiente base de datos de calificaciones. Calcula la mediana de calificaciones, media, varianza y el IQR.
26. A partir de la siguiente gráfica determina si los incisos son verdaderos o falsos o no se puede determinar:



- a. La correlación de Pearson de \vec{x} y \vec{y} es -1
 - Verdadero
 - Falso
 - No se puede determinar
 - b. La correlación de Spearman de \vec{x} y \vec{y} es -1
 - Verdadero
 - Falso
 - No se puede determinar
 - c. La tau de Kendall de \vec{x} y \vec{y} es -1
 - Verdadero
 - Falso
 - No se puede determinar
27. A partir de la siguiente gráfica determina si los incisos son verdaderos o falsos o no se puede determinar:



- a. El coeficiente de asimetría de \vec{x} es positivo.
 - Verdadero
 - Falso
 - No se puede determinar
 - b. El coeficiente de curtosis de \vec{x} es positivo.
 - Verdadero
 - Falso
 - No se puede determinar
 - c. La distribución de \vec{x} tiene una sola moda.
 - Verdadero
 - Falso
 - No se puede determinar
28. Considera el problema de mínimos cuadrados donde ahora suponemos que existe una submatriz $X_{n \times k}$ de la base de datos Z y un vector columna y tal que cada entrada de las y , y_i es una función lineal de las $x_{i,j}$:

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k}$$

Suponemos, además que las columnas de $X_{n \times k}$ son linealmente independientes (*i.e.* es de rango completo). En este caso la función de error a minimizar para estimar las β s es:

$$\text{SSR}(\beta_1, \beta_2, \dots, \beta_k) = \sum_{i=1}^n \left(y_i - \sum_{j=1}^k \beta_j x_{i,j} \right)^2$$

- a. Demuestra que en este caso el problema es equivalente a minimizar:

$$\text{SSR}(\vec{\beta}) = (\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta})$$

- b. Obtén entonces que:

$$\vec{\beta} = (X^T X)^{-1} X^T \vec{y}$$

- c. Finalmente, suponiendo que $y = \alpha + \gamma x + \eta z$ es un hiperplano de x y z calcula los coeficientes:

x	y	z
1	7	-1
2	9	-2
3	12	-3
1	14	1

- d. Verifica los coeficientes haciendo la regresión en R.

Chapter 4

Muestreo Aleatorio Simple

4.1 Inicio

Siempre que inicies un nuevo trabajo en R ¡no olvides borrar el historial!

```
rm(list = ls()) #Clear all
```

4.2 Librerías

Para este análisis vamos a tener que llamar a las siguientes librerías previamente instaladas (por única vez) con `install.packages`:

```
library(tidyverse)
library(dplyr)
library(imager)
library(rlist)
library(gridExtra)
library(kableExtra)
```

4.3 Notación

Supongamos una matriz de datos de tamaño $N \times L$ dada por:

$$U = (z_1 | z_2 | \dots | z_\ell)$$

donde las z_i representan las columnas de la matriz. La U será conocida como la **matriz universo** (el *universo* ó *la población*) si contiene *toda* la información de la población. Intuitivamente, la matriz U son todos los datos de un censo: esta es una matriz *ideal* donde están todos los datos.

A cualquier permutación en las filas de una submatriz S (tamaño $n \times \ell$ con $0 < n \leq N$ y $0 < \ell \leq L$) de U se le conoce como una *muestra* de U . Si S es una variable aleatoria (por ejemplo, porque se construyó a partir de un mecanismo de aleatoriedad) decimos que S es una **muestra aleatoria** (denotamos \mathcal{S} a la variable aleatoria y S a un valor específico de la misma).¹

En particular en esta sección (y hasta establecer lo contrario) consideraremos que el universo U es de tamaño $N \times 1$ y la submatriz S (resp. \mathcal{S}) es de tamaño $n \times 1$. En notación $U = (x_1, x_2, \dots, x_N)^T$ y *técnicamente* $S = (x_{i_1}, x_{i_2}, \dots, x_{i_n})^T$ para un conjunto de índices i_1, i_2, \dots, i_n . Sin embargo para simplificar la notación consideraremos que en S están los primeros n de los x_i ; es decir:

$$S = (x_1, x_2, \dots, x_n)^T$$

Cuando estemos hablando de la muestra *como variable aleatoria* \mathcal{S} y no como *valores observados (fijos)* S , denotaremos:

$$\mathcal{S} = (X_1, X_2, \dots, X_n)^T$$

donde \mathcal{S} representa la muestra posible y cada X_i es una variable aleatoria con el valor posible de la i -ésima entrada.

Un **esquema muestral** es una función \mathbb{P} de probabilidad definida en el conjunto de submatrices de U . Ésta es el punto medular de todas las estrategias de muestreo: distintos esquemas muestrales generan diferentes distribuciones y pueden llevar a distintas inferencias sobre un fenómeno.

4.3.1 Ejemplo

Considera la matriz universo con tres letras:

$$U = \begin{pmatrix} A \\ B \\ C \end{pmatrix}$$

Ésta es la matriz universo. Las submatrices² que pueden crearse a partir de dicho universo son:

1. De dimensión $n = 1$: $S^1 = (A)^T, S^2 = (B)^T, S^3 = (C)^T$.
2. De dimensión $n = 2$: $S^4 = (A, B)^T, S^5 = (A, C)^T, S^6 = (B, C)^T, S^7 = (B, A)^T, S^8 = (C, A)^T, S^9 = (C, B)^T$.
3. De dimensión $n = 3$: $S^{10} = (A, B, C)^T, S^{11} = (B, A, C)^T, S^{12} = (A, C, B)^T, S^{13} = (C, B, A)^T, S^{14} = (B, C, A)^T, S^{15} = (C, A, B)^T$,

¹En la literatura muchas referencias establecen una muestra aleatoria como un conjunto de valores. Yo utilizo vectores para poder hablar de repeticiones (por ejemplo si extraes el mismo valor varias veces en la muestra).

²Enumero las submatrices para luego poder hablar de ellas

Un esquema muestral sería la función de probabilidad:

$$\mathbb{P}(\mathcal{S} = S^k) = \begin{cases} 0.1 & \text{si } k = 1, \\ 0.2 & \text{si } k = 3, \\ 0.5 & \text{si } k = 11, \\ 0.2 & \text{si } k = 15, \\ 0 & \text{en otro caso.} \end{cases}$$

Otro esquema muestral posible sería:

$$\mathbb{P}(\mathcal{S} = S^k) = \begin{cases} \frac{1}{3} & \text{si } k = 1, \\ \frac{1}{3} & \text{si } k = 2, \\ \frac{1}{3} & \text{si } k = 3, \\ 0 & \text{en otro caso.} \end{cases}$$

Este último esquema, intuitivamente, corresponde a la selección aleatoria de un elemento de U con una probabilidad uniforme de que cada elemento salga.

A fin de simplificar el problema (y hasta que se diga lo contrario) agregaremos la hipótesis de **intercambiabilidad**; es decir, consideraremos es irrelevante el orden de las filas de las submatrices de datos. Por ejemplo, bajo intercambiabilidad, $S^4 = (A, B)^T$ es *la misma matriz* que $S^7 = (B, A)^T$.

Un ejemplo de muestra donde el orden sí importa (*i.e.* no son intercambiables) es cuando se realizan exámenes orales según una selección aleatoria de la lista. La tercera persona en presentar el examen estará informada por el *¿qué te preguntó el profe?, ¿estuvo difícil?* que las primeras dos le cuenten.

Bajo intercambiabilidad, los esquemas muestrales estarán definidos únicamente sobre los siguientes vectores:

1. De dimensión $n = 1$: $S^1 = (A)^T$, $S^2 = (B)^T$, $S^3 = (C)^T$.
2. De dimensión $n = 2$: $S^4 = (A, B)^T$, $S^5 = (A, C)^T$, $S^6 = (B, C)^T$.
3. De dimensión $n = 3$: $S^7 = (A, B, C)^T$.

En este caso un esquema muestral sería:

$$\mathbb{P}(\mathcal{S} = S^k) = \begin{cases} \frac{1}{16} & \text{si } k = 1, \\ \frac{3}{16} & \text{si } k = 2, \\ 0 & \text{si } k = 3, \\ \frac{7}{16} & \text{si } k = 4, \\ \frac{1}{16} & \text{si } k = 5, \\ \frac{4}{16} & \text{si } k = 6, \\ 0 & \text{en otro caso.} \end{cases}$$

Dado un elemento x_i del universo, podemos preguntarnos por la probabilidad de que dicho x_i esté en la muestra. Siguiendo el ejemplo anterior:

$$\mathbb{P}(A \in \mathcal{S}) = \mathbb{P}(\mathcal{S} = S^1) + \mathbb{P}(\mathcal{S} = S^4) + \mathbb{P}(\mathcal{S} = S^5) + \mathbb{P}(\mathcal{S} = S^7) = \frac{9}{16}.$$

Como notación, para una población $U = (x_1, x_2, \dots, x_N)^T$ y una muestra aleatoria \mathcal{S} denotamos la probabilidad de que x_k esté en la muestra como:

$$\pi_k = \mathbb{P}(x_k \in \mathcal{S})$$

Estas probabilidades (para $k = 1, 2, \dots, N$) se conocen como **probabilidades de inclusión de primer orden**. La probabilidad conjunta de que x_k y x_l (ambos) estén en la muestra (**probabilidad de inclusión de segundo orden**) está dada por:

$$\pi_{k,l} = \mathbb{P}(x_k \in \mathcal{S}, x_l \in \mathcal{S})$$

Notamos que por definición $\pi_{kk} = \pi_k$. Análogamente se pueden crear probabilidades de inclusión de cualquier orden deseado.

Finalmente, una población $U = (x_1, x_2, \dots, x_N)^T$ y una muestra aleatoria \mathcal{S} definimos la variable indicadora de que x_k esté en la muestra como:

$$\mathbb{I}_{\mathcal{S}}(x_k) = \begin{cases} 1 & \text{si } x_k \in \mathcal{S} \\ 0 & \text{si } x_k \notin \mathcal{S} \end{cases}$$

Notamos que para una muestra aleatoria \mathcal{S} las indicadoras tienen una distribución conocida:

$$\mathbb{I}_{\mathcal{S}}(x_k) \sim \text{Bernoulli}(\pi_k)$$

pues

$$\mathbb{P}(\mathbb{I}_{\mathcal{S}}(x_k) = 1) = \mathbb{P}(x_k \in \mathcal{S}) = \pi_k$$

Como las $\mathbb{I}_{\mathcal{S}}(x_k)$ son Bernoulli podemos calcular su varianza:

$$\text{Var}(\mathbb{I}_{\mathcal{S}}(x_k)) = \pi_k(1 - \pi_k)$$

Finalmente, recordamos que la covarianza entre dos variables aleatorias X, Y se define como:

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

Por lo que calculamos la covarianza entre dos indicadoras (de x_k y x_l):

$$\begin{aligned} \text{Cov}(\mathbb{I}_{\mathcal{S}}(x_k), \mathbb{I}_{\mathcal{S}}(x_l)) &= \mathbb{E}[\mathbb{I}_{\mathcal{S}}(x_k) \cdot \mathbb{I}_{\mathcal{S}}(x_l)] - \mathbb{E}[\mathbb{I}_{\mathcal{S}}(x_k)] \mathbb{E}[\mathbb{I}_{\mathcal{S}}(x_l)] \\ &= 1 \cdot \mathbb{P}(\mathbb{I}_{\mathcal{S}}(x_k) \cdot \mathbb{I}_{\mathcal{S}}(x_l) = 1) + 0 \cdot \mathbb{P}(\mathbb{I}_{\mathcal{S}}(x_k) \cdot \mathbb{I}_{\mathcal{S}}(x_l) = 0) - \pi_k \pi_l \\ &= \mathbb{P}(\mathbb{I}_{\mathcal{S}}(x_k) = 1, \mathbb{I}_{\mathcal{S}}(x_l) = 1) - \pi_k \pi_l \\ &= \pi_{k,l} - \pi_k \pi_l \end{aligned} \tag{4.1}$$

La cantidad $\pi_{k,l} - \pi_k \pi_l$ usualmente se denota $\Delta_{k,l}$:

$$\Delta_{k,l} = \pi_{k,l} - \pi_k \pi_l$$

A continuación hablaremos de algunos esquemas de muestreo comunmente utilizados y, finalmente, llegaremos a una generalización de los mismos.

4.3.2 Ejercicio

Demuestra las siguientes propiedades de los π_k para un diseño muestral \mathbb{P} con tamaño fijo de la muestra $n \in \mathbb{N}$:

1. $\sum_{k=1}^N \pi_k = n$
2. $\sum_{k=1}^N \sum_{\substack{l=1 \\ k \neq l}}^N \pi_{k,l} = n(n-1)$
3. $\sum_{\substack{l=1 \\ l \neq k}}^N \pi_{k,l} = (n-1)\pi_k$

4.4 Muestreo Aleatorio Simple sin Reemplazo (MAS/sR)

Vamos a considerar una de las formas más sencillas de muestreo: el aleatorio simple *sin reemplazo*. Para ello seleccionamos de $U = (x_1, x_2, \dots, x_N)^T$ a $n \in \mathbb{N}$ (fijo) observaciones asignándole la probabilidad de ser seleccionada a cada una de $\frac{1}{N}$. Una vez se selecciona la primera, se selecciona una de las que restan de U con probabilidad $\frac{1}{N-1}$. El proceso se repite hasta extraer n elementos.

Comencemos por un ejemplo, supongamos tenemos una población de cinco personas:

$$U = \left(\text{Ana, Beto, Carlos, Diana, Enriqueta} \right)^T$$

Si queremos tomar una muestra de 3 personas sin reemplazo, las muestras posibles son:

1. $\left(\text{Ana, Beto, Carlos} \right)^T$
2. $\left(\text{Ana, Carlos, Diana} \right)^T$
3. $\left(\text{Ana, Beto, Diana} \right)^T$
4. $\left(\text{Ana, Beto, Enriqueta} \right)^T$

5. $\left(\text{Ana}, \text{Carlos}, \text{Enriqueta} \right)^T$
6. $\left(\text{Ana}, \text{Diana}, \text{Enriqueta} \right)^T$
7. $\left(\text{Beto}, \text{Carlos}, \text{Diana} \right)^T$
8. $\left(\text{Beto}, \text{Diana}, \text{Enriqueta} \right)^T$
9. $\left(\text{Beto}, \text{Carlos}, \text{Enriqueta} \right)^T$
10. $\left(\text{Carlos}, \text{Diana}, \text{Enriqueta} \right)^T$

Obtener una muestra aleatoria se puede hacer en R con un vector mediante `sample`:

```
#Vector de nombres
nombres <- c("Ana", "Beto", "Carlos", "Diana", "Enriqueta")

#Muestra
sample(nombres, 3, replace = FALSE)

## [1] "Beto"  "Ana"   "Diana"
```

Formalmente, un esquema de muestreo es **aleatorio simple sin reemplazo** si dada una constante $n \in \mathbb{N}$ (con $0 < n \leq N$) se tiene:

$$\mathbb{P}(\mathcal{S} = S) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{si } \#\mathcal{S} = n \\ 0 & \text{en otro caso.} \end{cases}$$

En el caso de muestreo aleatorio simple sin reemplazo podemos calcular las probabilidades de inclusión como siguen:

$$\pi_k = \mathbb{P}(x_k \in \mathcal{S}) = \sum_{i=1}^{M_1} \frac{1}{\binom{N}{n}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} = f$$

donde la tercera igualdad se sigue de que hay $M_1 = \binom{N-1}{n-1}$ muestras que contienen al x_k . (La lógica es, fijo el x_k y entonces me quedan $N-1$ valores de x a acomodar en $n-1$ espacios). Por otro lado:

$$\pi_{k,j} = \mathbb{P}(x_k \in \mathcal{S}, x_j \in \mathcal{S}) = \sum_{i=1}^{M_2} \frac{1}{\binom{N}{n}} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

pues hay $M_2 = \binom{N-2}{n-2}$ muestras conteniendo a x_k y x_j a la vez.

Para estimar el total poblacional dado por:

$$t = \sum_{i=1}^N x_i$$

bajo *MAS/sR* podemos tomar:

$$\hat{t} = N \cdot \bar{x}_{\mathcal{S}} = N \frac{1}{n} \sum_{k=1}^n x_k = \sum_{k=1}^n \frac{x_k}{n/N} = \sum_{k=1}^N \frac{x_k}{\pi_k} \cdot \mathbb{I}_{\mathcal{S}}(x_k)$$

Notamos entonces que el estimador \hat{t} es una variable aleatoria pues depende de las indicadoras de la muestra. En particular:

$$\mathbb{E}[\hat{t}] = \mathbb{E}\left[\sum_{k=1}^N \frac{x_k}{\pi_k} \cdot \mathbb{I}_{\mathcal{S}}(x_k)\right] = \sum_{k=1}^N \frac{x_k}{\pi_k} \underbrace{\mathbb{E}\left[\mathbb{I}_{\mathcal{S}}(x_k)\right]}_{\pi_k} = t$$

de donde se sigue que en promedio el estimador \hat{t} vale el total.

Definición: [Insesgado] Un estimador $\hat{\theta}$ es un estimador insesgado de θ si:

$$\mathbb{E}[\hat{\theta} - \theta] = 0$$

En nuestro caso \hat{t} es *insesgado*. En general, la cantidad $\mathbb{E}[\hat{\theta} - \theta]$ se conoce como *el sesgo*.

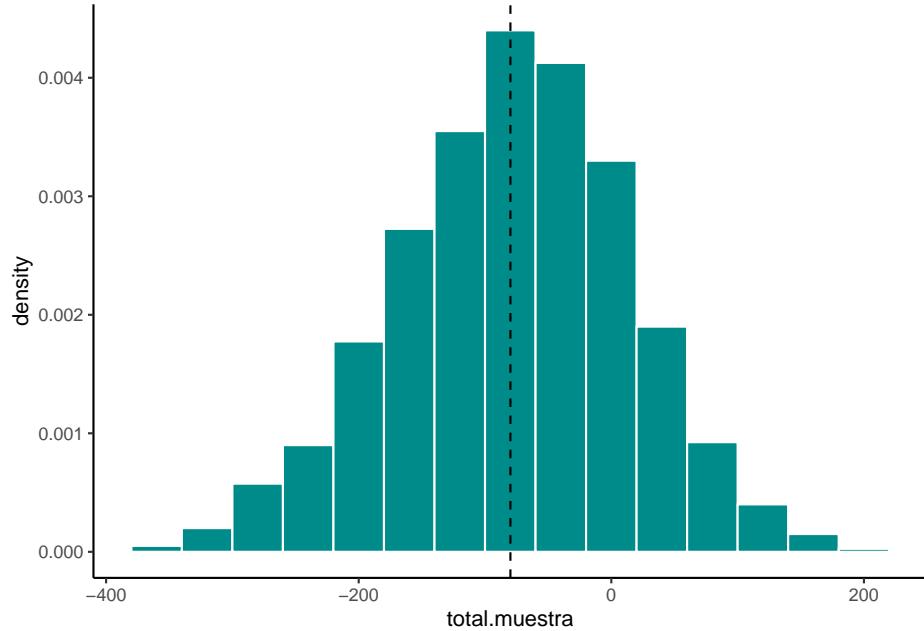
De manera numérica, podemos simular la estimación del total en 1000 simulaciones como sigue:

```
nsim <- 1000
N    <- 1000
n    <- 100
base.completa <- data.frame(x = rnorm(N))
total          <- sum(base.completa$x)
total.muestra <- rep(NA, nsim)
for (i in 1:nsim){
  muestra        <- sample(base.completa$x, n)
  total.muestra[i] <- N*mean(muestra)
}
mean(total.muestra)

## [1] -78.75412
```

Podemos ver las simulaciones como sigue:

```
ggplot() +
  geom_histogram(aes(x = total.muestra, y = ..density..), fill = "#008B8B",
                 color = "white", binwidth = 40) +
  geom_vline(aes(xintercept = total), linetype = "dashed") +
  theme_classic()
```



Como podrás notar la \hat{t} es una variable aleatoria y por tanto tiene varianza. De hecho:

$$\text{Var}(\hat{t}) = \sum_{k=1}^N \sum_{l=1}^N \Delta_{k,l} \frac{x_k x_l}{\pi_k \pi_l}$$

Para demostrarlo seguimos las igualdades:

$$\begin{aligned} \text{Var}(\hat{t}) &= \text{Var}\left(\sum_{k=1}^N \frac{x_k}{\pi_k} \cdot \mathbb{I}_S(x_k)\right) \\ &= \sum_{k=1}^N \frac{x_k^2}{\pi_k^2} \cdot \text{Var}\left(\mathbb{I}_S(x_k)\right) + \sum_{k=1}^N \sum_{l=1}^N \underbrace{\frac{x_k x_l}{\pi_k \pi_l}}_{l \neq k} \cdot \text{Cov}\left(\mathbb{I}_S(x_k), \mathbb{I}_S(x_l)\right) \\ &= \sum_{k=1}^N \frac{x_k^2}{\pi_k^2} \cdot \underbrace{\pi_k(1-\pi_k)}_{\Delta_{k,k}} + \sum_{k=1}^N \sum_{l=1}^N \underbrace{\frac{x_k x_l}{\pi_k \pi_l}}_{l \neq k} \cdot \underbrace{\text{Cov}\left(\mathbb{I}_S(x_k), \mathbb{I}_S(x_l)\right)}_{\Delta_{k,l}} \\ &= \sum_{k=1}^N \sum_{l=1}^N \Delta_{k,l} \frac{x_k x_l}{\pi_k \pi_l} \end{aligned}$$

Numéricamente, en el ejemplo anterior la varianza (simulada) de \hat{t} es:

```
var(total.muestra)
```

```
## [1] 8515.982
```

mientras que la *real* está dada por (ver ejercicio más adelante):

```
f      <- n/N
varianza <- N^2*(1 - f)/n*var(base.completa$x)
print(varianza)
```

```
## [1] 8278.693
```

Nota que tenemos un problema: para estimar $\text{Var}(\hat{t})$ necesitamos conocer todas las x_k de la población ¡lo cual es imposible! Entonces necesitamos un estimador de la varianza de \hat{t} para lo cual proponemos:

$$\widehat{\text{Var}}(\hat{t}) = \sum_{k=1}^n \sum_{l=1}^n \frac{\Delta_{k,l}}{\pi_{k,l}} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l}$$

Para demostrar que el estimador es insesgado tomamos el valor esperado y agregamos las variables indicadoras correspondientes:

$$\widehat{\text{Var}}(\hat{t}) = \sum_{k=1}^N \sum_{l=1}^N \frac{\Delta_{k,l}}{\pi_{k,l}} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l} \mathbb{I}_{\mathcal{S}}(x_k) \mathbb{I}_{\mathcal{S}}(x_l)$$

Se sigue la demostración:

$$\begin{aligned} \mathbb{E}[\widehat{\text{Var}}(\hat{t})] &= \mathbb{E}\left[\sum_{k=1}^N \sum_{l=1}^N \frac{\Delta_{k,l}}{\pi_{k,l}} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l} \mathbb{I}_{\mathcal{S}}(x_k) \mathbb{I}_{\mathcal{S}}(x_l)\right] \\ &= \sum_{k=1}^N \sum_{l=1}^N \frac{\Delta_{k,l}}{\pi_{k,l}} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l} \underbrace{\mathbb{E}[\mathbb{I}_{\mathcal{S}}(x_k) \mathbb{I}_{\mathcal{S}}(x_l)]}_* \end{aligned}$$

donde notamos que:

$$\begin{aligned} * &= \mathbb{E}[\mathbb{I}_{\mathcal{S}}(x_k) \mathbb{I}_{\mathcal{S}}(x_l)] = \text{Cov}(\mathbb{I}_{\mathcal{S}}(x_k), \mathbb{I}_{\mathcal{S}}(x_l)) + \mathbb{E}[\mathbb{I}_{\mathcal{S}}(x_k)] \mathbb{E}[\mathbb{I}_{\mathcal{S}}(x_l)] \\ &= \pi_{k,l} - \pi_k \pi_l + \pi_k \pi_l \\ &= \pi_{k,l} \end{aligned}$$

de donde se sigue:

$$\begin{aligned}\mathbb{E}[\widehat{\text{Var}}(\hat{t})] &= \sum_{k=1}^N \sum_{l=1}^N \frac{\Delta_{k,l}}{\pi_{k,l}} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l} \underbrace{\pi_{k,l}}_* \\ &= \sum_{k=1}^N \sum_{l=1}^N \Delta_{k,l} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l} = \text{Var}(\hat{t})\end{aligned}$$

Podemos calcular la varianza estimada para una muestra aleatoria simple sin reemplazo como sigue (ver ejercicio):

```
f      <- n/N
varianza <- N^2*(1 - f)/n*var(muestra)
print(varianza)
```

```
## [1] 7811.204
```

Observaciones

1. La media muestral $\bar{x}_{\mathcal{S}} = \frac{1}{n} \sum_{i=1}^n x_i$ es un estimador insesgado de la media poblacional $\bar{x}_{\mathcal{U}} = \frac{1}{N} \sum_{i=1}^N x_i$. Se sigue de una factorización de n del total (t y \hat{t} respectivamente).
2. Se puede obtener $\text{Var}(\bar{x}_{\mathcal{S}})$ y $\widehat{\text{Var}}(\bar{x}_{\mathcal{S}})$ factorizando las n de manera cuadrática del \hat{t} .

4.4.1 Ejercicio

Definimos:

$$s_{x,\mathcal{U}}^2 = \frac{1}{N-1} \sum_{k=1}^N (x_k - \bar{x}_{\mathcal{U}})^2$$

como la **varianza poblacional ajustada** y

$$s_{x,\mathcal{S}}^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x}_{\mathcal{S}})^2$$

como la **varianza muestral ajustada**. Sea $f = \frac{n}{N}$ la **fracción muestral**. Demuestra que en el caso de muestreo aleatorio simple sin reemplazo:

$$\text{Var}(\hat{t}) = N^2 \frac{1-f}{n} s_{x,\mathcal{U}}^2$$

mientras que el estimador insesgado se transforma en:

$$\widehat{\text{Var}}(\hat{t}) = N^2 \frac{1-f}{n} s_{x,\mathcal{S}}^2$$

4.5 Teorema del Límite Central (Aplicación)

En esta sección hablaremos del teorema central del límite correspondiente a muestreo aleatorio simple con poblaciones finitas. Éste no es el mismo que el de Proba 2 (en términos de hipótesis) aunque las conclusiones sean las mismas. El teorema de Proba 2 establece que si se tiene una colección $\{X_i\}$ de variables aleatorias independientes idénticamente distribuidas (todas con distribución acumulada F_X) con media μ y varianza $\sigma^2 < \infty$, entonces, si definimos Z como:

$$Z = \lim_{n \rightarrow \infty} \sqrt{\frac{n}{\sigma^2}} \cdot \left(\frac{1}{n} \sum_{i=1}^n X_i - \mu \right)$$

se tiene que $Z \sim \text{Normal}(0, 1)$.

En este teorema central podemos observar que hay algo muy parecido a la media muestral embebido en el teorema (la $\frac{1}{n} \sum_{i=1}^n X_i$) pero no es exactamente la media muestral (aquí se supone que todas las X_i son independientes con distribución F_X y en el caso de muestreo aleatorio sin reemplazo se sabe que las indicadoras **NO** son independientes y que de hecho tampoco son idénticamente distribuidas cuando analizamos $\sum_{i=1}^n x_i \mathbb{I}_{\mathcal{S}}(x_i)$). Entonces *técnicamente* no podemos aplicar el teorema central del límite así como está a nuestra muestra. Sin embargo, H  jek (y m  s tarde Rosen) encontraron condiciones *sin tener que pedir independencia ni distribuci  n id  ntica* que permiten sustituir las X_i por las de la media muestral ($x_i \mathbb{I}_{\mathcal{S}}(x_i)$) y que, cuando N y n tienden a infinito “de buena manera,” se tiene algo similar a esta expresi  n (**OJO** no es una expresi  n *correcta* pero es la idea):

$$Z = \lim_{N, n \rightarrow \infty} \sqrt{\frac{1}{\text{Var}(\bar{x}_{\mathcal{S}})}} \cdot \left(\frac{1}{n} \sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i) - \bar{x}_{\mathcal{U}} \right)$$

donde $\mu = \sum_{k=1}^N x_k$ es la media poblacional y $\sigma^2 = \frac{1}{N} \sum_{k=1}^N (x_k - \mu)^2$ la varianza poblacional no ajustada. La demostraci  n propia de este teorema la posponemos para una secci  n posterior. Por ahora, exemplificaremos el teorema del l  mite central en R, utilizaremos la expresi  n anterior para deducir y explicar el concepto de intervalo de confianza y, finalmente, haremos un ejemplo de estimaci  n de intervalo.

4.5.1 Estimaci  n de intervalos de confianza para el total

Un intervalo de confianza de $(1 - \alpha) \times 100\%$ de un estimador poblacional desconocido $\theta = \theta(x_1, x_2, \dots, x_N)$ (constante) es un intervalo aleatorio de la forma $[L(\mathcal{S}), U(\mathcal{S})]$ (donde L, U son variables aleatorias que dependen de la muestra) tal que

$$\mathbb{P}\left(\theta \in [L(\mathcal{S}), U(\mathcal{S})]\right) = 1 - \alpha$$

Notamos que lo aleatorio del intervalo son las cotas del mismo y que, dadas distintas muestras \mathcal{S} el valor de inter  s θ no siempre va a caer ah  . La idea de un

intervalo es poder dar una cota de más o menos dónde anda un valor. Veamos un ejemplo con el total.

Recordamos que el estimador del total es insesgado $\mathbb{E}[\hat{t}] = t$ y que por definición:

$$\hat{t} = N \frac{1}{n} \sum_{i=1}^N x_i \cdot \mathbb{I}_{\mathcal{S}}(x_i)$$

luego usando la versión de muestreo finito del teorema central del límite (factorizando N) tenemos que:

$$\sqrt{\frac{1}{\text{Var}(\bar{x}_{\mathcal{S}})}} \cdot \left(\frac{1}{n} \sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i) - \bar{x}_{\mathcal{U}} \right) = \cdot N \frac{\left(\frac{1}{n} \sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i) - \bar{x}_{\mathcal{U}} \right)}{N \sqrt{\text{Var}(\bar{x}_{\mathcal{S}})}} = \frac{\hat{t} - t}{\sqrt{\text{Var}(\hat{t})}} \stackrel{\sim}{\rightarrow} \text{Normal}(0, 1)$$

De donde se sigue que si se desea tener un intervalo de tamali $(1 - \alpha) \times 100\%$ lo que hay que hacer es buscar $L(\mathcal{S})$ y $U(\mathcal{S})$ tales que:

$$\mathbb{P}\left(L(\mathcal{S}) \leq \frac{\hat{t} - t}{\sqrt{\text{Var}(\hat{t})}} \leq U(\mathcal{S})\right) = 1 - \alpha$$

En este caso las probabilidades (por aproximación asintótica) se modelan bajo la hipótesis de normalidad. Y tomamos ventaja de que la normal es simétrica respecto a la media para proponer que $L(\mathcal{S}) = -U(\mathcal{S})$ y ambas correspondan a $\pm\Phi^{-1}(\alpha/2)$ (la función de distribución acumulada inversa de la normal). Es decir, ambos deben corresponder a los cuantiles con probabilidad $\alpha/2$ y $1 - \alpha/2$, denotados $z_{\alpha/2}$ y $z_{1-\alpha/2}$. Por simetría de la normal tenemos que: $z_{\alpha/2} = -z_{1-\alpha/2}$ y por tanto:

$$\mathbb{P}\left(z_{\alpha/2} \leq \frac{\hat{t} - t}{\sqrt{\text{Var}(\hat{t})}} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

de donde despejamos:

$$\mathbb{P}\left(z_{\alpha/2} \sqrt{\text{Var}(\hat{t})} \leq \hat{t} - t \leq z_{1-\alpha/2} \sqrt{\text{Var}(\hat{t})}\right) = \mathbb{P}\left(\hat{t} - z_{1-\alpha/2} \sqrt{\text{Var}(\hat{t})} \leq t \leq \hat{t} + z_{\alpha/2} \sqrt{\text{Var}(\hat{t})}\right) = 1 - \alpha$$

Notamos que como no conocemos $\text{Var}(\hat{t})$ la podemos aproximar mediante $\widehat{\text{Var}}(\hat{t})$ (hay mejores aproximaciones mediante una t de Student asintótica pero no lo usaremos ahora) y tener intervalos aproximados de la forma:

$$\begin{aligned} L(\mathcal{S}) &= \hat{t} - z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{t})} \\ U(\mathcal{S}) &= \hat{t} + z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{t})} \end{aligned} \tag{4.2}$$

de manera concisa muchas veces los escribimos como:

$$\hat{t} \pm z_{1-\alpha/2} \sqrt{\text{Var}(\hat{t})}$$

4.5.2 Ejemplo con simulación:

Veamos cómo se ven múltiples intervalos simulados con confianza del 90% y suponiendo la varianza es conocida

```

nsim <- 100
n    <- 100

total.muestra <- rep(NA, nsim)
confianza.bajo <- rep(NA, nsim)
confianza.alto <- rep(NA, nsim)
f <- n/N
z <- qnorm(1 - 0.1/2)

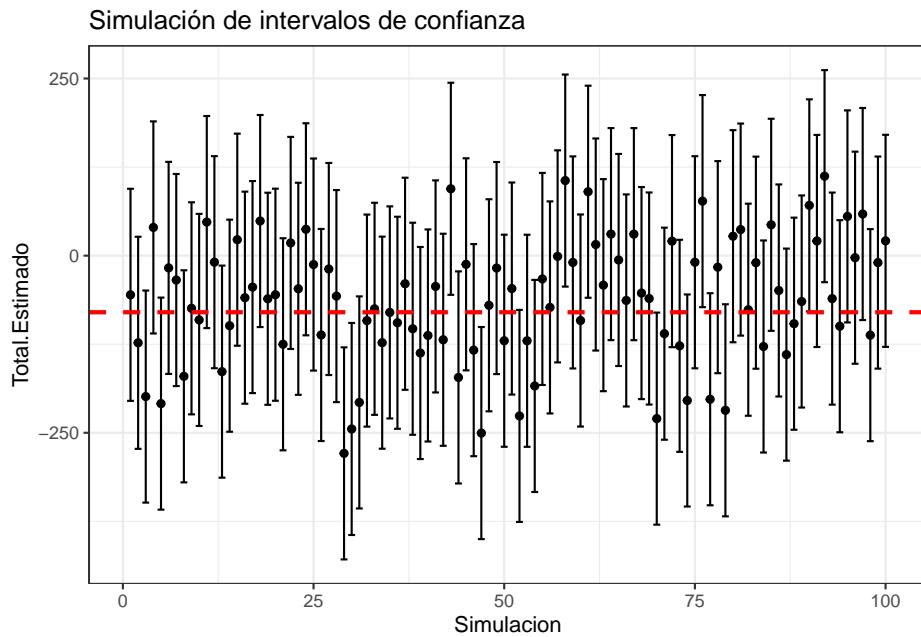
var.total      <- N^2*(1 - f)/n*var(base.completa$x)

for (i in 1:nsim){
  muestra           <- sample(base.completa$x, n, replace = FALSE)
  total.muestra[i] <- N*mean(muestra)
  #var.total[i]     <- N^2*(1 - f)/n*var(muestra)
  confianza.bajo[i] <- total.muestra[i] - z*sqrt(var.total)
  confianza.alto[i] <- total.muestra[i] + z*sqrt(var.total)
}

intervalos.simulados <- data.frame(
  Simulacion = 1:nsim,
  Intervalo.Bajo = confianza.bajo,
  Total.Estimado = total.muestra,
  Intervalo.Alto = confianza.alto
)

ggplot(intervalos.simulados) +
  geom_point(aes(x = Simulacion, y = Total.Estimado)) +
  geom_errorbar(aes(x = Simulacion, ymin = Intervalo.Bajo,
                    ymax = Intervalo.Alto)) +
  geom_hline(aes(yintercept = sum(base.completa$x)),
             linetype = "dashed",
             size = 1, color = "red") +
  theme_bw() +
  ggtitle("Simulación de intervalos de confianza")

```



Nota que estos intervalos son aproximados y no siempre van a funcionar. (¿Puedes hallar un ejemplo donde no sirvan a pesar de que n y N sean grandes?) Luego veremos correcciones a esto; por ahora, supondremos que la aproximación es buena.

4.6 Ejemplo Resumen: Estimación de una proporción bajo muestreo aleatorio simple sin reemplazo

Se realiza una encuesta mediante muestreo aleatorio simple sin reemplazo a la población del ITAM $N = 5000$ donde interesa conocer la proporción de gente que apoya al gobierno en turno p . Implícitamente, se supone que alguien apoya (proporción p de toda la población) o no lo apoya (proporción $1 - p$), que dichos conjuntos son disjuntos y que no hay una tercera opción (como NO RESPONDE / DESCONOCE QUIÉN GOBIERNA). La pregunta es: ¿a cuántas personas hay que encuestar si interesa estimar p con un error máximo de tamaño $\epsilon = 0.05$ al 99% de confianza (es decir, que el estimador \hat{p} de la proporción esté, a lo más, a ± 0.05 de distancia del valor verdadero p con un intervalo de confianza al 99%)?

Supongamos tomamos una muestra de tamaño n dada por $\mathcal{S} = (x_1, x_2, \dots, x_n)^T$ de una población $\mathcal{U} = (x_1, x_2, \dots, x_N)^T$ de tamaño N . Pensemos, además, existen N_1 personas que aprueban al gobierno actual y $N - N_1$ que desaprueban

4.6. EJEMPLO RESUMEN: ESTIMACIÓN DE UNA PROPORCIÓN BAJO MUESTREO ALEATORIO SIMPLE S

del mismo y por tanto la proporción que nos interesa estimar es:

$$p = \frac{N_1}{N}$$

Por otro lado, la proporción muestral de personas que aprueban está dada por:

$$\hat{p} = \frac{\sum_{i=1}^n \mathbb{I}_{\text{Aprueba}}(x_i)}{n}$$

donde si definimos $H = \frac{\sum_{i=1}^n \mathbb{I}_{\text{Aprueba}}(x_i)}{n}$ notamos que la distribución de H está dada por una variable Hipergeométrica (pues de una población de N se seleccionan n donde N_1 cumplen la categoría deseada). Su media y varianza están dadas respectivamente por:

$$\mathbb{E}[H] = n \frac{N_1}{N} = np$$

así como por:

$$\text{Var}[H] = n \frac{N_1}{N} \left(1 - \frac{N_1}{N}\right) \left(\frac{N-n}{N-1}\right) = np(1-p) \left(\frac{N-n}{N-1}\right)$$

Se sigue entonces que $\mathbb{E}[\hat{p}] = p$ y por tanto \hat{p} es un estimador insesgado. La varianza por otro lado es:

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n} \left(\frac{N-n}{N-1}\right)$$

Finalmente, el estimador de la varianza es:

$$\widehat{\text{Var}}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1}\right)$$

el cual también cumple que es insesgado (demuéstralos).

Podemos aplicar el Teorema Central del Límite para la proporción³ notando que la definición de \hat{p} coincide con una media (de las indicadoras):

$$\frac{\hat{p} - p}{\sqrt{\underbrace{\frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1}\right)}_{\widehat{\text{Var}}(\hat{p})}}} \sim \text{Normal}(0, 1)$$

De donde se tiene que:

$$\begin{aligned} & \mathbb{P}\left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1}\right)}} \leq z_{\alpha/2}\right) \approx 1 - \alpha \\ & \Rightarrow \mathbb{P}\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1}\right)} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1}\right)}\right) \approx 1 - \alpha \end{aligned} \tag{4.3}$$

³Una mejor distribución sería una t de Student; empero eso lo verás en Estadística Matemática.

Nota Es común encontrar en Internet que para los intervalos de confianza la gente supone una población muy grande N respecto a la muestra n y entonces eliminan el término $\frac{N-n}{N-1}$ argumentando que $\frac{N-n}{N-1} \approx 1$ y obtienen la siguiente fórmula:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

esto simplifica algunos cálculos (a mano) pero nosotros tenemos R y podemos hacer cálculos más exactos sin tener que suponer semejantes atrocidades.

Como el error deseado es de tamaño ϵ queremos $|p - \hat{p}| \leq \epsilon$ esto se traduce en:

$$|p - \hat{p}| \leq \underbrace{z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1} \right)}}_{\epsilon}$$

de donde igualamos para despejar la n :

$$\begin{aligned} \epsilon &= z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1} \right)} \\ &= \frac{\epsilon^2}{z_{\alpha/2}^2} = \frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1} \right) \\ &= \frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} = \frac{N-n}{n} = \frac{N}{n} - 1 \\ &= \frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1 = \frac{N}{n} \\ \Rightarrow n &= \frac{N}{\frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1} = \frac{\frac{z_{\alpha/2}^2}{\epsilon^2} \hat{p}(1-\hat{p})}{\frac{N-1}{N} + \frac{1}{N} \frac{z_{\alpha/2}^2}{\epsilon^2} \hat{p}(1-\hat{p})} = \frac{m}{1 + \frac{m-1}{N}} \end{aligned}$$

donde

$$m = \frac{z_{\alpha/2}^2}{\epsilon^2} \hat{p}(1-\hat{p})$$

Ahora el problema es que el tamaño de muestra n depende de la muestra a través de \hat{p} ¡y no hemos tomado la muestra! Para ello entonces analizamos el peor caso que puede ocurrir de \hat{p} de tal forma que obtengamos la n que puede salir con la

peor proporción \hat{p} posible. Para ello maximizamos con derivadas:

$$\begin{aligned}
 \frac{\partial n}{\partial \hat{p}} &= \frac{\partial}{\partial \hat{p}} \left(\frac{N}{\frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1} \right) \\
 &= N \left(\frac{1}{\frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1} \right)^2 \cdot \frac{\partial}{\partial \hat{p}} \left(\frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1 \right) \\
 &= N(N-1) \underbrace{\frac{\epsilon^2}{z_{\alpha/2}^2}}_C \left(\frac{1}{\frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1} \right)^2 \cdot \frac{\partial}{\partial \hat{p}} \left(\frac{1}{\hat{p}(1-\hat{p})} \right) \\
 &= C \left(\frac{1}{\frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1} \right)^2 \left(\frac{1}{\hat{p}(1-\hat{p})} \right)^2 \frac{\partial}{\partial \hat{p}} \hat{p}(1-\hat{p}) \\
 &= C \left(\frac{1}{\frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1} \right)^2 \left(\frac{1}{\hat{p}(1-\hat{p})} \right)^2 (1 - 2\hat{p}) = 0
 \end{aligned}$$

de donde se sigue que $\hat{p} = \frac{1}{2}$ es un punto crítico. De hecho puede verificarse que es el máximo (por ejemplo a través de la segunda derivada). Luego, podemos estimar la n de la muestra mediante:

$$n = \left\lceil \frac{m}{1 + \frac{m-1}{N}} \right\rceil$$

donde $m = \frac{1}{4} \frac{z_{\alpha/2}^2}{\epsilon^2}$. En el caso particular de este ejercicio, $N = 5000$, $\epsilon = 0.05$, $\alpha = 0.01$ y $z_{\alpha/2}^2 \approx \text{qnorm}(0.9)$. Luego podemos calcular:

```

alpha <- 0.01
z <- qnorm(1 - alpha/2)
epsilon <- 0.05
m <- (1/4)*(z/epsilon)^2
N <- 5000
n <- ceiling(m/(1 + (m-1)/N))

print(paste0("El tamaño de muestra es ", n))
## [1] "El tamaño de muestra es 586"

```

4.7 Ejemplo Resumen: Estimación del total de individuos en una fotografía

En este ejercicio vamos a determinar cuánta gente aparece en la siguiente foto:



Figure 4.1: Imagen de un concierto extraída de <https://www.youtube.com/watch?v=pJ1YKwyH5bk>

Hay varias opciones para determinar la cantidad de gente que está en dicha foto. Una sería contar todas las cabecitas que aparecen; otra, diseñar un modelo de redes neuronales (o de convolución porque a la gente le encanta eso) que identifique una cabeza y la cuente. Nosotros lo que haremos (por ser un curso de estadística) será muestrear. Como investigador me interesa responder la siguiente pregunta:

¿Cuánta gente está en la fotografía con un intervalo de error de ± 50 casos al 95%?

Para ello dividiremos la fotografía en N pedazos (a determinar), muestrearemos n de ellos y contaremos la cantidad de personas que aparecen en cada pedazo. Finalmente, generamos intervalos de confianza y de muestreo. Para ello repetimos el ejercicio anterior de despejar la n del intervalo de confianza; por el teorema del límite central tenemos:

$$\frac{\hat{t} - t}{\sqrt{\text{Var}(\hat{t})}} \sim \text{Normal}(0, 1)$$

de donde obtenemos intervalos (¡verifícalo!) de la forma:

$$\hat{t} \pm z_{1-\alpha/2} \cdot \sqrt{\text{Var}(\hat{t})}$$

Donde podemos aproximar la varianza mediante $\widehat{\text{Var}}(\hat{t}) = N^2 \frac{1-f}{n} s_{x,S}^2$ donde recordamos que $f = n/N$ y $s_{x,S}^2$ es la varianza muestral. Tomamos $\epsilon = 50$ y despejamos:

$$\begin{aligned}
\epsilon &= z_{1-\alpha/2} \cdot \sqrt{\text{Var}(\hat{t})} \\
\Rightarrow \frac{\epsilon^2}{z_{1-\alpha/2}^2} &= N^2 \frac{1-f}{n} s_{x,S}^2 \\
\Rightarrow \frac{\epsilon^2}{z_{1-\alpha/2}^2 s_{x,S}^2 N^2} &= \frac{1 - \frac{n}{N}}{n} \\
\Rightarrow \frac{\epsilon^2}{z_{1-\alpha/2}^2 s_{x,S}^2 N^2} &= \frac{1}{n} - \frac{1}{N} \\
\Rightarrow \frac{\epsilon^2}{z_{1-\alpha/2}^2 s_{x,S}^2 N^2} + \frac{1}{N} &= \frac{1}{n} \\
\Rightarrow \frac{1}{N} \left(\frac{\epsilon^2}{z_{1-\alpha/2}^2 s_{x,S}^2 N} + 1 \right) &= \frac{1}{n} \\
\Rightarrow \frac{1}{N} \left(\frac{\epsilon^2 + z_{1-\alpha/2}^2 s_{x,S}^2 N}{z_{1-\alpha/2}^2 s_{x,S}^2 N} \right) &= \frac{1}{n} \\
\Rightarrow \left(\frac{(z_{1-\alpha/2} s_{x,S} N)^2}{\epsilon^2 + z_{1-\alpha/2}^2 s_{x,S}^2 N} \right) &= n
\end{aligned}$$

El problema aquí es que la n depende de la varianza muestral $s_{x,S}^2$ (actualmente desconocida) así como de la cantidad de cuadritos originales N en los que dividimos la foto. Hay en la literatura varias técnicas que se pueden utilizar para estimar el $s_{x,S}^2$:

1. Realizar un estudio piloto (es decir un pequeño ejemplo de lo que vas a hacer en una población chica y de ahí tener la varianza). Esta es la mejor opción.
2. Buscar otros estudios similares donde se analicen objetos similares de estudio y ver sus varianzas; suponer que la de este estudio es similar. Esta es la segunda mejor opción.
3. Inventártela (sí, es una opción pero no la mejor). Vamos, ¿cuál es la probabilidad de que nadie en todo el mundo haya hecho un análisis similar al tuyo? Si realmente estás haciendo algo completamente nuevo *sin estudio piloto* pues... podrías inventarla. ¿Lo recomiendo? No; pero pasa.

En nuestro caso utilizaremos la varianza estimada de este artículo reportada en 1.02; luego $s_{x,S}^2 \approx 1.02$ para nuestro análisis.

Finalmente, como éste es sólo un ejercicio de clase tomaremos $N = 100$ (dividir la foto en 100 cuadritos). De manera profesional, de nuevo habría que ver diferencias en los resultados de las estimaciones en función de los cuadritos, o bien asignar un costo a la cantidad de cuadros. Concluimos entonces que para nuestro estudio:

$$n = \left\lceil \frac{(z_{1-\alpha/2} s_{x,\mathcal{S}} N)^2}{\epsilon^2 + z_{1-\alpha/2}^2 s_{x,\mathcal{S}}^2 N} \right\rceil = \left\lceil \frac{(1.95 \cdot \sqrt{1.02} \cdot 100)^2}{50^2 + 1.95^2 \cdot 1.02 \cdot 100} \right\rceil$$

Podemos calcular en R:

```
n <- ceiling((qnorm(0.975)*sqrt(1.02)*100)^2/(50^2 + (qnorm(0.975)^2*1.02*100)))
print(paste0("El tamaño de muestra es ", n))
```

```
## [1] "El tamaño de muestra es 14"
```

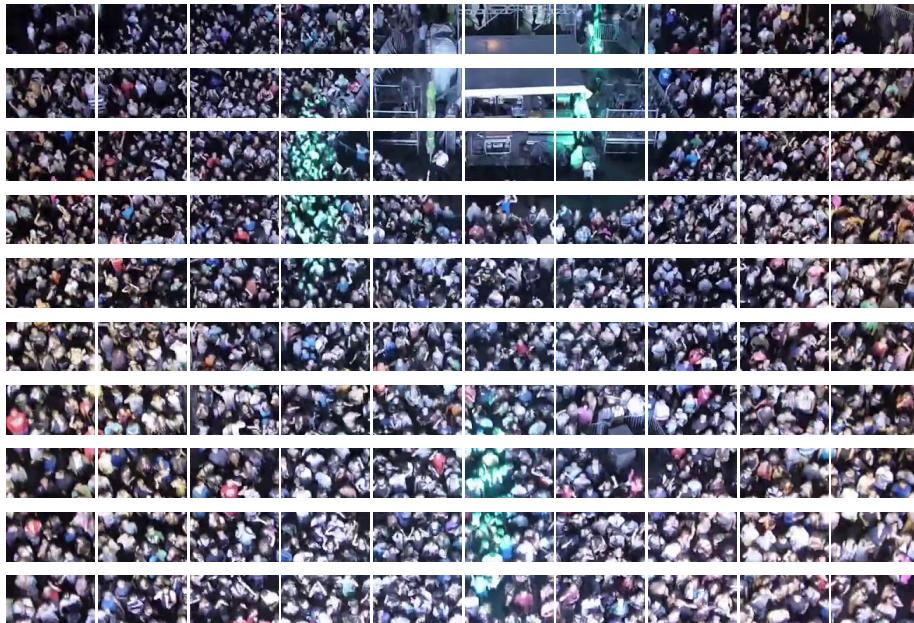
Podemos proceder a dividir la foto en los $N = 100$ pedazos:

```
#División con base en el siguiente link:
#https://rpubs.com/issactoast/cutimage
library(imager)

#Cargamos la imagen
img <- load.image("images/concierto.jpg")

#Función auxiliar del link superior
make.vr <- function( x, name ){
  assign( name, x, envir = .GlobalEnv)
}

#División en N
N <- 100
par(mfrow=c(sqrt(N),sqrt(N)), mar = c(0.1,0.1,0.1,0.1))
k <- 1
for (j in 1:sqrt(N)){
  for (i in 1:sqrt(N)){
    vr.name <- paste0("sub", k)
    k      <- k + 1
    imsub(img, (width/sqrt(N))*(i-1) < x & x < i * (width/sqrt(N)),
           (height/sqrt(N))*(j-1) < y & y < j * (height/sqrt(N))) %>%
      make.vr(name = vr.name) %>%
      # save.image( file = paste0(vr.name, ".jpg")) %>%
      plot(axes = FALSE,
            xaxt="n", yaxt="n",
            xlab = "", ylab = "", ann = FALSE )
  }
}
```



Podemos acceder a cada una de las imágenes que se tienen a través de su nombre (sub seguido de un número entre 0 y 100). Muestreamos entonces los nombres de las 15 imágenes:

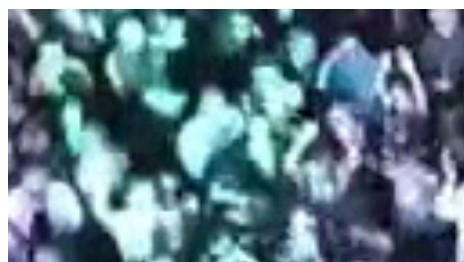
```
#Obtenemos los dígitos a muestrear
imagenes.muestreadas <- sample(1:100, n, replace = FALSE)

#Agregamos el prefijo sub
imagenes.muestreadas <- paste0("sub", imagenes.muestreadas)
```

Y graficamos cada una de ellas:

```
par(mfrow = c(1,1))

for (imagen in imagenes.muestreadas){
  plot(get(imagen), main = imagen, axes = FALSE)
}
```

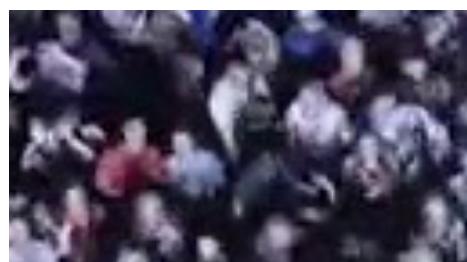
sub52**sub43****sub34****sub35**

4.7. EJEMPLO RESUMEN: ESTIMACIÓN DEL TOTAL DE INDIVIDUOS EN UNA FOTOGRAFÍA 115

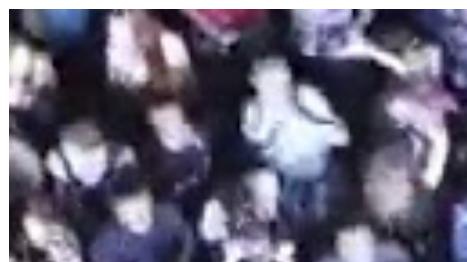
sub54



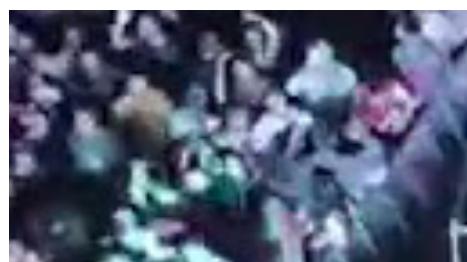
sub23



sub83



sub14



sub79



sub5



sub30



sub36

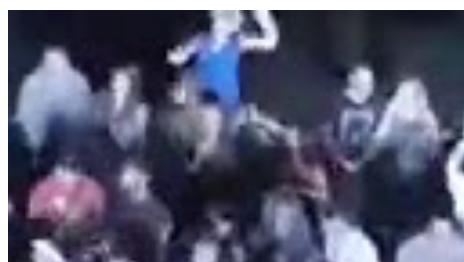
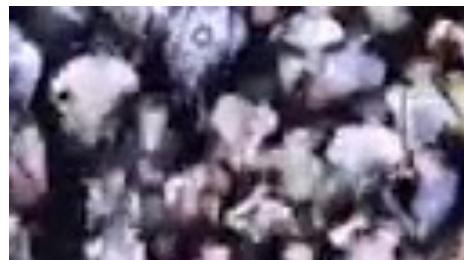


Imagen	Conteo
sub52	13
sub43	11
sub34	9
sub35	14
sub54	9
sub23	15
sub83	14
sub14	10
sub79	1
sub5	22
sub30	8
sub36	9
sub39	17
sub16	16

sub39



sub16



Para cada una de las imágenes contamos las cabecitas que aparecen:

```
datos <- data.frame(
  Imagen = imagenes.muestreadas,
  Conteo = c(13, 11, 9, 14, 9, 15, 14, 10, 1, 22, 8, 9, 17, 16)
)
kable(datos) %>% kable_styling(latex_options = "striped")
```

Tenemos entonces que la estimación del total \hat{t} es: 1200, por otro lado la varianza

muestral es $s_{x,S}$ está dada por: 25.2307692. Podemos entonces establecer un intervalo de confianza para el total:

```
x <- c(13, 11, 9, 14, 9, 15, 14, 10, 1, 22, 8, 9, 17, 16, 10)
s2           <- var(x)
N            <- 100
n            <- 15
total.muestra <- N*mean(x)
ci            <- qnorm(0.975)*sqrt(N^2*(1 - n/N)/n*s2)
ci_low        <- round(total.muestra - ci,2)
ci_up         <- round(total.muestra + ci,2)

print(paste0("Se estiman ", round(total.muestra,2), " personas con intervalo de ",
             "confianza al 95% de [", ci_low, " , ", ci_up,"]"))

## [1] "Se estiman 1186.67 personas con intervalo de confianza al 95% de [959.55 ,1413
```

4.8 Ejercicio:

Cuando se registra un paquete de R en CRAN estos se registran junto con sus autores como muestra la imagen:

La información de un paquete puede encontrarse en la página de CRAN dando clic en **Packages** y luego en **Table of available packages, sorted by name** y buscando el paquete deseado.

Se desea conocer el número promedio de autores por paquete registrado en CRAN con un intervalo de confianza al 80% y un error de ± 1 . Obtén la n necesaria para muestrear, calcula un estimador de la media y obtén intervalos de confianza. Justifica tu elección de la varianza para la n mediante un estudio piloto (muestreando de manera inicial 10 y calculando la varianza de ellos).

Hint Para obtener una lista (censo) de todos los paquetes de R puedes utilizar la función `available.packages()` la cual devuelve una matriz con todos los paquetes e incluye la `url` de donde se encuentra.

4.9 Ejemplo Resumen: Estimación de una región crítica

En una elección existen dos candidatas A y B . Se realiza una encuesta de opinión mediante muestreo aleatorio simple sin reemplazo donde se les pregunta a una cantidad suficiente de votantes por quién votarían de las dos. En este análisis no hay **NO SABE / NO RESPONDE** sino que todos los individuos indican su preferencia. Se desea determinar la cantidad de puntos porcentuales que debe haber de diferencia entre la proporción de individuos que reportan apoyan al candidato A y los que reportan que apoyan al B de tal forma que el 95% de las veces podamos declarar de manera adecuada al ganador.

Nota Si A no es el ganador entonces $p_A < 50\%$ (la proporción de votantes que van a elegir a A es menor a la mitad) ¿cierto?

Para ello el análisis es como sigue: sea \hat{p}_A un estimador de la proporción de individuos que van a elegir a A y p_A la verdadera proporción. Sin pérdida de generalidad supondremos que B es el ganador; es decir que $p_A < 0.5$. El problema puede traducirse en determinar una c tal que:

$$\mathbb{P}(\hat{p}_A > c | p_A < 0.5) \leq 0.05$$

Notamos que el evento $\{p_A < 50\%\}$ es por definición conocido (con probabilidad 0 ó 1) pues está dado por la población (constante). Notamos que por el teorema del límite central podemos escribir:

$$\frac{\hat{p}_A - p_A}{\sqrt{\text{Var}(\hat{p}_A)}} \sim \text{Normal}(0, 1)$$

donde $\hat{p}_A = \frac{1}{N} \sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i)$ como anteriormente hicimos para proporciones y su varianza está dada por:

$$\text{Var}(\hat{p}_A) = \frac{p_A(1-p_A)}{n} \left(\frac{N-1}{N-n} \right)$$

donde el cálculo se hizo en el primer ejemplo de esta sección. Podemos transformar el problema entonces en hallar c tal que:

$$\mathbb{P}\left(\underbrace{\frac{\hat{p}_A - p_A}{\sqrt{\text{Var}(\hat{p}_A)}}}_{Z \sim \text{Normal}(0,1)} > \frac{c - p_A}{\sqrt{\text{Var}(\hat{p}_A)}} \middle| p_A < 0.5 \right) \leq 0.05$$

Notamos que el lado izquierdo tiene una aproximación normal y entonces podemos reescribir el problema como hallar c tal que:

$$\mathbb{P}\left(Z > \frac{c - p_A}{\sqrt{\text{Var}(\hat{p}_A)}} \middle| p_A < 0.5 \right) \leq 0.05 \quad \text{donde } Z \sim \text{Normal}(0, 1).$$

Recordando la expresión para la varianza sustituyo:

$$\mathbb{P} \left(Z > \frac{c - p_A}{\sqrt{\frac{p_A(1-p_A)}{n} \left(\frac{N-1}{N-n} \right)}} \middle| p_A < 0.5 \right) \leq 0.05 \quad \text{donde } Z \sim \text{Normal}(0, 1).$$

En función del análisis pasado, observamos que $\frac{c - p_A}{\sqrt{\frac{p_A(1-p_A)}{n} \left(\frac{N-1}{N-n} \right)}}$ es una función decreciente en términos de p_A (¡compruébalo!) y que el mínimo valor se alcanza en el máximo de la p_A en el intervalo; es decir cuando $p_A = \frac{1}{2}$. Luego el problema se transforma en hallar c tal que:

$$\mathbb{P} \left(Z > \frac{c - \frac{1}{2}}{\sqrt{\frac{\frac{1}{2}(1-\frac{1}{2})}{n} \left(\frac{N-1}{N-n} \right)}} \right) \leq 0.05 \quad \text{donde } Z \sim \text{Normal}(0, 1).$$

donde eliminamos el evento $p_A < 0.5$ por ser un evento seguro. Reescribimos el evento:

$$\underbrace{\mathbb{P} \left(Z < \frac{c - \frac{1}{2}}{\sqrt{\frac{\frac{1}{2}(1-\frac{1}{2})}{n} \left(\frac{N-1}{N-n} \right)}} \right)}_{\Phi(x)} \geq 0.95 \quad \text{donde } x = \frac{c - \frac{1}{2}}{\sqrt{\frac{\frac{1}{2}(1-\frac{1}{2})}{n} \left(\frac{N-1}{N-n} \right)}}$$

de tal forma que descubrimos la acumulada de la normal; terminamos de escribir todo:

$$\Phi(x) \geq 0.95$$

donde aplicamos la función inversa de la acumulada de la normal para descubrir:

$$\frac{c - \frac{1}{2}}{\sqrt{\frac{\frac{1}{2}(1-\frac{1}{2})}{n} \left(\frac{N-1}{N-n} \right)}} \geq \phi^{-1}(0.95) \Rightarrow c = \frac{1}{2} + \phi^{-1}(0.95) \sqrt{\frac{\frac{1}{2}(1-\frac{1}{2})}{n} \left(\frac{N-1}{N-n} \right)}$$

de donde se sigue que:

$$\hat{p}_A > \frac{1}{2} \left(1 + \phi^{-1}(0.95) \sqrt{\frac{N-1}{n(N-n)}} \right) \Rightarrow 2\hat{p}_A = 1 + \phi^{-1}(0.95) \sqrt{\frac{N-1}{n(N-n)}}$$

Notando que los puntos porcentuales de B estimados mediante \hat{p}_B tienen la forma:

$$\hat{p}_B = 1 - \hat{p}_A$$

se tiene entonces que la diferencia entre puntos para determinar quien gana es:

$$\hat{p}_A - \hat{p}_B = 2\hat{p}_A - 1 \geq \phi^{-1}(0.95) \sqrt{\frac{N-1}{n(N-n)}}$$

El mismo análisis se seguiría bajo la hipótesis de que el perdedor es B ; por tanto se tiene que cumplir que:

$$|\hat{p}_A - \hat{p}_B| \geq \phi^{-1}(0.95) \sqrt{\frac{N-1}{n(N-n)}}$$

para poder declarar como ganador a aquél con más puntos porcentuales de manera correcta con una confianza del 95%.

4.10 Ejemplo Resumen: Estimación del total de una población

Consideremos una población de tiburones donde se desconoce el tamaño total de la población N . Algunas veces para determinar el tamaño poblacional se utiliza un modelo de *captura y recaptura*. En él se capturan ℓ individuos los cuales se identifican (mediante etiquetas, por ejemplo) y se devuelven a convivir entre la población de N para mezclarse de vuelta. Una vez mezclados, seleccionamos n nuevos individuos por muestreo aleatorio simple sin reemplazo donde descubrimos que K están marcados. Suponiendo que $K \neq 0$, determinaremos un estimador \hat{N} del total poblacional (en el caso $K = 0$ tuvimos muy mala suerte y seguimos recapturando tiburones hasta encontrar alguno).

En primer lugar notamos que los K marcados que surgen en la segunda muestra siguen una distribución hipergeométrica:

$$\mathbb{P}(K = x) = \frac{\binom{\ell}{x} \binom{N-\ell}{n-x}}{\binom{N}{n}}$$

donde $x \in [\max\{0, \ell - N + n\}, \min\{n, \ell\}] \cap \mathbb{N}$. Para construir el estimador notamos que:

$$\mathbb{E}(K) = n \frac{\ell}{N}$$

de donde podemos despejar N :

$$N = n \frac{\ell}{\mathbb{E}(K)}$$

Ahora bien, dada una muestra donde se obtuvieron K (de n) marcados se propone un estimador de N dado por:

$$\hat{N} = \ell \cdot \frac{n}{K}$$

donde $K = \sum_{i=1}^n x_i$ donde las $x_i = 1$ si estaba marcado y $x_i = 0$ si no lo estaba. La K de hecho depende de la muestra y se puede escribir como:

$$K = \sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i)$$

Para estimar si \hat{N} es insesgado, habría que calcular su valor esperado condicional en que $K > 0$. Para ello notamos que:

$$\mathbb{E}[\hat{N}|K > 0] = (\ell n) \cdot \mathbb{E}\left[\frac{1}{K}|K > 0\right]$$

Sabemos (por la desigualdad de Jensen) que $\mathbb{E}\left[\frac{1}{K}\right] \neq \frac{1}{\mathbb{E}[K]}$ por lo cual aproximamos el valor esperado mediante una expansión de Taylor; es decir para una función $f \in \mathcal{C}^2$:

$$\mathbb{E}[f(X)] \approx \mathbb{E}[f(\mu) + (X - \mu)f'(\mu) + (X - \mu)^2 f''(\mu)] = f(\mu) + \text{Var}[X] f''(\mu)$$

donde $\mu = \mathbb{E}[X]$. En nuestro caso $f(k) = \frac{1}{k}$ y por tanto:

$$\mathbb{E}\left[\frac{1}{K}|K > 0\right] \approx \frac{1}{\mathbb{E}[K|K > 0]} + 2 \cdot \frac{\text{Var}[K|K > 0]}{(\mathbb{E}[K|K > 0])^3} = \frac{1}{\mu} + 2 \frac{\sigma^2}{\mu^3}$$

Calculamos los valores esperados:

$$\mathbb{E}[K] = \underbrace{\mathbb{E}[K|K = 0]\mathbb{P}(K = 0)}_{=0} + \mathbb{E}[K|K > 0]\mathbb{P}(K > 0) \Rightarrow \mathbb{E}[K|K > 0] = \frac{\ell n}{N} \frac{1}{\mathbb{P}(K > 0)}$$

de donde se sigue que:

$$\mathbb{E}[K|K > 0] = \frac{\ell n}{N} \frac{1}{1 - \mathbb{P}(K = 0)} = \frac{\ell n}{N} \frac{1}{1 - \frac{\binom{N-\ell}{n}}{\binom{N}{n}}} = \frac{\ell n}{N} \cdot \frac{\binom{N}{n}}{\binom{N}{n} - \binom{N-\ell}{n}} = \mu$$

Por otro lado el cálculo de la varianza:

$$\begin{aligned}
 \text{Var}[K|K > 0] &= \mathbb{E}[K^2|K > 0] - \mathbb{E}[K|K > 0]^2 \\
 &= \frac{\mathbb{E}[K^2]}{\mathbb{P}(K > 0)} - \mu^2 \\
 &= \frac{\text{Var}[K] + \mathbb{E}[K]^2}{1 - \mathbb{P}(K = 0)} - \mu^2 \\
 &= \frac{\text{Var}[K] + \left(n \frac{\ell}{N}\right)^2}{1 - \mathbb{P}(K = 0)} - \mu^2 \\
 &= \frac{\frac{n\ell}{N} \cdot \frac{(N-\ell)}{N} \cdot \left(\frac{N-n}{N-1}\right) + \left(n \frac{\ell}{N}\right)^2}{1 - \mathbb{P}(K = 0)} - \mu^2 \\
 &= \frac{\frac{n\ell}{N} \cdot \frac{(N-\ell)}{N} \cdot \left(\frac{N-n}{N-1}\right) + \left(n \frac{\ell}{N}\right)^2}{1 - \frac{\binom{N-\ell}{n}}{\binom{N}{n}}} - \mu^2 \\
 &= \binom{N}{n} \frac{\frac{n\ell}{N} \cdot \frac{(N-\ell)}{N} \cdot \left(\frac{N-n}{N-1}\right) + \left(n \frac{\ell}{N}\right)^2}{\binom{N}{n} - \binom{N-\ell}{n}} - \mu^2 = \sigma^2
 \end{aligned}$$

Donde se tiene entonces que:

$$\mathbb{E}[\hat{N}|K > 0] \approx (\ell n) \left[\cdot \frac{1}{\mathbb{E}[K|K > 0]} + 2 \cdot \frac{\text{Var}[K|K > 0]}{(\mathbb{E}[K|K > 0])^3} \right]$$

con los valores estimados en los renglones anteriores. En particular, \hat{N} no es insesgado pero puede demostrarse que en el límite $\lim_{N \rightarrow \infty} \lim_{n \rightarrow \infty}$ lo es.

De manera similar puede obtenerse (ver Lohr capítulo 13):

$$\text{Var}[\hat{N}|K > 0] \approx \left(\frac{n\ell}{K}\right)^2 \frac{\ell - K}{K(\ell - 1)}$$

Misma que puede utilizarse para los intervalos de confianza.

4.11 Demostración del Teorema del Límite Central para Muestras Finitas

PRONTO

4.12 Muestreo Aleatorio Simple Bernoulli (BE)

En un esquema de muestreo Bernoulli (BE) se tiene una población de tamaño $N \in \mathbb{N}$ (constante) la cual se enlista de manera ordenada $U = (x_1, x_2, \dots, x_N)^T$.

Se recorre la lista de 1 hasta N . Cada elemento de la población, se selecciona y se mide con probabilidad $\pi \in (0, 1)$ para generar una muestra $\mathcal{S} = (x_1, x_2, \dots, x_n)^T$ de tamaño $n = n(\mathcal{S})$ aleatorio (con $0 \leq n(\mathcal{S}) \leq N$).

Un ejemplo de muestreo Bernoulli ocurre en las aduanas del Sistema de Administración Tributaria (SAT) donde con probabilidad π se revisa la mercancía de un viajero (de un total predefinido de N viajeros) para verificar no haya contrabando y con probabilidad $1 - \pi$ se le deja entrar al país sin revisar su mercancía.

Un muestreo Bernoulli no necesariamente tiene muestras del mismo tamaño: como el que cada elemento esté en la muestra depende de π entonces $n(\mathcal{S})$ es una variable aleatoria con distribución Binomial:

$$n(\mathcal{S}) \sim \text{Binomial}(N, \pi)$$

con media y varianza dadas por:

$$\mathbb{E}[n(\mathcal{S})] = N\pi \quad \text{y} \quad \text{Var}[n(\mathcal{S})] = N\pi(1 - \pi)$$

Una forma de muestrear de un muestreo Bernoulli es recorrer uno a uno los elementos de la muestra y generar una variable aleatoria $B_i \sim \text{Bernoulli}(\pi)$ de tal forma que si $B_i = 1$ se incluye el elemento en la muestra. Este esquema está programado en R como sigue:

```
datos <- data.frame(Edad = c(10, 12, 5, 4, 1, 3, 14),
                     Raza = c("Labrador", "Pomeranio", "Labrador",
                             "Pastor Alemán", "Bulldog", "Bulldog", "Chihuahua"))
datos$en_muestra <- 0
proba <- 3/4
for (i in 1:nrow(datos)){
  Bi <- sample(c(0,1), 1, prob = c(1 - proba, proba))
  datos$en_muestra[i] <- Bi
}
muestra <- datos %>% filter(en_muestra == 1)
```

Bajo este esquema se tiene que:

$$\pi_k = \mathbb{P}(x_k \in \mathcal{S}) = \pi \quad \forall k$$

Además en este caso las $\{\mathbb{I}_{\mathcal{S}}(x_k)\}_k$ son independientes y por tanto:

$$\pi_{k,l} = \pi^2$$

En caso de muestreo aleatorio Bernoulli tenemos que un estimador del total es de la misma forma que en el caso de muestreo aleatorio simple:

$$\hat{t}_\pi = \frac{1}{\pi} \sum_{i=1}^{n(\mathcal{S})} x_i$$

El cual es insesgado pues usando indicadoras reescribimos $\hat{t}_\pi = \frac{1}{\pi} \sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i)$ y tomamos valor esperado:

$$\mathbb{E}[\hat{t}_\pi] = \frac{1}{\pi} \sum_{i=1}^N x_i \mathbb{E}[\mathbb{I}_{\mathcal{S}}(x_i)] = \frac{1}{\pi} \sum_{i=1}^N x_i \pi = \sum_{i=1}^N x_i = t$$

por otro lado su varianza está dada por:

$$\text{Var}_{\text{BE}}(\hat{t}_\pi) = \left(\frac{1}{\pi} - 1\right) \sum_{i=1}^N x_i^2$$

la cual puede estimarse de manera insesgada mediante:

$$\widehat{\text{Var}}_{\text{BE}}(\hat{t}_\pi) = \frac{1}{\pi} \left(\frac{1}{\pi} - 1\right) \sum_{i=1}^{n(\mathcal{S})} x_i^2$$

4.12.1 Ejercicio

1. Demuestra la expresión para $\text{Var}_{\text{BE}}(\hat{t}_\pi)$
2. Demuestra que $\widehat{\text{Var}}_{\text{BE}}(\hat{t}_\pi)$ es un estimador insesgado de $\text{Var}_{\text{BE}}(\hat{t}_\pi)$.

4.12.2 Ejemplo

Consideraremos un ejemplo presentado por *Särndal et al.* Un profesor corrige 600 exámenes. Quiere tener un estimado de la calificación de sus alumnos y para ello cada que aparece un examen tira un dado justo de 6 caras y si sale un 6 corrige dicho examen; en caso contrario lo deja pasar. Al final del análisis el profe obtiene una muestra de 90 estudiantes de los cuales 60 pasaron. Asignamos $x_i = 0$ si un alumno no pasó y $x_i = 1$ si pasó; de esta forma la estimación de la cantidad de alumnos que pasaron es un total dado por:

$$\hat{t} = \frac{1}{\pi} \sum_{i=1}^{90} x_i = \frac{1}{\frac{1}{6}} 60 = 360$$

El profe, después de pensar un rato se le ocurre otra manera de estimar la proporción de los alumnos que pasaron. Si pasaron $60/90$ se tiene entonces que $2/3$ de los alumnos pasan; aplicando el $2/3$ a los 600 alumnos que tiene un estimador alternativo del total sería:

$$\hat{t}_{\text{Alt}} = \frac{2}{3} \cdot 600 = 400$$

El cual escrito en términos de las variables utilizadas es:

$$\hat{t}_{\text{Alt}} = \begin{cases} \frac{N}{n(\mathcal{S})} \cdot \sum_{i=1}^{n(\mathcal{S})} x_i & \text{si } n(\mathcal{S}) > 0 \\ 0 & \text{si } n(\mathcal{S}) = 0 \end{cases}$$

La pregunta obligada es ¿cuál es un mejor estimador si \hat{t} o bien \hat{t}_{Alt} ?

4.12.3 Un mejor estimador: el proporcional al tamaño

Para decidir si \hat{t}_{Alt} es un mejor estimador que \hat{t} calculemos su valor esperado y su varianza. En ambos casos tenemos dos cosas aleatorias: los elementos que sí quedaron en la muestra (las x_i) y el tamaño de muestra (la n). Para ello utilizamos la propiedad de torre de la esperanza condicional:

$$\begin{aligned}
 \mathbb{E}[\hat{t}_{\text{Alt}}] &= \mathbb{E}\left[\mathbb{E}[\hat{t}_{\text{Alt}}|n(\mathcal{S}) = k]\right] \\
 &= \sum_{k=0}^N \mathbb{E}\left[\hat{t}_{\text{Alt}}|n(\mathcal{S}) = k\right] \cdot \mathbb{P}(n(\mathcal{S}) = k) \\
 &= \sum_{k=1}^N \mathbb{E}\left[\frac{N}{n(\mathcal{S})} \cdot \sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i) | n(\mathcal{S}) = k\right] \cdot \mathbb{P}(n(\mathcal{S}) = k) \\
 &= \sum_{k=1}^N \mathbb{E}\left[\frac{N}{k} \cdot \sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i) | n(\mathcal{S}) = k\right] \cdot \mathbb{P}(n(\mathcal{S}) = k) \\
 &= \sum_{k=1}^N \frac{N}{k} \mathbb{E}\left[\sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i) | n(\mathcal{S}) = k\right] \binom{N}{k} \pi^k (1-\pi)^{N-k} \\
 &= \sum_{k=1}^N \left(\frac{N}{k} \sum_{i=1}^N x_i \mathbb{E}[\mathbb{I}_{\mathcal{S}}(x_i) | n(\mathcal{S}) = k] \right) \binom{N}{k} \pi^k (1-\pi)^{N-k} \\
 &= \sum_{k=1}^N \left(\frac{N}{k} \sum_{i=1}^N x_i \frac{k}{N} \right) \binom{N}{k} \pi^k (1-\pi)^{N-k} \\
 &= \left(\sum_{i=1}^N x_i \right) \cdot \sum_{k=1}^N \left(\binom{N}{k} \pi^k (1-\pi)^{N-k} \right) \\
 &= t \cdot (1 - (1-\pi)^N)
 \end{aligned}$$

en este caso el estimador *no* es insesgado y su sesgo es $(1-\pi)^N$. Este sesgo es prácticamente ignorable pues para aplicaciones con N grande $(1-\pi)^N \approx 0$ y no habrá mucha variación en el resultado.

Definición Dado $\hat{\theta}$ estimador de θ definimos el **sesgo** de $\hat{\theta}$ como:

$$\text{Sesgo}(\hat{\theta}) = \mathbb{E}[\hat{\theta} - \theta]$$

Podemos calcular la varianza de nuestro estimador; para ello denotamos

$$H(\pi, N) = \sum_{k=1}^N \frac{1}{k} \binom{N}{k} \pi^N (1-\pi)^{N-k} - \frac{(1 - (1-\pi)^N)}{N}$$

Luego:

$$\begin{aligned}
\text{Var}[\hat{t}_{\text{Alt}}] &= \mathbb{E}[\hat{t}_{\text{Alt}}^2] - \mathbb{E}[\hat{t}_{\text{Alt}}]^2 \\
&= \mathbb{E}[\hat{t}_{\text{Alt}}^2] - \left(t \cdot (1 - (1 - \pi)^N) \right)^2 \\
&= \mathbb{E}\left[\mathbb{E}[\hat{t}_{\text{Alt}}^2 | n(\mathcal{S}) = k] \right] - \left(t \cdot (1 - (1 - \pi)^N) \right)^2 \\
&= \sum_{k=1}^N \mathbb{E}[\hat{t}_{\text{Alt}}^2 | n(\mathcal{S}) = k] \cdot \mathbb{P}(n(\mathcal{S}) = k) - \left(t \cdot (1 - (1 - \pi)^N) \right)^2 \\
&= \sum_{k=1}^N \frac{N^2}{k^2} \mathbb{E}\left[\left(\sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i) \right)^2 \middle| n(\mathcal{S}) = k \right] \cdot \mathbb{P}(n(\mathcal{S}) = k) - \left(t \cdot (1 - (1 - \pi)^N) \right)^2
\end{aligned}$$

Notamos que:

$$\begin{aligned}
& \mathbb{E}\left[\left(\sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i)\right)^2 \middle| n(\mathcal{S}) = k\right] \\
&= \text{Var}\left[\left(\sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i)\right) \middle| n(\mathcal{S}) = k\right] + \mathbb{E}\left[\left(\sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i)\right) \middle| n(\mathcal{S}) = k\right]^2 \\
&= \sum_{i=1}^N x_i^2 \text{Var}\left[\mathbb{I}_{\mathcal{S}}(x_i) \middle| n(\mathcal{S}) = k\right] + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N x_i x_j \text{Cov}\left[\mathbb{I}_{\mathcal{S}}(x_i), \mathbb{I}_{\mathcal{S}}(x_j) \middle| n(\mathcal{S}) = k\right] \\
&\quad + \left(\sum_{i=1}^N x_i \mathbb{E}\left[\mathbb{I}_{\mathcal{S}}(x_i) \middle| n(\mathcal{S}) = k\right]\right)^2 \\
&= \sum_{i=1}^N x_i^2 \frac{k}{N} \left(1 - \frac{k}{N}\right) + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N x_i x_j \left(\frac{k(k-1)}{N(N-1)} - \frac{k^2}{N^2}\right) + \left(\frac{k}{N} \sum_{i=1}^N x_i\right)^2 \\
&= \frac{k}{N} \left[\sum_{i=1}^N x_i^2 \left(1 - \frac{k}{N}\right) + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N x_i x_j \left(\frac{k-1}{N-1} - \frac{k}{N}\right) \right] + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= \frac{k}{N} \left[\sum_{i=1}^N x_i^2 \left(\frac{N-k}{N}\right) - \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N x_i x_j \left(\frac{N-k}{N(N-1)}\right) \right] + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= \frac{k}{N}(N-k) \left[\frac{1}{N} \sum_{i=1}^N x_i^2 - \frac{1}{N} \frac{1}{N-1} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N x_i x_j \right] + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= \frac{k}{N}(N-k) \frac{1}{N-1} \sum_{i=1}^N \left[\frac{N-1}{N} x_i^2 - \frac{1}{N} x_i \sum_{\substack{j=1 \\ j \neq i}}^N x_j \right] + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= \frac{k}{N}(N-k) \frac{1}{N-1} \sum_{i=1}^N \left[x_i^2 - \frac{1}{N} x_i \sum_{j=1}^N x_j \right] + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= \frac{k}{N}(N-k) \frac{1}{N-1} \left[\sum_{i=1}^N x_i^2 - \frac{1}{N} \left(\sum_{i=1}^N x_i\right) \left(\sum_{j=1}^N x_j\right) \right] + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= \frac{k}{N}(N-k) \frac{1}{N-1} \left[\sum_{i=1}^N x_i^2 - \frac{1}{N} \left(\sum_{i=1}^N x_i\right)^2 \right] + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= \frac{k}{N}(N-k) \frac{1}{N-1} \sum_{i=1}^N \left(x_i - \frac{1}{N} \sum_{j=1}^N x_j\right)^2 + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= k \frac{(N-k)}{N} \frac{1}{N-1} \sum_{i=1}^N \left(x_i - \bar{x}_{\mathcal{U}}\right)^2 + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= k \frac{(N-k)}{N} s_{\mathcal{U}}^2 + k^2 \bar{x}_{\mathcal{U}}^2
\end{aligned}$$

por lo cual si sustituimos en la ecuación anterior: %

$$\begin{aligned}
\text{Var}[\hat{t}_{\text{Alt}}] &= \sum_{k=1}^N \frac{N^2}{k^2} \mathbb{E}\left[\left(\sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i)\right)^2 \middle| n(\mathcal{S}) = k\right] \cdot \mathbb{P}(n(\mathcal{S}) = k) - \left(t \cdot (1 - (1 - \pi)^N)\right)^2 \\
&= \sum_{k=1}^N \frac{N^2}{k^2} \left[k \frac{(N-k)}{N} s_{\mathcal{U}}^2 + k^2 \bar{x}_{\mathcal{U}}^2\right] \cdot \binom{N}{k} \pi^k (1 - \pi)^{N-k} - \left(t \cdot (1 - (1 - \pi)^N)\right)^2 \\
&= N^2 \sum_{k=1}^N \left[\frac{(N-k)}{kN} s_{\mathcal{U}}^2 + \bar{x}_{\mathcal{U}}^2\right] \cdot \binom{N}{k} \pi^k (1 - \pi)^{N-k} - t^2 \cdot (1 - (1 - \pi)^{2N}) \\
&= N^2 s_{\mathcal{U}}^2 \sum_{k=1}^N \left(\frac{1}{k} - \frac{1}{N}\right) \binom{N}{k} \pi^k (1 - \pi)^{N-k} + N^2 \bar{x}_{\mathcal{U}}^2 \sum_{k=1}^N \binom{N}{k} \pi^k (1 - \pi)^{N-k} \\
&\quad - N^2 \bar{x}_{\mathcal{U}}^2 (1 - (1 - \pi)^{2N}) \\
&= N^2 s_{\mathcal{U}}^2 \sum_{k=1}^N \frac{1}{k} \binom{N}{k} \pi^k (1 - \pi)^{N-k} - \frac{1}{N} (1 - (1 - p)^N) + \\
&\quad (1 - (1 - p)^N) N^2 \bar{x}_{\mathcal{U}}^2 (1 - (1 - (1 - p)^N)) \\
&= N^2 [H(N, \pi) s_{\mathcal{U}}^2 + (1 - p)^N (1 - (1 - p)^N) \bar{x}_{\mathcal{U}}^2]
\end{aligned}$$

En nuestro caso para elegir el mejor estimador entre \hat{t} y \hat{t}_{alt} se calculan las varianzas de ambos. Una posible elección es aquél que tiene menos varianza (podría estar más cercano al valor dado que el sesgo de \hat{t}_{alt} es pequeñísimo⁴). Se puede demostrar (ver Särndal) que en general

$$\text{Var}[\hat{t}_{\text{Alt}}] \ll \text{Var}[\hat{t}]$$

Y usualmente se prefiere el estimador \hat{t}_{Alt} .

4.13 Ejemplo Resumen: Aduana

Se sabe que de manera diaria fluyen por un punto de la aduana 1000 cargamentos. Cada cargamento que entra debe ser analizado para buscar contrabando con probabilidad p (y con probabilidad $(1 - p)$ se deja pasar sin mayor análisis). Determina la probabilidad p si se desea estimar el total de cargamentos con contrabando que pasan por la aduana y, a la vez, se busca que el 75% de las ocasiones no se analicen más de 200 cargamentos.

Para encontrar la probabilidad p (correspondiente al π) recordamos que el tamaño de la muestra n tiene una distribución Binomial:

$$n(\mathcal{S}) \sim \text{Binomial}(1000, p)$$

⁴Calcúlalo.

Buscamos entonces una p tal que

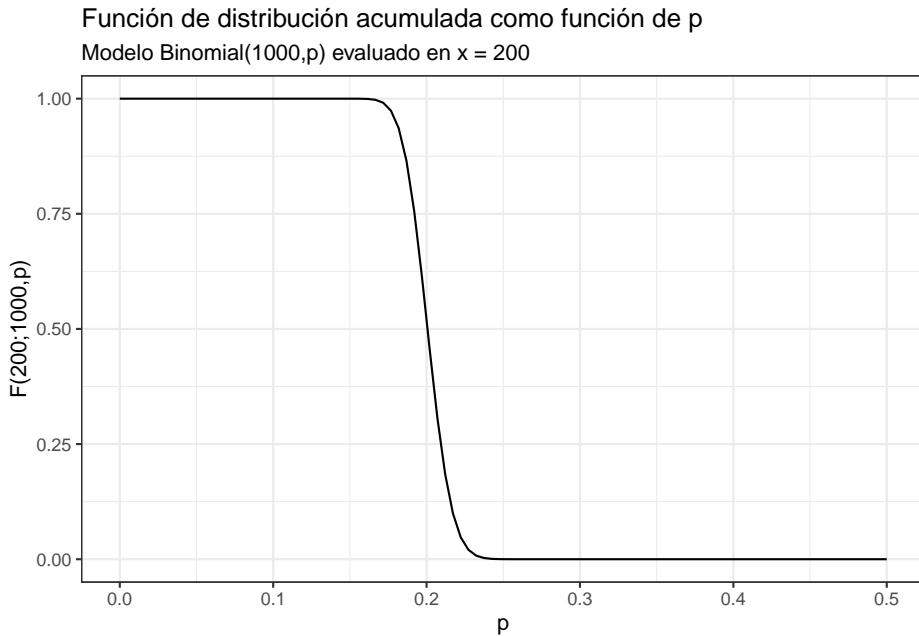
$$\mathbb{P}(B \leq 200) = 0.75 \quad \text{donde } B \sim \text{Binomial}(1000, p)$$

En particular, notamos que del lado izquierdo tenemos a la función de distribución acumulada $F_B(200) = \mathbb{P}(B \leq 200)$ la cual depende (de manera implícita) de p . ¡Hagamos explícita la dependencia de los parámetros p y N :

$$F_B(200; 1000, p) = 0.75 \quad \text{donde } B \sim \text{Binomial}(1000, p)$$

Podemos graficar la función de distribución acumulada como función de p :

```
p.val <- seq(0, 0.5, length.out = 100)
ggplot() +
  geom_line(aes(x = p.val, y = pbinom(200, 1000, p.val))) +
  theme_bw() +
  labs(
    x = "p",
    y = "F(200;1000,p)",
    title = "Función de distribución acumulada como función de p",
    subtitle = "Modelo Binomial(1000,p) evaluado en x = 200"
  )
```



Notamos entonces que lo que necesitamos es hallar la p donde la función de distribución acumulada (como función de p) toca al 0.75. Para ello, como no podemos despejar, utilizamos un método numérico a través de `uniroot` para encontrar el 0 de la función $g(p) = F_B(200; 1000, p) - 0.75$ (pues la p^* tal que $g(p^*) = 0$ es la respuesta):

```

g.fun <- function(p){pbinom(200, 1000, p) - 0.75}
raiz <- uniroot(g.fun, lower = 0, upper = 0.5)
print(paste0("El valor de p es ", raiz$root))

## [1] "El valor de p es 0.192159774829166"

```

De donde obtenemos el p necesario.

4.14 Muestreo Aleatorio Simple con Reemplazo (MAS/cR)

El muestreo aleatorio simple con reemplazo es idéntico al muestreo aleatorio sin reemplazo *pero* en este caso no se extrae un elemento de la muestra sino que se permite que se seleccione múltiples veces. En cada selección hay una probabilidad $1/N$ de que un individuo de la población sea seleccionado. Cada selección es independiente de la pasada. Aquí consideraremos un universo de tamaño constante $N \in \mathbb{R}$ dado por $U = (x_1, x_2, \dots, x_N)^T$ y las variables aleatorias N_k que denotan la cantidad de veces que x_k fue seleccionado para incluirse en la muestra⁵. El orden en el que fueron seleccionados los elementos no importa.

En el caso de muestreo aleatorio simple con reemplazo se fija un tamaño de muestra m y hay por tanto N^m muestras posibles. Cada una de las muestras sigue la siguiente función de probabilidad uniforme:

$$\mathbb{P}(\mathcal{S} = S) = \begin{cases} \frac{1}{N^m} & \text{si } \#S = m \\ 0 & \text{en otro caso} \end{cases}$$

Dado un elemento x_k la probabilidad de que dicho x_k aparezca r veces en la muestra de tamaño m está dada por:

$$\binom{m}{r} \left(\frac{1}{N}\right)^r \left(1 - \frac{1}{N}\right)^{m-r}$$

En particular se tiene que la probabilidad de que x_k no esté en la muestra es:

$$\left(1 - \frac{1}{N}\right)^m$$

o bien de que esté en la muestra:

$$\pi_k = 1 - \left(1 - \frac{1}{N}\right)^m$$

lo cual se calcula por el complemento. Por otro lado, la probabilidad conjunta $\pi_{k,l}$ de que x_k y x_l estén en la muestra se puede computar usando inclusión exclusión:

⁵Observa que las variables aleatorias N_k generalizan a las variables indicadoras.

$$\pi_{k,l} = 1 - \underbrace{\left(1 - \frac{1}{N}\right)^m}_{\text{No está } x_k} - \underbrace{\left(1 - \frac{1}{N}\right)^m}_{\text{No está } x_l} + \underbrace{\left(1 - \frac{2}{N}\right)^m}_{\text{No está ni } x_k \text{ ni } x_l}$$

En R puedes obtener un muestreo aleatorio simple con reemplazo cambiando en `sample` el `replace`:

```
sample(c("A", "B", "C"), 10, replace = TRUE)
```

```
## [1] "B" "A" "C" "A" "C" "B" "B" "A" "A" "C"
```

Una observación es bastante relevante aquí:

$$\begin{aligned} \pi_k &= 1 - \left(1 - \frac{1}{N}\right)^m = 1 - \sum_{j=0}^m \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j} \\ &= 1 - \left[\sum_{j=0}^{m-2} \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j} - \frac{m}{N} + 1 \right] \\ &= \frac{m}{N} - \sum_{j=0}^{m-2} \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j} \\ &= \frac{m}{N} + \mathcal{O}\left(\frac{m^2}{N^2}\right) \end{aligned}$$

donde $\mathcal{O}\left(\frac{m^2}{N^2}\right)$ es notación que implica que una función $f(n)$ es de orden $g(n)$ (denotado $f(n) = \mathcal{O}(g(n))$) si y sólo si existe M tal que para cualquier $n \in \mathbb{N}$ tal que $|f(n)|/g(n) \leq M$. Escrito con palabras en este caso esto significa que si m/N es pequeño entonces $\frac{m^2}{N^2}$ es casi 0 y entonces muestrear con reemplazo es casi lo mismo que muestrear sin reemplazo (lo cual tiene sentido: si tu población es muy grande $N \gg 0$ entonces está bien difícil que vuelvas a capturar a uno en tu encuesta y por tanto es casi lo mismo muestrear **con** que **sin** reemplazo en términos prácticos).

Para el análisis del muestreo aleatorio simple con reemplazo podemos generalizar la idea de variables indicadoras. Como en este tipo de muestreo pueden aparecer **varias veces** los mismos valores x_i utilizaremos unas variables $\mathbb{A}_{\mathcal{S}}(x_i)$ para denotar cuántas veces aparece el valor x_i en la muestra aleatoria \mathcal{S} ; es decir:

$$\mathbb{A}_{\mathcal{S}}(x_i) = \begin{cases} 0 & \text{si } x_i \notin \mathcal{S} \\ k & \text{si } x_i \in \mathcal{S} \quad k \text{ veces} \end{cases}$$

Observamos que la distribución de las a_i es multinomial:

$$\begin{aligned}
 \mathbb{P}(\mathbb{A}_{\mathcal{S}}(x_1) = a_1, \mathbb{A}_{\mathcal{S}}(x_2) = a_2, \dots, \mathbb{A}_{\mathcal{S}}(x_N) = a_N) &= \\
 &= \binom{m}{a_1} \left(\frac{1}{N}\right)^{a_1} \cdot \binom{m-a_1}{a_2} \left(\frac{1}{N}\right)^{a_2} \cdot \binom{m-a_1-a_2}{a_3} \left(\frac{1}{N}\right)^{a_3} \cdots \left(\frac{1}{N}\right)^{a_N} \\
 &= \frac{m!}{a_1! a_2! \cdots a_N!} \frac{1}{(m - \sum_{l=1}^N a_l)!} \left(\frac{1}{N}\right)^{\sum_{l=1}^N a_l} \\
 &= \frac{m!}{a_1! a_2! \cdots a_N!} \left(\frac{1}{N}\right)^m
 \end{aligned} \tag{4.4}$$

donde tomamos que $\sum_{l=1}^N a_l = m$. Al ser una distribución multinomial se tienen las marginales:

$$\mathbb{E}[\mathbb{A}_{\mathcal{S}}(x_i)] = \frac{m}{N}$$

y por otro lado:

$$\text{Var}[\mathbb{A}_{\mathcal{S}}(x_i)] = \frac{m(N-1)}{N^2}$$

con:

$$\text{Cov}[\mathbb{A}_{\mathcal{S}}(x_i), \mathbb{A}_{\mathcal{S}}(x_j)] = -\frac{m}{N^2}$$

Un estimador de la media es la suma de los N valores únicos en el universo (*i.e.* $N = \#\mathcal{U}$) multiplicados por la cantidad de veces que aparecen en la muestra (las $\mathbb{A}_{\mathcal{S}}(x_i)$):

$$\bar{x}_{\mathcal{S}} = \frac{1}{m} \sum_{i=1}^m x_i = \frac{1}{m} \sum_{i=1}^N x_i \cdot \mathbb{A}_{\mathcal{S}}(x_i)$$

en este caso se tiene que el estimador es insesgado:

$$\mathbb{E}[\bar{x}_{\mathcal{S}}] = \frac{1}{m} \sum_{i=1}^N x_i \cdot \mathbb{E}[\mathbb{A}_{\mathcal{S}}(x_i)] = \frac{1}{m} \sum_{i=1}^N x_i \frac{m}{N} = \bar{x}_{\mathcal{U}}$$

Su varianza está dada por lo siguiente:

$$\begin{aligned}
\text{Var}[\bar{x}_{\mathcal{S}}] &= \frac{1}{m^2} \left(\sum_{i=1}^N x_i^2 \text{Var}[\mathbb{A}_{\mathcal{S}}(x_i)] + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N \text{Cov}[\mathbb{A}_{\mathcal{S}}(x_i), \mathbb{A}_{\mathcal{S}}(x_j)] x_i x_k \right) \\
&= \frac{1}{m^2} \left(\frac{m(N-1)}{N^2} \sum_{i=1}^N x_i^2 - \frac{m}{N^2} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N x_i x_k \right) \\
&= \frac{N-1}{mN} \left(\frac{1}{N} \sum_{i=1}^N x_i^2 - \frac{1}{N} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N x_i x_k \right) \\
&= \frac{N-1}{mN} \left(\frac{1}{N-1} \sum_{i=1}^N \left[\frac{N-1}{N} x_i^2 - \frac{1}{N} x_i \sum_{\substack{j=1 \\ j \neq i}}^N x_k \right] \right) \\
&= \frac{N-1}{mN} \left(\frac{1}{N-1} \sum_{i=1}^N \left[x_i^2 - \frac{1}{N} x_i \sum_{j=1}^N x_k \right] \right) \\
&= \frac{N-1}{mN} \left(\frac{1}{N-1} \left[\sum_{i=1}^N x_i^2 - \frac{1}{N} \sum_{i=1}^N x_i \sum_{j=1}^N x_k \right] \right) \\
&= \frac{N-1}{mN} \left(\frac{1}{N-1} \left[\sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i \right)^2 \right] \right) \\
&= \frac{N-1}{mN} s_{\mathcal{U}}^2
\end{aligned}$$

Donde la última igualdad se sigue de la misma que se hizo con Bernoulli. Un estimador de la varianza es:

$$\widehat{\text{Var}}[\bar{x}_{\mathcal{S}}] = \frac{N-1}{mN} s_{\mathcal{S}}^2$$

4.15 Ejemplo Resumen: Proporción de trabajadores enfermos con o sin reemplazo

Nos interesa estimar la proporción de trabajadores P afectados por una enfermedad en su trabajo en un negocio que emplea a 1500 personas. Además sabemos que en población general 3 de cada 10 personas enferman. Para ello obtenemos una muestra aleatoria con reemplazo donde además buscamos un intervalo de confianza al 0.95 con un error a lo más de 0.01.

Proponemos el estimador de la proporción dado por:

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

donde $x_i = 1$ si el trabajador tiene la enfermedad (0 en otro caso). En este caso tenemos que la varianza está dada por:

$$\widehat{\text{Var}}(\bar{x}_m) = \frac{N-1}{N \cdot m} s_{\mathcal{U}}^2$$

Tenemos entonces que el error es:

$$0.01 \geq \epsilon = Z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\bar{x}_m)}$$

de donde se tiene:

$$\frac{0.01^2}{Z_{1-\alpha/2}^2} \geq \frac{N-1}{N \cdot m} s_{\mathcal{U}}^2$$

o de manera equivalente:

$$\frac{0.01^2}{Z_{1-\alpha/2}^2 \cdot s_{\mathcal{U}}^2} \frac{N}{N-1} \geq \frac{1}{m}$$

de donde se sigue que:

$$\left(\frac{Z_{1-\alpha/2} s_{\mathcal{U}}}{0.01} \right)^2 \cdot \frac{N-1}{N} \leq m$$

donde finalmente como sabemos que 3 de cada 10 personas lo padecen notamos que la probabilidad en el mundo real de obtener a una persona que lo tenga es $p = \frac{1}{3}$ y podemos modelar la variable `estar enfermo` mediante una Bernoulli y por tanto una buena aproximación a $s_{\mathcal{U}}^2$ es:

$$s_{\mathcal{U}}^2 \approx \underbrace{\frac{1}{3} \left(1 - \frac{1}{3} \right)}_{p \cdot (1-p)}$$

donde sustituimos:

$$\frac{1}{3} \left(1 - \frac{1}{3} \right) \cdot \left(\frac{Z_{1-\alpha/2}}{0.01} \right)^2 \cdot \frac{N-1}{N} \leq m$$

Concluimos que:

```
m <- ceiling(1/3*(1 - 1/3)*(qnorm(0.975)/0.01)^2*(1500 - 1)/1500)
print(paste0("m = ", m))
## [1] "m = 8531"
```

Lo cual es un sinsentido estadístico: ¿para qué muestrear más de lo que se tiene en población total? En este ejercicio lo mejor sería hacer un censo.

Nota Si se repite el ejercicio con muestreo aleatorio simple sin reemplazo acabamos con $n \approx 1300$ lo cual sí tiene sentido como muestreo. En general la m de muestreo con reemplazo es mayor que la n de sin reemplazo (las varianzas son mayores). Intuitivamente esto tiene sentido pues si estás muestrando con la posibilidad de repetidos a fuerza necesitas muestrear más para obtener la misma cantidad de datos únicos.

4.16 Ejemplo Resumen: Captura-Recaptura con reemplazo

Se realiza un estudio para determinar la cantidad de ratas en la CDMX. Para ello se pone una trampa en algún lugar aleatorio de la ciudad. Si se atrapa una rata se le marca y se le deja ir. Si para 50 ratas capturadas, contamos 42 marcadas determina el número de ratas en la isla suponiendo que las 50 fueron con reemplazo.

Para ello denotamos $p_N(r)$ a la probabilidad de tener r ratas distintas en m intentos con reemplazos ($m = 50$ es determinado por nosotros y no es aleatorio) en una población de tamaño desconocido N . Una vez que se fijan las r ratas que van a salir hay $\binom{N}{r}$ formas de elegirlas. Entonces:

$$p_N(r) = \frac{N!}{r!(N-r)!} q_N(r)$$

donde $q_N(r)$ es la probabilidad de obtener r distintas ratas en m intentos con reemplazo. Fijado el número de ratas, el universo Ω de posibilidades se forma por el grupo de mapeos de $\{1, 2, \dots, m\} \rightarrow \{1, 2, \dots, N\}$ (todas las formas de haber acomodado las ratas). Tenemos $m \geq r$ y de hecho:

$$q_N(r) = \sum_{\omega \in \text{Fav}} p(\omega)$$

dpmde $p(\omega)$ es la probabilidad de obtener un mapeo favorable ($\omega \in \text{Fav}$). Tenemos que $p(\omega) = N^{-m}$ para toda ω . La cantidad de casos favorables es lo mismo que preguntarse por la cantidad de mapeos suprayectivos del conjunto $\{1, 2, \dots, m\}$ en $\{1, 2, \dots, r\}$ lo cual está dado por $r!$ multiplicado por el número de Stirling de segundo tipo $s_m^{(r)}$:

$$s_m^{(r)} = \frac{1}{r!} \sum_{i=1}^r \binom{r}{i} i^m (-1)^{r-i}$$

donde $s_m^{(r)}$ es la forma de encontrar de un grupo de m elementos r partes no vacías. Tenemos entonces:

$$p_N(r) = \frac{N!}{(N-r)!N^m} s_m^{(r)} \quad \text{parar} = 1, 2, \dots, \min\{m, N\}.$$

Lo que vamos a hacer es pensar que la N que generó los datos es la N máxima (*criterio de máxima verosimilitud*) y entonces lo que hay que maximizar es:

$$\frac{N!}{(N-r)!N^m} = \frac{\prod_{i=0}^{r-1}(N-i)}{N^m}$$

Hay dos formas de hacer esta maximización: enlistando todas las N y r para mi población o bien derivando el logaritmo:

$$\frac{d}{dN} \ln \left(\frac{\prod_{i=0}^{r-1}(N-i)}{N^m} \right) = \frac{d}{dN} \left[\sum_{i=0}^{r-1} \ln(N-i) - m \ln(N) \right] = \sum_{i=0}^{r-1} \frac{1}{N-i} - \frac{m}{N} = 0$$

de donde se sigue que:

$$\sum_{i=0}^{r-1} \frac{N}{N-i} = m$$

La cual es una ecuación no lineal que se puede resolver mediante `uniroot` como la pasada.

```
m.val <- function(N){
  r <- 42
  m <- 50
  suma <- 0
  N    <- floor(N) #Esto es para garantizar sale un entero; no es la mejor opción de optimizar
  for (i in 1:r){
    suma <- suma + N/(N - (i - 1))
  }
  return(suma - m)
}
floor(uniroot(m.val, lower = 100, upper = 200)$root)

## [1] 136
```

4.17 Muestreo Aleatorio Simple Ponderado (MAS/P)

En el caso más general posible cada uno de los elementos x_k tiene una probabilidad π_k de aparecer en la muestra. Análogamente se tienen probabilidades conjuntas de la forma $\pi_{k,l}$ donde no necesariamente hay independencia. El estimador del total está dado por el estimador Horvitz Thompson:

$$\hat{t} = \sum_{k=1}^n \frac{x_k}{\pi_k}$$

donde su varianza y sus estimadores fueron ya calculados desde muestreo aleatorio simple. En el siguiente capítulo comenzaremos a variar mucho más las π_k cuando entremos a muestreo estratificado. ¡Nos vemos pronto!

Edad	Peso
15	50
30	70
5	20
25	90
10	35
45	85

4.18 Ejercicios

1. Bajo muestreo aleatorio Bernoulli se propone el siguiente estimador para el total:

$$\hat{t}_{\text{BE}} = \frac{N}{\mathbb{E}[n(\mathcal{S})]} \sum_{i=1}^{n(\mathcal{S})} x_i$$

- a. Demuestra que es insesgado
 - b. Obtén su varianza
 - c. Para el ejemplo del profesor con los exámenes ¿es \hat{t}_{BE} una mejor opción que \hat{t}_{Alt} ? Justifica calculando en R todos los posibles casos (desde que 0 pasan hasta que los 600 pasan el examen) y analizando cuántas de esas veces la varianza de \hat{t}_{Alt} es menor que la de \hat{t}_{BE} .
2. Se tiene una muestra aleatoria simple con reemplazo y se calcula la varianza del estimador de la media como sigue:

$$\text{Var}(\bar{x}_{\mathcal{S}}) = \frac{1}{n} \sum_{i=1}^N x_i^2 \cdot \text{Var}(\mathbb{I}_{\mathcal{S}}(x_i)) = \frac{1}{n} \sum_{i=1}^N x_i^2 \cdot \frac{1}{N} \left(1 - \frac{1}{N}\right) = \frac{1}{N \cdot n} \sum_{i=1}^N x_i^2 \cdot \underbrace{\left(1 - \frac{1}{N}\right)}_{(N-1)/N} = \frac{N-1}{N \cdot n} s_{\mathcal{U}}^2$$

El resultado está bien pero el razonamiento tiene un error. Identifica los errores.

3. Sea \mathcal{S} una muestra bajo diseño Bernoulli con parámetro π . Sea $n(\mathcal{S}) = \#\mathcal{S}$ la variable aleatoria del tamaño de \mathcal{S} . Demuestra que condicional en que $n(\mathcal{S}) = n$ la probabilidad de que $\mathcal{S} = S$ es la misma que bajo muestreo aleatorio simple.
4. Encuentra un estimador insesgado de la varianza poblacional bajo muestreo aleatorio simple sin reemplazo. *Hint* Demuestra que:

$$\frac{1}{N} \sum_{k=1}^N (x_k - \bar{x}_{\mathcal{U}})^2 = \frac{1}{2N^2} \sum_{k=1}^N \sum_{\substack{l=1 \\ l \neq k}}^N (x_k - \bar{x}_l)^2$$

5. Considera la población de la tabla siguiente:

Se sabe que en una muestra aleatoria sin reemplazo que se obtenga de dicha población es dos veces más factible seleccionar a alguien de menos de 20 años que a alguien de más de 20 años. a. Determina un estimador insesgado de la media de peso para dicha población. b. Se obtiene la muestra aleatoria simple de peso dada por $S = \{50, 35, 85\}$. Determina la probabilidad de haber obtenido dicha muestra. c. A partir de la muestra S del inciso anterior, estima el peso total poblacional (la suma de los pesos) y da un intervalo asintótico de confianza de 75% para dicho total.

6. De una población de $N = 1000$ personas se entrevistaron a $n = 100$. Se les preguntó el tipo de música que preferían; 30 personas respondieron *reguetón*.
 - a. Estima el total de personas de la población que prefieren *reguetón*.
 - b. Genera un intervalo de confianza asintótico de 50% para dicha estimación.
7. Un estimador $\hat{\theta}$ de θ es Fisher consistente si cuando $n = N$ el estimador $\hat{\theta} = \theta$.
 - a. Demuestra que \bar{x}_S es Fisher consistente.
 - b. Da un ejemplo de un estimador que no sea Fisher consistente.
8. ¿Cuál de los siguientes diseños de muestreo simple sin reemplazo tiene la mayor precisión para estimar la media poblacional? Suponiendo que todas las poblaciones tienen la misma varianza σ^2 :
 - a. Tomar $n = 400$ de $N = 4000$
 - b. Tomar $n = 300$ de $N = 4000$
 - c. Tomar $n = 3000$ de $N = 300,000,000$
9. En una ciudad hay 10 millones de habitantes. Interesa saber el porcentaje de personas que utilizan bicicleta en dicha ciudad con un margen de error de 4 puntos porcentuales y una confianza del 75%. Determina el tamaño de muestra mínimo para hacer una encuesta de usuarios de bicicleta.
10. Bajo **muestreo Bernoulli**, define el total de la población como $t = \sum_{i=1}^N x_i$. Un estimador del total es:

$$\hat{t}_1 = \frac{1}{p} \sum_{k=1}^N x_k \mathbb{I}_S(x_k)$$

- a. Obtén $\mathbb{E}[\hat{t}_1]$
- b. Demuestra que $\text{Var}[\hat{t}_1] = (\frac{1}{p} - 1) \sum_{k=1}^N x_k^2$.
- c. Obtén un estimador de $\text{Var}[\hat{t}_1]$. ¿Es insesgado?
10. Un estimador distinto del total de la población (bajo Bernoulli) está dado por:

$$\hat{t}_2 = N\bar{x}$$

donde $\bar{x} = \frac{1}{\#S} \sum_{k=1}^N x_k \mathbb{I}_k$.

- Demuestra que $\mathbb{E}[\hat{t}_2] = (1 - (1-p)^N) \cdot t$. Hint Utiliza la propiedad de torre de proba 2: $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$ condicionando en el tamaño de la muestra.
- Demuestra que $\text{Var}[\hat{t}_2] = \frac{N^3}{N-1} [H\sigma^2 + (1-p)^N (1 - (1-p)^N) \bar{x}^2]$ donde $H = \sum_{k=1}^N \frac{1}{k} \binom{N}{k} p^k (1-p)^{N-k} - (1 - (1-p)^N)/N$.
- En muestreo aleatorio simple con reemplazo sea $n_{\text{únicos}}$ la cardinalidad del conjunto $\{x_k | x_k \in S\}$; es decir, el número de valores únicos que se obtuvieron en dicho muestreo. Determina $\mathbb{E}[n_{\text{únicos}}]$
- En un esquema de muestreo Poisson (PO) se tiene una población de tamaño $N \in \mathbb{N}$ (constante) la cual se enlista de manera ordenada $U = \{x_1, x_2, \dots, x_N\}$. Se recorre la lista de 1 hasta N . Cada elemento de la población, se selecciona y se mide con probabilidad $\pi_k \in (0, 1)$ para $k = 1, 2, \dots, N$ a fin de generar una muestra $S = \{x_1, x_2, \dots, x_n\}$ de tamaño $\#S$ (con $0 \leq \#S \leq N$) donde en este caso la cardinalidad de S , $\#S$, es una variable aleatoria. En este caso se asume que \mathbb{I}_k es independiente de \mathbb{I}_j para $k \neq j$. Responde los siguientes incisos considerando que la muestra S fue generada por un esquema Poisson.

- Demuestra que $\mathbb{E}[\mathbb{I}_k] = \pi_k$. ¿Cuánto vale es $\text{Cov}(\mathbb{I}_k, \mathbb{I}_j)$?
- Demuestra $\mathbb{E}[\#S] = \sum_{k=1}^N \pi_k$ y $\text{Var}[\#S] = \sum_{k=1}^N \pi_k(1 - \pi_k)$.
- Define la media de la población como $\mu_X = \frac{1}{N} \sum_{i=1}^N x_i$. Se propone un estimador de la media como:

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n \frac{x_k}{\pi_k}$$

- Determina el sesgo de $\hat{\mu}$
- ¿Es $\hat{\mu}$ Fisher-consistente?
- Obtén el error cuadrático medio de $\hat{\mu}$
- Demuestra que $\widehat{\text{Var}}[\hat{\mu}] = \frac{1}{N^2} \sum_{i \in S} (1 - \pi_i) \pi_i^{-2} x_i^2$ es estimador insesgado de la varianza $\text{Var}[\hat{\mu}]$.
- Una aplicación de el muestreo Poisson es en estimación de cantidad de madera en un bosque. Los investigadores van al bosque y estiman a ojo de buen cubero el tamaño de un árbol en una de las categorías: *pequeño*, *mediano*, *grande*. Una vez que a ojo se estimaron los tamaños se seleccionan los árboles pequeños con probabilidad p_1 para cada árbol, los medianos con probabilidad p_2 y los grandes con probabilidad p_3 ($0 < p_1 < p_2 < p_3 < 1$). La selección de cada árbol es independiente de que otro haya sido seleccionado. Se busca estimar el total de madera midiendo los árboles (área de la base por altura). Supongamos que un bosque tiene N árboles de los cuales N_1 son pequeños, N_2 son medianos y N_3 son grandes ($\sum N_i = N$). Supongamos además que medir un árbol pequeño cuesta C_1 , un árbol

mediano cuesta C_2 y un árbol grande cuesta C_3 ($0 < C_1 < C_2 < C_3$). Específicamente, supongamos que $p_i = C_i / \sum_k C_k$. En promedio para $N_1 = N_2 = 300$ y $N_3 = 400$ y $C_1 = 10$, $C_2 = 20$ y $C_3 = 70$ ¿es más barato un muestreo Poisson (parámetros p_i) o un muestreo aleatorio simple sin reemplazo de los árboles del bosque de tamaño $n = \lceil \sum_k n_k \pi_k \rceil$?

13. Supongamos que se tienen dos muestras aleatorias simples sin reemplazo \mathcal{S} para \mathcal{U} y \mathcal{S}_- para $\mathcal{U} \setminus \mathcal{S}$ (de tamaños n y n_-). Encuentra la covarianza entre $\bar{x}_{\mathcal{S}}$ y $\bar{x}_{\mathcal{S}_-}$.
14. Queremos estimar el área cultivada por granjas en una localidad. De $N = 2010$ granjas seleccionamos 100 mediante muestreo aleatorio simple sin reemplazo. Medimos x_k el área cultivada en hectáreas y obtenemos que:

$$\sum_{i=1}^n x_k = 2907 \quad \text{y} \quad \sum_{i=1}^n x_k^2 = 154593$$

Estima el promedio de hectáreas cultivadas y da un intervalo al 90%.

15. Determina el tamaño de muestra bajo muestreo aleatorio simple para hallar con una precisión de al menos dos puntos porcentuales y un intervalo al 95% la proporción de personas que usan lentes en una población de tamaño N
16. De una población de 4000 individuos nos interesan dos proporciones: P_1 la proporción de individuos con lavadora y P_2 la proporción de individuos con laptop. Se sabe además de un estudio previo que que:

$$45\% \leq P_1 \leq 65\% \quad \text{y} \quad 5\% \leq P_2 \leq 10\%$$

De qué tamaño tiene que ser la muestra si queremos conocer *a la vez* P_1 con una precisión $\pm 2\%$ y P_2 con un error de 1% y una confianza de 95%

17. Nos interesa conocer el precio por litro en una población de 10 gasolineras. Los precios para mayo y junio de las mismas aparecen en la tabla 2. En particular, queremos estimar la evolución del precio por litro entre dichos meses. Para ello se proponen dos métodos
 - a. Muestrear n estaciones en mayo y n en junio de manera independiente y calcular la diferencia en precios.
 - b. Muestrear n estaciones en mayo e ir a verlas en junio de nuevo (a las mismas) y calcular la diferencia en precios.

Determina cuál de los métodos es mejor (en términos de varianza)

18. En una población de tamaño N consideramos n individuos mediante muestreo aleatorio simple sin reemplazo. Consideremos D una subpoblación de tamaño N_D de donde n_D es el tamaño de los muestreados que son elementos de D . La muestra la podemos dividir en dos partes: $\mathcal{S} =$

Mes	Gas1	Gas2	Gas3	Gas4	Gas5	Gas6	Gas7	Gas8	Gas9	Gas10
Mayo	5.82	5.33	5.76	5.98	6.2	5.89	5.68	5.55	5.69	5.81
Junio	5.89	5.34	5.92	6.05	6.2	6	5.79	5.63	5.78	5.84

$(\mathcal{S}_D, \mathcal{S}_{\bar{D}})^T$ según se esté o no en D . Encuentra la distribución de \mathcal{S}_D condicional en n_D :

19. Dada una población $\mathcal{U} = \{1, 2, 3\}$ se tiene el diseño:

$$p(\{1, 2\}) = 1/2, \quad p(\{1, 3\}) = 1/4, \quad p(\{2, 3\}) = 1/4$$

determina los π_k los $\pi_{k,l}$ y los $\Delta_{k,l}$

Chapter 5

Muestreo Aleatorio Estratificado

5.1 Introducción a Muestreo Aleatorio Estratificado (MAE)

Vamos a considerar una población \mathcal{U} la cual suponemos podemos particionar en una cantidad finita (no vacía) de subpoblaciones: $\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_H$ de tamaños N_1, N_2, \dots, N_H con $\bigcup_{h=1}^H \mathcal{U}_h = \mathcal{U}$ (lo que se traduce en $\sum_{h=1}^H N_h = N$). Lo que busca el muestreo aleatorio estratificado es *estimar* en cada uno de los estratos así como de manera global. Por ejemplo, puede interesarnos conocer la estatura en hombres y mujeres, las ganancias en empresas agrícolas, ganaderas, de servicios y de transformación, la cantidad de enfermos que hay en cada entidad federativa, etc. En cada uno de estos casos estamos hablando de estratos de en los cuales interesa realizar la estimación. Un punto importante aquí es que *los estratos son conocidos y decididos por la investigadora*. La extracción en cada uno de los estratos se realiza de manera independiente obteniéndose muestras $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_H$ de tamaños n_1, n_2, \dots, n_H para cada uno de ellos.

Notación Si $x_i \in \mathcal{U}_h$ lo denotaremos como $x_{i,h}$ para de esta manera distinguir el x_i que está en \mathcal{U}_h del que está en \mathcal{U}_k

La muestra total es:

$$\mathcal{S} = (\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_H)^T$$

un vector de tamaño $n = \sum_{h=1}^H n_h$. Por independencia, se tiene:

$$\mathbb{P}(\mathcal{S} = S) = \mathbb{P}_1(\mathcal{S}_1 = S_1) \cdot \mathbb{P}_2(\mathcal{S}_2 = S_2) \cdots \mathbb{P}_H(\mathcal{S}_H = S_H)$$

donde cada \mathbb{P}_h es un esquema muestral para el estrato h . En el caso del muestreo

aleatorio estratificado tenemos un estimador de la media dada por la media ponderada de las medias:

$$\bar{x}_{\mathcal{S}} = \sum_{h=1}^H \frac{N_h}{N} \cdot \bar{x}_{\mathcal{S}_h}$$

donde por comodidad denotaremos

$$\bar{x}_h = \bar{x}_{\mathcal{S}_h}$$

En particular, por la independencia se tiene:

$$\text{Var}(\bar{x}_{\mathcal{S}}) = \sum_{h=1}^H \frac{N_h^2}{N^2} \cdot \text{Var}(\bar{x}_h)$$

donde un estimador de la varianza es:

$$\widehat{\text{Var}}(\bar{x}_{\mathcal{S}}) = \sum_{h=1}^H \frac{N_h^2}{N^2} \cdot \widehat{\text{Var}}(\bar{x}_h)$$

En particular tenemos que para muestreo aleatorio simple tenemos un estimador insesgado:

$$\mathbb{E}[\bar{x}_{\mathcal{S}}] = \sum_{h=1}^H \frac{N_h}{N} \mathbb{E}[\bar{x}_h] = \sum_{h=1}^H \frac{N_h}{N} \cdot \frac{1}{N_h} \sum_{i=1}^{N_h} x_{i,h} = \frac{1}{N} \sum_{h=1}^H \sum_{i=1}^{N_h} x_{i,h} = \bar{x}_{\mathcal{U}}$$

Su varianza es:

$$\text{Var}(\bar{x}_{\mathcal{S}}) = \sum_{h=1}^H N_h^2 \frac{1-f_h}{n_h} s_{\mathcal{U}_h}^2$$

donde $f_h = n_h/N_h$ es la fracción de muestreo del estrato h y:

$$s_{\mathcal{U}_h}^2 = \frac{1}{N_h - 1} \sum_{\mathcal{U}_h} (x_k - \bar{x}_{\mathcal{U}_h})^2$$

es la varianza del estrato con

$$\bar{x}_{\mathcal{U}_h} = \sum_{\mathcal{U}_h} x_k$$

siendo la media del mismo. El estimador insesgado de la varianza en este caso es:

$$\widehat{\text{Var}}(\bar{x}_{\mathcal{S}}) = \sum_{h=1}^H N_h^2 \frac{1-f_h}{n_h} s_{\mathcal{S}_h}^2$$

donde

$$s_{\mathcal{S}_h}^2 = \frac{1}{n_h - 1} \sum_{\mathcal{S}_h} (x_k - \bar{x}_{\mathcal{S}_h})^2$$

es la varianza muestral ajustada.

5.1.1 Ejemplo

Dada la población $\mathcal{U} = \{x_1, x_2, x_3, x_4\}$ con $x_1 = x_2 = 0, x_3 = 1, x_4 = -1$:

1. Calcula la varianza del estimador de la media de un muestreo aleatorio simple sin reemplazo de tamaño 2
2. Calcula la varianza del estimador de la media de un muestreo estratificado de donde se selecciona una unidad por cada estrato y los estratos están dados por $U_1 = \{x_1, x_2\}$ y $U_2 = \{x_3, x_4\}$.

Solución Tenemos que:

$$\bar{x} = 0$$

mientras que por otro lado,

$$s_{\mathcal{S}}^2 = \frac{1}{4-1} (1^2 + (-1)^2) = \frac{2}{3}$$

Finalmente:

$$\text{Var}(\bar{x}) = \frac{N-n}{N} \frac{s_{\mathcal{S}}^2}{n} = \frac{1}{6}$$

Por otro lado para resolver 2:

$$\bar{x}_1 = \bar{x}_2 = 0$$

Además de que:

$$s_{\mathcal{S}_1}^2 = 0$$

y:

$$s_{\mathcal{S}_2}^2 = 2$$

Tenemos entonces que:

$$\text{Var}(\bar{x}) = \frac{N-n}{nN} \sum_{h=1}^2 \frac{N_h}{N} s_{\mathcal{S}_h}^2 = \frac{1}{4}$$

Notamos que la varianza del muestreo estratificado es mayor a la varianza del muestreo simple. Por lo que concluimos que **estratificar no necesariamente reduce la varianza**.

	Bajo	Medio	Alto
\$N_h\$	3500.00	2000.00	2e+03
\$n_h\$	500.00	300.00	2e+02
\$p_h\$	0.13	0.45	5e-01

5.1.2 Ejercicio sugerido

De entre 7500 empleados de una compañía deseamos conocer la proporción P que tiene un vehículo por lo menos. Se construyeron 3 estratos para la población según el ingreso (bajo, medio, alto). Se tiene N_h el total del estrato, n_h el total muestreado y p_h el estimador del total de vehículos para cada estrato $h = 1, 2, 3$ según la muestra.

Determina un estimador \hat{p} y su intervalo de confianza.

5.2 Alocación

Una pregunta importante para el caso de muestreo estratificado es el cálculo de la(s) n . En este caso ¿cómo determinar cuánto muestrear de cada población? Veamos un ejemplo:

Supongamos que se desean muestrear hombres y mujeres en México.
En este país el 48% de los habitantes son hombres y el 52% son mujeres.

Una opción en este caso podría ser tomar una muestra que refleje exactamente esas proporciones. Ésta se conoce como *proporcional al tamaño*.

5.2.1 Alocación proporcional al tamaño

Dada una población de tamaño N con H estratos de tamaños N_1, N_2, \dots, N_H para $h = 1, 2, \dots, H$ la alocación proporcional consiste en tomar n_h como:

$$n_h = n \cdot \frac{N_h}{N}$$

Ésta forma de asignar variables no necesariamente es la mejor (mucho menos para estudios con costo) por lo cual se tienen otras alocaciones.

Un alocación proporcional al tamaño representa usualmente una ganancia en la precisión (ver último ejemplo, el de los doctores)

5.2.2 Alocación óptima

Si consideramos muestreo aleatorio simple sin reemplazo y analizamos su varianza podemos reescribirla:

$$V = \text{Var}(\bar{x}_S) = \sum_{h=1}^H N_h^2 \frac{1-f_h}{n_h} s_{U_h}^2 = \sum_{h=1}^H \frac{A_h}{n_h} + B$$

donde

$$A_h = N_h^2 s_{\mathcal{U}_h}^2$$

y

$$B = - \sum_{h=1}^H N_h s_{\mathcal{U}_h}^2$$

(Ésta no es la única que se puede escribir de esta forma, también la de la Bernoulli, por ejemplo). Supongamos, además que asociado a muestrear cada estrato h hay un costo diferenciado c_h para cada elemento muestreado de h . El costo total sería:

$$C = c_0 + \sum_{h=1}^H n_h c_h$$

El problema de alocación de muestras es determinar las n_h que minimizan las varianzas V sujetas a los costos C (o puede verse de igual forma como hallar aquellas n_h que dadas varianzas predefinidas V minimizan los costos C).

Teorema Bajo un muestreo aleatorio estratificado donde V puede escribirse como:

$$V = \sum_{h=1}^H \frac{A_h}{n_h} + B$$

y con una función de costo lineal $C = c_0 + \sum_{h=1}^H n_h c_h$ la muestra óptima se alcanza tomando n_h proporcional a $(A_h/c_h)^{1/2}$.

Demostración Sea $V^* = V - B$ y $C^* = C - c_0$. El problema de optimización se puede reescribir como minimizar el producto:

$$V^* C^* = \left(\sum_{h=1}^H \frac{A_h}{n_h} \right) \cdot \left(\sum_{h=1}^H n_h c_h \right)$$

Utilizamos la desigualdad de Cauchy

$$\left(\sum_h a_h^2 \right) \left(\sum_h b_h^2 \right) \geq \left(\sum_h a_h b_h \right)^2$$

con $a_h = (A_h/n_h)^{1/2}$ y $b_h = (n_h c_h)^{1/2}$ tenemos:

$$V^* C^* \geq \left[\sum_{h=1}^H (A_h c_h)^{1/2} \right]^2$$

Recordamos que la igualdad en el caso de Cauchy se mantiene cuando b_h/a_h es constante para toda h :

$$\left(\frac{n_h c_h}{A_h/n_h} \right)^{1/2} = \text{Constante}$$

De donde se sigue el resultado que $n_h \propto (A_h/c_h)^{1/2}$

Ventas en millones	Cantidad de negocios
0 a 1	1000
1 a 10	100
Más de 10	10

Nota Minimizar la varianza para un costo fijo C nos lleva a :

$$n_h = \frac{(C - c_0)(A_h/c_h)^{1/2}}{\sum_{h=1}^H (A_h c_h)^{1/2}}$$

en particular para muestreo aleatorio simple sin reemplazo puede demostrarse:

$$n_h = \frac{(C - c_0)N_h s_{S_h}/c_h^{1/2}}{\sum_{h=1}^H N_h s_{S_h} c_h^{1/2}}$$

Por otro lado, minimizar el costo para una varianza fija V nos lleva a:

$$n_h = \left(\frac{A_h}{c_h} \right)^{1/2} \left[\sum_{h=1}^H (A_h c_h)^{1/2} \right] / (V - B)$$

Nota 2 Cuando se toman todas las c_h idénticas y constantes se le conoce como alocación de Neymann o sólo alocación óptima.

5.2.3 Ejemplo

Se quiere estimar las ventas promedio de una población de empresas. Las empresas se enlistan según tres clases: según sus ventas en la siguiente tabla:

```
tibble(`Ventas en millones` = c("0 a 1", "1 a 10", "Más de 10"),
      `Cantidad de negocios` = c(1000, 100, 10)) %>%
  kable(booktabs = T) %>% kable_styling()
```

Se sabe que se quieren estimar 111 empresas. Se supone, además que dentro de cada clase la distribución de ventas es uniforme. Obtén las varianzas de los estimadores cuando se toma alocación proporcional y cuando se toma óptima con costos constantes (Neyman).

Solución Como la distribución intra-clase es uniforme podemos completar la tabla recordando que la varianza de una variable uniforme es:

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

de donde obtenemos la tabla actualizada:

Ventas en millones	Cantidad de negocios	Varianza
0 a 1	1000	1/12
1 a 10	100	81/12
Más de 10	10	8100/12

Table 5.1: ACTUALIZACIÓN Tabla de datos de empresas

de donde se sigue que (para convertir a $1/N - 1$ de $1/N$):

$$\begin{aligned}s_{h_1}^2 &= \frac{1}{12} \cdot \frac{1000}{999} \approx 0.0834168 \\ s_{h_2}^2 &= \frac{81}{12} \cdot \frac{100}{99} \approx 0.81818 \\ s_{h_3}^2 &= \frac{8100}{12} \cdot \frac{10}{9} \approx 750\end{aligned}$$

Luego:

Estratificación proporcional al tamaño

$$\text{Var}(\bar{x}) = \frac{N-n}{nN} \sum_{h=1}^3 \frac{N_h}{N} s_h^2 \approx 0.0604$$

Estratificación óptima Por un lado tenemos que:

$$\begin{aligned}N_1 s_1^2 &= 288.82 \\ N_2 s_2^2 &= 261.116 \\ N_3 s_3^2 &= 273.861\end{aligned}$$

Lo que nos da las alocaciones óptimas:

$$\begin{aligned}n_1 &= \frac{nN_1 s_1}{\sum_{h=1}^3 N_h s_h} = 38.9161 \\ n_2 &= \frac{nN_2 s_2}{\sum_{h=1}^3 N_h s_h} = 35.18 \\ n_3 &= \frac{nN_3 s_3}{\sum_{h=1}^3 N_h s_h} = 36.90\end{aligned}$$

En el caso del tercer estrato $n_3 > N_3$ por lo que seleccionamos $n_3 = 10$. En este caso es necesario redistribuir de manera óptima los restantes 101 entre los estratos 1 y 2 por lo que recalculamos las n :

$$\begin{aligned}n_1 &= \frac{101N_1 s_1}{N_1 s_1 + N_2 s_2} = 53.0439 \\ n_2 &= \frac{101N_2 s_2}{N_1 s_1 + N_2 s_2} = 47.9561\end{aligned}$$

La distribución óptima entonces es $n_1 = 53, n_2 = 48, n_3 = 10$. Finalmente la varianza está dada por:

$$\text{Var}(\bar{x}_S) = \sum_{h=1}^3 \frac{N_h^2}{N^2} \frac{1-f_h}{n_h} s_h^2 = 0.0018$$

5.3 Ejercicio de clase:

En una ciudad grande se estudia el número promedio de pacientes que ven los médicos en su día laboral. Comenzamos con algunas hipótesis *a priori*: entre más experiencia tiene un médico más pacientes ve. Esto nos lleva a clasificar a la población de médicos en tres grupos: **recién graduados** (grupo 1), **intermedios** (grupo 2) y **experimentados** (grupo 3). Tenemos una lista de 500 doctores en el grupo 1, 1000 en el grupo 2 y 2500 en el grupo 3. Seleccionamos mediante muestreo aleatorio simple sin reemplazo 200 doctores por cada clase y calculamos el número de pacientes por día y por doctor: 10 para el grupo 1, 15 para el grupo 2 y 20 para el grupo 3. Finalmente calculamos las varianzas del número de pacientes por doctor en cada una de las siguientes muestras y encontramos respectivamente que son 4 (grupo 1), 7 (grupo 2) y 10 (grupo 3).

1. Estima la media del número de pacientes que ve un doctor por día y obtén un intervalo de confianza.
2. Si al año siguiente se volviera a repetir el mismo análisis con 600 médicos (de nuevo) una hipótesis usual es que las varianzas no cambian. Determina la alocación de Neyman y la proporcional en este caso.
3. Determina la ganancia en precisión de hacer alocación proporcional al tamaño por encima de hacer muestreo aleatorio simple

Solución 1. Consideramos el estimador de la media:

$$\bar{x}_S = \sum_{h=1}^3 \frac{N_h}{N} \bar{x}_h = 17.5$$

Por otro lado, su varianza está estimada por:

$$\widehat{\text{Var}}(\bar{x}_S) = \sum_{h=1}^3 \left(\frac{N_h}{N} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \frac{s_h^2}{n_h} \approx 0.0199$$

De donde tenemos el intervalo de confianza:

$$\bar{Y} \pm 1.95 \sqrt{\widehat{\text{Var}}} \Rightarrow \text{IC}_{95\%} = [17.5 - 0.28, 17.5 + 0.28]$$

2. Una alocación proporcional al tamaño nos lleva a que $n_h = N_h/N$ en este caso, $n_1 = 75, n_2 = 150, n_3 = 375$. Por otro lado, si utilizamos para la de Neyman las varianzas del actual y suponemos serán similares el próximoa no podemos estimar: $N_1 s_1 = 1000, N_2 s_2 = 2646$ y $N_3 s_3 = 7906$. En este caso, $n_1 = 52, n_2 = 137$ y $n_3 = 411$.

3. A partir de la fórmula de descomposición de la varianza (demuestra) podemos aproximar la varianza poblacional a partir de las de la muestra:

$$s_{\mathcal{U}}^2 \approx \sum_{h=1}^3 \frac{N_h}{N} s_{\mathcal{U}_h}^2 + \sum_{h=1}^3 \frac{N_h}{N} (\bar{x}_{\mathcal{U}_h} - \bar{x}_{\mathcal{U}})^2$$

Sabemos que $\mathbb{E}[s_{\mathcal{S}_h}^2] = s_{\mathcal{U}_h}^2$ (el estimador es insesgado en cada estrato). Nos interesa el valor esperado de:

$$A = \sum_{h=1}^3 \frac{N_h}{N} (\bar{x}_{\mathcal{S}_h} - \bar{x}_{\mathcal{S}})^2 = \sum_{h=1}^3 \frac{N_h}{N} \bar{x}_{\mathcal{S}_h}^2 - \bar{x}_{\mathcal{S}}^2$$

Luego:

$$\begin{aligned} \mathbb{E}[A] &= \sum_{h=1}^3 \frac{N_h}{N} \mathbb{E}[\bar{x}_{\mathcal{S}_h}^2] - \mathbb{E}[\bar{x}_{\mathcal{S}}^2] \\ &= \sum_{h=1}^3 \frac{N_h}{N} (\text{Var}(\bar{x}_{\mathcal{S}_h}^2) + \bar{x}_{\mathcal{S}_h}^2) - (\text{Var}(\bar{x}_{\mathcal{S}}^2) + \bar{x}_{\mathcal{S}}^2) \\ &= \sum_{h=1}^3 \frac{N_h}{N} (\bar{x}_{\mathcal{U}_h} - \bar{x}_{\mathcal{U}})^2 + \sum_{h=1}^3 \frac{N_h}{N} \text{Var}(\bar{x}_{\mathcal{S}_h}) - \text{Var}(\bar{x}_{\mathcal{S}}) \end{aligned}$$

En nuestro caso

$$\widehat{\text{Var}}(\bar{x}_{\mathcal{S}_h}) = \left(1 - \frac{n_h}{N_h}\right) \frac{s_{\mathcal{S}_h}^2}{n_h}$$

es un estimador de $\text{Var}(\bar{x}_{\mathcal{S}_h})$. Si juntamos todo tenemos un estimador insesgado de $s_{\mathcal{U}}^2$ dado por:

$$\hat{S}_{\mathcal{U}}^2 = \sum_{h=1}^3 \frac{N_h}{N} s_{\mathcal{S}_h}^2 + \sum_{h=1}^3 \frac{N_h}{N} (\bar{x}_{\mathcal{S}_h} - \bar{x}_{\mathcal{S}})^2 - \sum_{h=1}^3 \frac{N_h}{N} \widehat{\text{Var}}(\bar{x}_{\mathcal{S}_h}) + \widehat{\text{Var}}(\bar{x}_{\mathcal{S}}) = 20.983$$

La varianza estimada entonces con muestreo aleatorio simple sin reemplazo es:

$$\hat{V}_{\text{MAS}} = \frac{1-f}{n} \hat{S}_{\mathcal{U}}^2$$

Mientras que la estimada con alocación proporcional:

$$\hat{V}_{\text{Prop}} = \frac{1-f}{n} \left(\sum_{h=1}^3 \frac{N_h}{N} s_{\mathcal{S}_h}^2 \right)$$

de donde la ganancia de la proporcional está dada por:

$$\frac{\hat{V}_{\text{Prop}}}{\hat{V}_{\text{MAS}}} = \frac{\sum_{h=1}^3 \frac{N_h}{N} s_{\mathcal{S}_h}^2}{\hat{S}_{\mathcal{U}}^2} \approx 40.5\%$$

Lo cual nos muestra que, en este caso, estratificar sí resulta en estimaciones más precisas.

Edad	Varianza
< 20	100
[20,60]	200
>60	500

5.4 Ejercicio en R tipo control

La base de datos `Base_a_estratificar` (en este link) contiene una base con un millón de entradas correspondientes a los registros de un millón de clientes de una empresa. Se registró el grupo de edad, la entidad federativa y el género de la persona. Interesa realizar un muestreo aleatorio simple para estudiar el ingreso promedio de los clientes estratificando por grupo de edad, entidad y género. Se sabe además que los costos de muestreo por cada persona muestreada varían según el estado y puedes encontrarlos en la base `Costos_x_entidad` este link.

1. Determina las n óptimas para el muestreo suponiendo que la varianza sólo varía por edad de acuerdo a la siguiente tabla (varianza son las s^2) pensando, además que nos interesa un error de ± 50 al 95%.
2. Suponiendo un costo basal de 500,000, ¿cuánto es el costo total de la encuesta?
3. La base de datos `Muestra` link contiene una muestra estratificada por muestreo aleatorio simple sin reemplazo de los datos. Obtén el estimador del ingreso promedio y su intervalo de confianza para el total y para cada uno de los estratos.

Solución

1. En primer lugar calculamos N_h de cada estrato así como el N :

```
base.datos <- read_rds("Base_a_estratificar.RDS")
base.nh     <- base.datos %>% group_by(Género, Entidad, Edad) %>% tally()
```

Por otro lado calculamos la varianza a partir del error usando que

$$\epsilon = Z_{1-\alpha/2} \sqrt{\text{Var}(\bar{x}_S)} \quad \text{con} \quad \alpha = 0.05$$

Entonces:

```
eps.error <- 50
zalpha    <- qnorm(0.975)
var.x     <- (eps.error/zalpha)^2
print(paste0("La varianza es ", var.x))
```

```
## [1] "La varianza es 650.794429067515"
```

De la ecuación calculamos el término $V^* = V - B$. Para ello recordamos que:

$$B = - \sum_{h=1}^H N_h s_{U_h}^2$$

Entonces:

```
edad      <- c("< 20", "[20,60]", ">60")
dats      <- data.frame(Edad = edad, Varianza = c(100, 200, 500))
base.nh   <- base.nh %>% left_join(dats, by = "Edad")
base.nh   <- base.nh %>% mutate(BSumandos = Varianza*n)
B         <- -sum(base.nh$BSumandos)
print(paste0("B = ", B))

## [1] "B = -21616500"
```

Por otro lado, obtenemos los costos:

```
base.costos <- read_rds("Costos_x_entidad.RDS")
base.nh     <- base.nh %>% left_join(base.costos, by = "Entidad")
```

Y calculamos los A_h :

```
base.nh <- base.nh %>% mutate(Ah = Varianza*n^2)
```

Finalmente obtenemos los n_h :

```
sumaAh  <- sum(sqrt(base.nh$Ah*base.nh$Costo))
base.nh <- base.nh %>% mutate(nh = sqrt(Ah/Costo)*sumaAh/(var.x - B))
base.nh <- base.nh %>% mutate(nh = ceiling(nh))
```

Verificamos que no haya ningún $n_h > N_h$:

```
base.nh <- base.nh %>% mutate(nh = ifelse(nh > n, n, nh))
```

2. Para determinar el costo total de la encuesta,

```
base.nh <- base.nh %>% mutate(Costo_estrato = Costo*nh)
costo   <- 500000 + sum(base.nh$Costo_estrato)
print(paste0("El costo es de $", scales::comma(costo)))
```

```
## [1] "El costo es de $985,409"
```

3. Analizamos la base de datos muestra:

```
muestra <- read_rds("Muestra_estratificada.RDS")
```

Obtenemos los estimadores puntuales de cada uno

```
promedios.muestra <- muestra %>% group_by(Género, Entidad, Edad) %>%
  summarise(Media = mean(Ingreso), S_h = var(Ingreso), n = n())
```

Agrego los N_h y los N :

```
promedios.muestra <- promedios.muestra %>%
  left_join(base.nh, by = c("Género", "Entidad", "Edad")) %>%
  rename(`N_mayusc_h` = n.y) %>% rename(`n_minusc_h` = n.x)
```

```
Ntotal <- sum(promedios.muestra$N_mayusc_h)
```

El estimador total es el promedio ponderado de los de cada grupo:

```
promedios.muestra <- promedios.muestra %>%
  mutate(factor_pop = N_mayusc_h/!Ntotal)
promedios.muestra <- promedios.muestra %>%
  mutate(sumando_media = factor_pop*Media)
xbarra <- sum(promedios.muestra$sumando_media)
print(paste0("La media se estima con ", xbarra))
```

```
## [1] "La media se estima con 1193.65573405234"
```

Mientras que la varianza se estima mediante:

```
promedios.muestra <- promedios.muestra %>%
  mutate(varianza_intra_clase = (1 - n_minusc_h/N_mayusc_h)/n_minusc_h*S_h)
promedios.muestra <- promedios.muestra %>%
  mutate(sumando_var = (factor_pop^2)*varianza_intra_clase)
varianza.est <- sum(promedios.muestra$sumando_var)
```

Luego el intervalo está dado por:

```
c(
  Lower = xbarra - zalpha*sqrt(varianza.est),
  Upper = xbarra + zalpha*sqrt(varianza.est)
)

##      Lower      Upper
## 1178.268 1209.043
```

Para los intervalos de cada estrato usamos la varianza específica de los mismos:

```
promedios.muestra <- promedios.muestra %>%
  mutate(ic_lower = Media - !zalpha*sqrt(varianza_intra_clase)) %>%
  mutate(ic_upper = Media + !zalpha*sqrt(varianza_intra_clase))

kable(promedios.muestra) %>% kable_styling(latex_options = "striped")
```

Género	Entidad	Edad	Media	S_h	n_minusc_h	N_mayusc_h	Varianza
Hombre	Aguascalientes	[20,60]	1160.9109	64549.920	7	766	200
Hombre	Aguascalientes	< 20	1157.7719	72988.914	11	535	100
Hombre	Aguascalientes	>60	1215.9932	115311.993	20	260	500
Hombre	Baja California Norte	[20,60]	1249.1890	111605.495	13	751	200
Hombre	Baja California Norte	< 20	1225.9103	194636.825	14	475	100
Hombre	Baja California Norte	>60	1227.2721	102351.266	12	248	500
Hombre	Baja California Sur	[20,60]	1207.0631	105388.939	15	752	200
Hombre	Baja California Sur	< 20	1330.1951	151332.205	9	537	100
Hombre	Baja California Sur	>60	1169.8939	105405.420	11	239	500
Hombre	Campeche	[20,60]	1190.7680	109617.289	10	793	200
Hombre	Campeche	< 20	1296.5288	183909.715	15	536	100
Hombre	Campeche	>60	1100.7067	77398.895	6	275	500
Hombre	Chiapas	[20,60]	1058.5052	112526.529	11	758	200
Hombre	Chiapas	< 20	1182.7344	108902.265	11	499	100
Hombre	Chiapas	>60	1186.5285	146464.343	15	248	500
Hombre	Chihuahua	[20,60]	1221.0971	179155.101	17	726	200
Hombre	Chihuahua	< 20	1196.7915	66692.794	10	540	100
Hombre	Chihuahua	>60	1273.7945	62308.032	8	242	500
Hombre	Ciudad de México	[20,60]	1201.1904	81969.714	14	772	200
Hombre	Ciudad de México	< 20	1206.2439	132304.246	10	532	100
Hombre	Ciudad de México	>60	1247.1323	228599.912	10	227	500
Hombre	Coahuila	[20,60]	1167.4726	170040.368	15	753	200
Hombre	Coahuila	< 20	1173.5195	153515.205	20	514	100
Hombre	Coahuila	>60	1170.0741	124586.496	16	249	500
Hombre	Colima	[20,60]	1061.6194	81084.975	4	745	200
Hombre	Colima	< 20	1214.0360	275520.034	14	459	100
Hombre	Colima	>60	1208.3185	152782.922	11	256	500
Hombre	Durango	[20,60]	1143.4237	102923.725	16	783	200
Hombre	Durango	< 20	1212.6097	98808.579	14	512	100
Hombre	Durango	>60	1244.8630	13953.458	7	251	500
Hombre	Estado de México	[20,60]	1030.0824	109373.665	14	762	200
Hombre	Estado de México	< 20	1281.5619	167622.673	12	487	100
Hombre	Estado de México	>60	1285.7485	123599.194	10	240	500
Hombre	Guanajuato	[20,60]	1068.8483	70920.056	11	752	200
Hombre	Guanajuato	< 20	1359.0351	232779.860	10	504	100
Hombre	Guanajuato	>60	1321.4735	91641.590	10	250	500
Hombre	Guerrero	[20,60]	1215.0994	159936.739	12	738	200
Hombre	Guerrero	< 20	1278.6104	282360.300	8	488	100
Hombre	Guerrero	>60	1201.3422	124844.360	13	231	500
Hombre	Hidalgo	[20,60]	1129.2966	195408.174	14	739	200
Hombre	Hidalgo	< 20	1175.5184	117419.642	15	518	100
Hombre	Hidalgo	>60	1161.8690	127316.410	10	255	500
Hombre	Jalisco	[20,60]	1188.7770	6525.069	7	811	200
Hombre	Jalisco	< 20	1154.1446	74633.458	9	482	100
Hombre	Jalisco	>60	1422.8185	82379.428	14	243	500
Hombre	Michoacán	[20,60]	1237.4226	76144.062	16	720	200
Hombre	Michoacán	< 20	1162.9977	143248.797	11	537	100
Hombre	Michoacán	>60	1035.5259	122454.904	9	274	500
Hombre	Morelos	[20,60]	1096.9048	66995.607	12	752	200
Hombre	Morelos	< 20	1234.5305	76673.538	8	561	100
Hombre	Morelos	>60	1522.8539	169507.504	8	246	500
Hombre	Nayarit	[20,60]	1052.8052	202408.640	7	709	200
Hombre	Nayarit	< 20	1293.7826	65185.724	8	514	100

Chapter 6

Intervalos de Confianza mediante bootstrap

6.1 Inicio

En esta sección analizaremos más a fondo una técnica (llamada *bootstrap*) para realizar intervalos de confianza. Comenzaremos haciéndolo para el caso de experimentos y luego veremos cómo es distinto para el caso de muestreo aleatorio en poblaciones finitas y realizaremos ese caso.

6.2 Intervalos asintóticos

Los intervalos de confianza que hemos construido hasta ahora son con la idea de normalidad asintótica; se basan en la idea de que:

$$Z = \lim_{N-n,n \rightarrow \infty} \sqrt{\frac{1}{\text{Var}(\bar{x}_{\mathcal{S}})}} \cdot \left(\frac{1}{n} \sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i) - \bar{x}_{\mathcal{U}} \right) \sim \text{Normal}(0, 1)$$

Es decir, si tuviéramos una población infinita N y una muestra infinita n entonces la diferencia entre la media muestral y la media verdadera (normalizadas por la varianza) tienen una distribución normal. Empero, esto no siempre funciona como el siguiente con $n = 15$ muestras nos enseña:

```
#Tomar muestras de tamaño pequeño
n      <- 15
lambda <- 0.1
Z      <- qnorm(0.975)
nsim   <- 100    #Cantidad sims
N      <- 1000000 #Tamaño población
```

```

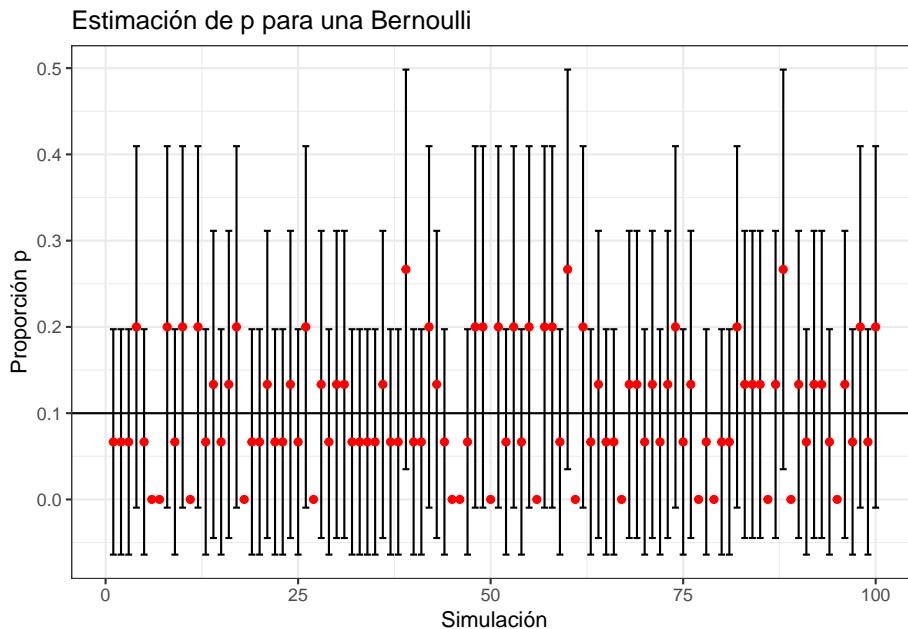
pop      <- sample(c(0,1), N, replace = TRUE, prob = c(1 - lambda, lambda))

#Creamos las bases donde guardar los datos
datos <- data.frame(matrix(NA, ncol = 3, nrow = nsim))
colnames(datos) <- c("IC_Bajo","Media","IC_Alto")
datos$Sim      <- 1:nsim

#Obtenemos 100 muestras
for (i in 1:nsim){
  muestra      <- sample(pop, n)
  datos$Media[i] <- mean(muestra)
  varianza     <- (1 - n/N)/n*var(muestra)
  datos$IC_Bajo[i] <- datos$Media[i] - Z*sqrt(varianza)
  datos$IC_Alto[i] <- datos$Media[i] + Z*sqrt(varianza)
}

ggplot(datos, aes(x = Sim)) +
  geom_errorbar(aes(ymin = IC_Bajo, ymax = IC_Alto)) +
  geom_point(aes(y = Media), color = "red") +
  geom_hline(aes(yintercept = mean(pop) )) +
  theme_bw() +
  labs(
    x = "Simulación",
    y = "Proporción p",
    title = "Estimación de p para una Bernoulli"
  )

```



En el caso de la Bernoulli, por ejemplo, se puede un mejor intervalo de confianza para cuando la proporción $\hat{p} = 0$ bajo una hipótesis distribucional. Esta se encuentra en el paquete `binom` de R:

```
library(binom)
binom.confint(0, n, conf.level = 0.95, method = 'exact')

##   method x  n mean lower      upper
## 1 exact 0 15  0    0 0.2180194
```

La deducción de dichos intervalos se hace a través de un esquema distinto de muestreo: **muestreo basado en modelos**.

6.3 Muestreo basado en modelos

Quizá en otras clases de estadística viste muestreo de otra manera. Lo usual en Estadística Matemática es suponer se tienen un conjunto de n variables aleatorias independientes idénticamente distribuidas que representan la cantidad de interés $\{Y_1, Y_2, \dots, Y_n\}$. De dichas variables se observa que toman los valores $\{y_1, y_2, \dots, y_n\}$ que son los medidos en la muestra. Aquí no se considera que haya un conjunto finito (universal) dado por la **población** sino que todas las variables provienen de una **metapoblación** (universo infinito de posibilidades). Así, por ejemplo, las alturas de individuos podrían tener una distribución gamma truncada y las alturas de personas en particular corresponden a realizaciones *independientes* de dichas Y_i . Este enfoque lo que hace es convertir el problema de estimación en un problema de predicción. No es que *la altura* de un individuo sea aleatoria en el sentido de que siempre cambie sino que *por nuestra ignorancia*

nosotros modelamos dicha altura con una variable aleatoria que representa, en esta **metapoblación** la altura de una cantidad infinita de individuos posibles. Bajo este esquema, una muestra aleatoria está dada por:

$$\mathcal{Y}_{(n)} = \{Y_1, Y_2, \dots, Y_n\}$$

y la muestra observada es:

$$\dagger_{(n)} = \{y_1, y_2, \dots, y_m\}$$

Donde suponemos que lo que se observó fue que $Y_i = y_i$. Un punto relevante aquí es que como las $\{Y_i\}$ son variables aleatorias éstas siguen algún modelo. En particular, se eligen modelos paramétricos para estos casos en los que la distribución asintótica no funciona. El problema de estimación se convierte ya sea en estimar el parámetro de la función de densidad (por ejemplo el $\Theta = (\mu, \sigma)^T$ de la normal o el λ de una Poisson) o una función del mismo (como puede ser la mismísima función de densidad o distribución acumulada).

Nota El enfoque basado en modelos es muy bueno cuando se tienen pocos datos (o mal medidos) y las observaciones por sí mismas no son suficientes para poder generar información. En ese caso se hacen hipótesis adicionales (como un modelo) para los procesos de estimación.

Veamos un ejemplo de generación de intervalos para la media (λ) de una variable Poisson.

6.3.1 Ejemplo Poisson

Consideremos una muestra aleatoria $\mathcal{Y}_{(n)} = \{Y_1, Y_2, \dots, Y_n\}$ de variables que se distribuyen Poisson(λ). Sabemos de proba (y si no lo sabemos no te apures, este ejemplo es sólo para ilustrar, no vamos a hacer más de esto) que la suma de variables aleatorias Poisson es Poisson, de donde:

$$\sum_{i=1}^n Y_i \sim \text{Poisson}(n \cdot \lambda)$$

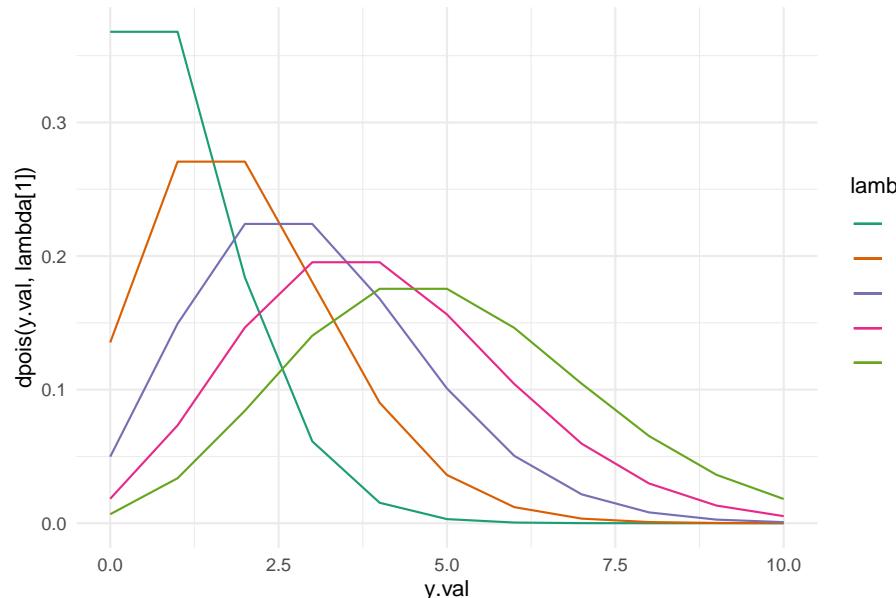
por lo cual:

$$n\bar{Y} = \sum_{i=1}^n Y_i \sim \text{Poisson}(n\lambda)$$

tenemos entonces que $\mathbb{E}[\bar{Y}] = \lambda$ (es insesgado). En este caso tomaremos que en particular observamos

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

Observamos que para y fijo, la distribución Poisson es decreciente antes de la media y creciente después de la media:



Podemos entonces obtener un estimador λ consideraremos los valores λ_1 y λ_2 tales que:

$$\sum_{k \leq \hat{y}} \frac{(n\lambda_1)^{\hat{y}} e^{-n\lambda_1}}{\hat{y}!} = \alpha/2 \quad \text{y} \quad \sum_{k \geq \hat{y}} \frac{(n\lambda_2)^{\hat{y}} e^{-n\lambda_2}}{\hat{y}!} = \alpha/2$$

El cual podemos encontrar facilmente con ayuda de R:

```
#Obtenemos la muestra
n      <- 20
ybarra <- sum(rpois(n, 1/5))
alpha.val <- 0.05

#Optimizamos
func.opt.1 <- function(lambda){ppois(ybarra, n*lambda) - alpha.val/2}
lambda.1   <- uniroot(func.opt.1, lower = 0, upper = 10, tol = 1.e-10)$root

func.opt.2 <- function(lambda){1 - ppois(ybarra, n*lambda) + dpois(ybarra, n*lambda) - alpha.val/2}
lambda.2   <- uniroot(func.opt.2, lower = 0, upper = 10, tol = 1.e-10)$root

c("Lower" = lambda.2, "Upper" = lambda.1)

##      Lower      Upper
## 0.03093361 0.43836365
```

Nota Estos intervalos se conocen como *fiducia*rios/*fiduciales*. La

idea es obtener los valores de λ tales que $\bar{Y} = Y$ con un intervalo de probabilidad $(1 - \alpha) \times 100\%$.

En nuestro caso para análisis de encuestas **no podemos usar directamente los resultados de estadística matemática que refieren más a experimentos** pues en ello se supone independencia (lo cual casi nunca es cierto *para encuestas*). Sin embargo, tras el modelo que hemos estudiado ahora sí podríamos generar variables *tipo* las de estadística matemática con una hipótesis distribucional dadas por $\{Z_i\}$ donde

$$Z_k = \mathbb{I}(x_k)$$

pero sin que Z_i sea independiente de Z_j . En este caso las Z_i siguen una distribución multivariada dictada por las π_k .

6.3.2 Ejercicio

Obtén los intervalos fiduciarios para una variable aleatoria Bernoulli. Considera que la muestra está dada por:

```
set.seed(646)
muestra <- sample(c(0,1), 13, replace = TRUE, prob = c(0.2, 0.8))
```

Recuerda Si $X_i \sim \text{Bernoulli}(p)$ entonces $\sum_{i=1}^n X_i \sim \text{Binomial}(n, p)$.

6.4 Intervalos Bootstrap

Los intervalos anteriores requieren de manera implícita de la función de distribución acumulada original la cual puede ser muy complicada o desconocida. Los intervalos Bootstrap buscan aproximar dicha función mediante un remuestreo. La demostración exacta (con medida) de por qué funciona el Boostrap la puedes ver en Asymptotic Statistics. La idea es *reconstruir* la función de distribución acumulada mediante remuestreo de la base (de una mejor manera aún que con la \hat{F} empírica) y calcular estadísticos a partir de ella. Algunos estadísticos que se pueden calcular a partir de ella es, por ejemplo, la media muestral o la varianza así como los cuantiles. Para ello definiremos un concepto importante:

Definición Un estadístico t es un funcional estadístico si t es simétrica para y_1, y_2, \dots, y_N . Dicho de otra forma, t depende solamente de los valores ordenados $y_{(1)}, y_{(2)}, \dots, y_{(N)}$ y no del orden preciso (de la muestra o la población). En este caso t depende sólo de la función de distribución y se escribe como $t(F)$ (en el caso poblacional) ó $t(\hat{F})$ en el caso muestral.

Un ejemplo de estadístico es la media pues (suponiendo F discreta) por ejemplo:

$$\mu = \sum_x x \left[\underbrace{F(x) - F(x^-)}_{\mathbb{P}(X=x)} \right]$$

Mientras que de manera muestral la estimación es idéntica mediante estimadores:

$$\hat{\mu} = \sum_x x \underbrace{[\hat{F}(x) - \hat{F}(x^-)]}_{\hat{\mathbb{P}}(X=x)}$$

donde la notación $F(x^-) = \lim_{y \rightarrow x^-} F(y)$.

Un ejemplo adicional es la mediana que corresponde a:

$$\text{Mediana} = F^{-1}(1/2)$$

donde

$$F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$$

es una pseudoinversa conocida como la función cuantil. (En el caso de F continua es una inversa “bien” pero discreta no lo es porque esa cosa ni es invertible).

La idea de Bootstrap (con una demostración bastante técnica) es que para una \hat{F}_B función de distribución empírica acumulada generada por Boostrap se tiene que:

$$\sqrt{n}(t(\hat{F}_B) - t(\hat{F}))$$

converge en distribución al mismo límite que:

$$\sqrt{n}(t(\hat{F}) - t(F))$$

para cualquier t función Hadamard-diferenciable..

Lo que en humano quiere decir es que muchísimos funcionales estadísticos convergen al verdadero. Lo interesante de Bootstrap (usualmente) no va a ser estimar el t sino estimar un intervalo para t a partir de remuestrear varias \hat{F} mediante simulación.

Nota que no sé dónde poner Un tipo de Bootstrap específico es el Jackknife donde se elimina aleatoriamente un elemento de la base. *Model Assisted Survey Sampling* lo pone como otra cosa pero en realidad es un boostrap bajo otro método.

6.4.1 Bootstrap bajo un modelo

Regresemos a nuestro ejemplo Poisson de la media pero esta vez usando Boostrap. Para ello consideraremos la muestra de tamaño $n = 20$.

```

n      <- 20
muestra <- rpois(n, 1/5)
alpha.val <- 0.05
nsim     <- 1000 #Mil simulaciones nada más para no trabar
media.completa <- mean(muestra)
media      <- rep(NA, nsim)
for (i in 1:nsim){
  
```

```

remuestreo <- sample(muestra, n, replace = TRUE)
media[i]   <- mean(remuestreo)
}

#Cálculo de cotas
lado.sup <- quantile(media - media.completa, alpha.val/2)
lado.inf <- quantile(media - media.completa, 1 - alpha.val/2)

print(paste0("El intervalo es [",
            media.completa - lado.inf,
            ",",
            media.completa - lado.sup, "])"))

## [1] "El intervalo es [-0.05,0.4]"

```

El algoritmo es como sigue: 1. Obtén una muestra aleatoria simple con reemplazo $X_1^B, X_2^B, \dots, X_n^B \sim \hat{F}$ y calcula $t(X_1^B, X_2^B, \dots, X_n^B)$ 2. Repite n veces (este n se decide de recalculando la n como si viniera de una muestra para tener la precisión deseada) o bien por convergencia. 3. Calcula los cuantiles simulados correspondientes $\delta_1 = (t_B - t)_{\alpha/2}$ y $\delta_2 = (t_B - t)_{1-\alpha/2}$ 4. El intervalo es $t - \delta_1$ y $t - \delta_2$

Nota Boostrap no es perfecto para todo. Por ejemplo si se tiene una muestra de variables aleatorias $X_i \sim \text{Unif}(0, 1)$ y se busca $t_n = \min\{X_i\}$ bajo Bootstrap se puede demostrar que

$$nt_n \rightarrow \text{Exponencial}(1)$$

lo cual no corresponde al valor verdadero

6.4.2 Bootstrap de una muestra

En el caso de una muestra aleatoria simple ponderada con estratos excluyentes S_1, S_2, \dots, S_H de tamaños N_1, N_2, \dots, N_H para cada estrato S_h seleccionamos una muestra bootstrap a partir de lo cual podemos calcular una función de distribución empírica para cada estrato \hat{F}_h y repetir el algoritmo (para cada estrato):

1. Crea una subpoblación U_h^f fija de tamaño $k = \left\lfloor \frac{N_h}{n_h} \right\rfloor$ repitiendo la original k veces.
2. Obtén U_h^c una muestra aleatoria de S_h de tamaño $N - nk$. Crea la pseudo población bootstrap como $U_h^{boot} = U_h^c \cup U_h^f$.
3. Obtén una muestra de U_h^{boot} de tamaño n_h y recálcula el parámetro de interés $t_B = t(\hat{F}_h^{boot})$.
4. Repite m veces (este m se decide de recalculando la n como si viniera de una muestra para tener la precisión deseada) o bien por convergencia.
5. Calcula los cuantiles simulados correspondientes $\delta_1 = (t_B - t)_{\alpha/2}$ y $\delta_2 = (t_B - t)_{1-\alpha/2}$
6. El intervalo es $t - \delta_1$ y $t - \delta_2$

El método anterior es el desarrollado por Booth, Butler y Hall; sin embargo no es el único que existe. Para poblaciones finitas te sugiero checar el compendio de Mashreghi, Haziza y Léger donde vienen varios métodos.

6.4.3 Ejemplo

Calcular la mediana de la siguiente población:

```
pop.total <- c(rnorm(100), rexp(100), rpois(100, 2))
N <- length(pop.total)

#Verdadera mediana
mediana.real <- median(pop.total)

#Obtenemos muestra de tamaño n = 25 por ejemplo
n <- 25
muestra      <- sample(pop.total, 25)
mediana.muestral <- median(muestra)
mediana       <- rep(NA, nsim)

print(paste0("El valor real es ", mediana.real))

## [1] "El valor real es 0.629704617234562"

#Tamaño de bootstrap
k.val <- floor(N/n)
Uf    <- rep(muestra, k.val)

for (i in 1:nsim){
  Uc      <- sample(muestra, N - n*k.val, replace = TRUE)
  Uboot   <- c(Uf, Uc)
  remuestreo <- sample(Uboot, n, replace = FALSE)
  mediana[i] <- median(remuestreo)
}

#Cálculo de cotas
lado.sup <- quantile(mediana - mediana.muestral, alpha.val/2)
lado.inf <- quantile(mediana - mediana.muestral, 1 - alpha.val/2)

print(paste0("El intervalo es [", mediana.muestral - lado.inf,
            ", ", mediana.muestral - lado.sup, "]"))

## [1] "El intervalo es [-0.12970966652169,0.838248776470654]"
```

Semilla	Nh
301	960
303	1008
305	968
307	1029
309	995
311	994
315	1025
319	979
321	1008
323	1046
325	1064
327	1012
329	996
331	1015

6.4.4 Ejercicio

De una población de árboles se midieron las alturas según distintos tipos de semilla. Esta información está contenida en R en una base de datos precargada:

```
data(Loblolly)
arboles <- Loblolly
```

1. Suponiendo que para la semilla 305 se realizó muestreo aleatorio simple sin reemplazo de un total $N = 100$ árboles, estima la mediana de altura y los cuantiles 0.25 y 0.75 de la misma junto con sus intervalos de confianza mediante *bootstrap*.
2. Suponiendo que para la semilla 307 se realizó muestreo aleatorio simple sin reemplazo de un total $N = 300$ árboles, estima la media y varianza de altura junto con sus intervalos de confianza mediante *bootstrap*.
3. Estima para **todos los árboles** (es decir no sólo para los estratos sino para la población completa) la altura promedio, la altura mediana, los cuantiles 0.25 y 0.75 de altura así como la curtosis y la asimetría de la misma junto con sus intervalos de confianza mediante *bootstrap* suponiendo las N_h son:
4. Considera una población de tamaño $N = 1000$ y una muestra aleatoria sin reemplazo de la misma de tamaño $n = 71$ dada por:

```
set.seed(64)
muestra <- rlnorm(71)
```

grafica la función de distribución empírica (asociada a la muestra) y grafica intervalos de confianza para la misma en el intervalo [0.1, 0.9].

Tip Genera intervalos de confianza mediante bootstrap asociados a $\hat{F}(0.1)$, $\hat{F}(0.2)$, $\hat{F}(0.3)$, $\hat{F}(0.4)$, $\hat{F}(0.5)$, $\hat{F}(0.6)$, $\hat{F}(0.7)$, $\hat{F}(0.8)$, y $\hat{F}(0.9)$

Chapter 7

Muestreo Aleatorio Multietápico

La idea de un muestreo aleatorio multietápico es un modelo donde la aleatoriedad ocurre más de una vez. El ejemplo común es el de encuestadores que van a una casa. De inicio la casa se selecciona al azar usando un censo o un mapa (primer fuente de aleatoriedad) y una vez que se llega a la casa y se le pregunta a la gente de la misma cuántos habitantes hay, se selecciona a una persona de la casa de manera aleatoria (segunda fuente de aleatoriedad) para realizarle la encuesta.

Nota La diferencia esencial entre una etapa y un estrato es que el estrato no es aleatorio. Un estrato es cómo decido yo dividir mi muestra que tenga sentido de manera explicativa (por ejemplo hombre/mujer/otrx o por estado civil, grupo de edad). Pero no es que *aleatoriamente* se seleccione un hombre o una mujer sino que es una característica de la población que se mide y en base a la cual se clasifica. Por otro lado, una etapa va a ser algo que se selecciona de manera aleatoria.

Hay dos razones principales para hacer muestreo multietápico:

- a. *Se desconoce la población total.* No hay un censo de toda la población. Pero sí hay forma de identificar otros grupos dentro de los cuales se puede muestrear y obtener un censo.
- b. *La muestra está repartida en un área demasiado grande* (piensa, por ejemplo, muestrear en una selva) por lo que es imposible cubrir toda el área por lo que se muestran regiones y luego subregiones a fin de tener toda la muestra (en la selva igual se seleccionarían al azar regiones y luego subregiones).

El primer ejemplo que analizaremos será un *muestreo aleatorizado por clusters*

de una sola etapa. De ahí saltaremos al bietápico y daremos las líneas para el multietápico (sin meternos mucho en ello). Para ello la definición de cluster será un subgrupo en el cual se divide la población donde la selección aleatoria ocurre entre los grupos (y no entre las unidades primarias de muestreo) y luego dentro del cluster se realiza un censo.

Por poner un ejemplo, de *muestreo aleatorizado por clusters de una sola etapa* podría interesar conocer el uso de drogas dentro de estudiantes de preparatoria. Para ello se divide a la población estudiantil en subgrupos (las preparatorias) y se muestrean las preparatorias (el cluster). Una vez seleccionadas algunas preparatorias dentro de las mismas se hace un censo de todos los estudiantes preguntándoles sobre consumo de drogas.

El algoritmo multietápico suele ser como sigue:

- a. Se agrupan elementos poblacionales en subpoblaciones disjuntas conocidas como **unidades primarias de muestreo** (PSU) dentro de las cuales se toma una muestra (*primera etapa*)
- b. Los elementos de cada PSU se consideran nuevas unidades de muestreo (SSU, unidades secundarias de muestreo) las cuales pueden ser elementos (de los cuales se toma un censo) o bien nuevos clusters
- c. En caso que la SSU sean nuevos clusters el proceso se repite en tantas etapas como sea necesario hasta tener un censo de elementos (en algún nivel). Estas se conocen como unidades terciarias de muestro (o penúltimas, n -ésimas, lo que sea que aplique según cuántas se tengan).

7.1 Muestreo aleatorio por clusters en una sola etapa

En este caso se tiene una población finita $U = (x_1, x_2, \dots, x_N)^T$ partitionada en N_t subpoblaciones llamadas *clusters* denotadas U_1, U_2, \dots, U_{N_t} . Generalmente identificamos al cluster por su subíndice:

$$\mathcal{U}_1 = \{1, 2, \dots, N_1\}$$

En este caso se tiene que:

$$U = \bigcup_{i \in \mathcal{C}_1} U_i \quad \text{y} \quad N = \sum_{i \in \mathcal{C}_1} N_i$$

El muestreo aleatorio por clusters en una sola etapa se define como sigue:

- a. Se obtiene una muestra~[La muestra puede ser estratificada, aleatoria simple, bernoulli]
- b. Se mide cada elemento de los clusters seleccionados.

Podemos hacer dicho esquema de muestreo en R, por ejemplo si consideramos una base de escuelas donde se tiene el registro de cada uno de los alumnos en la escuela:

```
#Creamos la base de datos
escuelas      <- paste0("Escuela ", 1:20)
datos.escuelas <- data.frame(Escuela = sample(escuelas, 1000, replace = TRUE),
                               Promedio_Alumno = runif(1000, 6, 10))

#Seleccionamos las escuelas al azar (los clusters)
escuelas.seleccionadas <- sample(escuelas, 5)

#Una vez se tiene el cluster se produce un censo de los clusters
muestra <- datos.escuelas %>% filter(Escuela %in% escuelas.seleccionadas)
```

En este caso denotaremos la muestra observada como:

$$\mathcal{S} = \bigcup_{i \in \mathcal{S}_1} U_i$$

con su tamaño respectivo:

$$n_{\mathcal{S}} = \sum_{i \in \mathcal{S}_1} N_i$$

> Toma en cuenta que aunque la muestra de clusters sea de tamaño fijo en general el número de elementos observados no va a ser fijo pues los tamaños de cluster pueden variar.

Agregamos unas nuevas probabilidades de pertenencia. Para un elemento $x_k \in U_i$ se tiene que:

$$\pi_k = \mathbb{P}(x_k \in \mathcal{S}) = \mathbb{P}(U_i \subseteq \mathcal{S}) = \pi_{1,i}$$

mientras que para $x_k \in U_i$ y $x_l \in U_j$ se tiene:

$$\pi_{k,l} = \mathbb{P}(x_k \in \mathcal{S}, x_l \in \mathcal{S}) = \mathbb{P}(U_i \subseteq \mathcal{S}, U_j \subseteq \mathcal{S}) = \pi_{1,i,j}$$

El total poblacional puede escribirse como:

$$t = \sum_{U_i} t_i = \sum_{k=1}^N x_k$$

donde $t_i = \sum_{x_k \in U_i} x_k$ son los totales de cada cluster. En este contexto tenemos que:

$$\hat{t} = \sum_{i \in \mathcal{S}_1} \frac{t_i}{\pi_{1,i}}$$

es un estimador insesgado del total poblacional. Su varianza está dada por:

$$\text{Var}(\hat{t}) = \sum_{j \in \mathcal{C}_1} \sum_{i \in \mathcal{C}_1} \frac{\Delta_{1ij}}{\pi_{1i}\pi_{1j}} t_i t_j$$

Cuadra	Casas.en.la.cuadra	Total.de.ingreso.en.la.cuadra
1	120	2100
2	100	2000
3	80	1500

mientras que un estimador insesgado es:

$$\widehat{\text{Var}}(\hat{t}) = \sum_{j \in \mathcal{S}_1} \sum_{i \in \mathcal{S}_1} \frac{1}{\pi_{1ij}} \frac{\Delta_{1ij}}{\pi_{1i}\pi_{1j}} t_i t_j$$

Las demostraciones de que dichos estimadores son insesgados son idénticas a las que ya hicimos con muestreo aleatorio simple y no las repetiremos.

En particular podemos tener que en el caso de muestreo aleatorio simple sin reemplazo se tiene:

$$\hat{t} = N_1 \bar{t}_{\mathcal{S}_1}$$

donde $\bar{t}_{\mathcal{S}_1} = \sum_{i \in \mathcal{S}_1} t_i / n_1$ es la media de los totales. En este caso las expresiones de las varianzas son idénticas a las que ya conocemos (justo porque se deducen de la misma manera):

$$\text{Var}(\hat{t}) = N_1^2 \frac{1 - f_1}{n_1} S_{t, \mathcal{U}_1}^2$$

donde $f_1 = n_1 / N_1$ es la fracción muestral del cluster y

$$S_{t, \mathcal{U}_1}^2 = \frac{1}{N_1 - 1} \sum_{i \in \mathcal{U}_1} (t_i - \bar{t}_{\mathcal{U}_1})^2$$

con $\bar{t}_{\mathcal{U}_1} = \sum_{i \in \mathcal{U}_1} t_i / N_1$. El estimador de la varianza corresponde al muestral:

$$\widehat{\text{Var}}(\hat{t}) = N_1^2 \frac{1 - f_1}{n_1} S_{t, \mathcal{S}_1}^2$$

donde

$$S_{t, \mathcal{S}_1}^2 = \frac{1}{n_1 - 1} \sum_{i \in \mathcal{S}_1} (t_i - \bar{t}_{\mathcal{S}_1})^2$$

7.1.1 Ejemplo

El objetivo es estimar la media de ingreso de hogares en una colonia de una ciudad que consiste de 60 cuadras de casas de tamaño variable. Para esto seleccionamos tres cuadras usando muestreo aleatorio simple sin reemplazo y entrevistamos a todos los hogares en dichas cuadras. Se sabe, además que hay 5000 casas en esta colonia y la tabla muestra los resultados de la encuesta

Estimar la media de ingreso y su varianza.

Solución Tenemos que $n_1 = 3$, $N = 5000$ y $N_1 = 60$. Las probabilidades de inclusión son:

$$\pi_{1i} = \frac{n_1}{N_1} = \frac{1}{20}$$

Se construye el estimador de la media mediante:

$$\hat{\bar{x}} = \frac{1}{N} \sum_{i \in S} \underbrace{\frac{t_i}{\frac{n_1}{N_1}}}_{\pi_{1i}} = \frac{N_1}{N} \frac{1}{n_1} \sum_{i \in S} t_i \approx 22.4$$

Por otro lado como $\hat{\bar{x}} = \hat{t}/N$ entonces se tiene que

$$\text{Var}(\hat{\bar{x}}) = \text{Var}(\hat{t})/N^2$$

de donde se obtiene:

$$\widehat{\text{Var}}(\hat{\bar{x}}) = \approx 4.7$$

(El resultado anterior sale sólo de sustituir).

7.2 Muestreo aleatorio por clusters bietápico (en dos etapas)

Usualmente es demasiado caro hacer el censo *dentro* del cluster por lo que se muestrea dentro del mismo una vez seleccionado. (Por ejemplo, en el caso de las drogas en las escuelas es más fácil seleccionar sólo unos alumnos al azar y no a todos los alumnos). En estos casos hay dos fuentes de aleatoriedad: las unidades primarias de muestreo (los clusters más grandes) y las secundarias (muestreo dentro del cluster), La notación se va a complicar pero el principio es el mismo.

La idea es que la población de elementos:

$$\mathcal{U} = \{x_1, x_2, \dots, x_N\}$$

es subdividida en N unidades primarias de muestreo denotadas U_1, \dots, U_{N_1} . El tamaño de U_i es N_u . El diseño muestral es como sigue:

- Se obtiene una muestra \mathcal{S}_1 de unidades primarias de muestreo de acuerdo con algún diseño \mathbb{P}_1 .
- Para cada $U_i \in \mathcal{S}_1$ se selecciona una muestra de S_i elementos de U_i de acuerdo al diseño condicional $\mathbb{P}_i(\cdot | \mathcal{S}_1)$

La muestra resultante de elementos se denota:

$$\mathcal{S} = \bigcup_{i \in \mathcal{S}_1} S_i$$

En este capítulo pediremos dos cosas para los muestreos condicionales:

Independencia Que el muestreo de U_i sea independiente del de U_j . Matemáticamente esto se escribe como:

$$\mathbb{P}\left(\bigcup_{i \in \mathcal{S}} S_i | \mathcal{S}_1\right) = \prod_{i \in \mathcal{S}} \mathbb{P}(S_i | \mathcal{S}_1)$$

Invarianza Una vez que se incluye el U_i en la muestra el muestreo siempre es igual independientemente de los U_j ; es decir: $\mathbb{P}_i(\cdot | \mathcal{S}_1) = \mathbb{P}_i(\cdot)$.

Los tamaños de muestra se definen como sigue: el número de unidades primarias de muestreo es S_i es n_i . El número total de elementos en \mathcal{S} está dado por:

$$n_{\mathcal{S}} = \sum_{i \in \mathcal{S}} n_i$$

Las probabilidades de inclusión en la primera etapa \mathbb{P}_1 son:

$$\Delta_{1ij} = \pi_{1ij} - \pi_{1i}\pi_{1j}$$

con

$$\Delta_{1ii} = \pi_{1i}(1 - \pi_{1i})$$

Para la segunda etapa (el muestreo adentro de un \mathcal{S}_i) las cantidades son:

$$\Delta_{kl|i} = \pi_{kl|i} - \pi_{k|i}\pi_{l|i}$$

con $\Delta_{kk|i} = \pi_{k|i}(1 - \pi_{k|i})$. Notamos que por invarianza e independencia:

$$\pi_k = \pi_{1i}\pi_{k|i} \quad \text{para } x_k \in U_i$$

mientras que:

$$\pi_{kl} = \begin{cases} \pi_{1i}\pi_{k|i} & \text{si } x_k = x_l \in U_i \\ \pi_{1i}\pi_{kl|i} & \text{si } x_k, x_l \in U_i \\ \pi_{1ij}\pi_{k|i}\pi_{l|j} & \text{si } x_k \in U_i \text{ y } x_l \in U_j \end{cases}$$

Un estimador del total poblacional (insesgado) es el ya usual estimador HT:

$$\hat{t} = \sum_{\mathcal{S}_1} \frac{\hat{t}_i}{\pi_{1i}}$$

donde \hat{t}_i es el estimador de t_i el total de U_i . La varianza de \hat{t} se puede escribir como:

$$\text{Var}(\hat{t}) = V_{PSU} + V_{SSU}$$

donde

$$V_{PSU} = \sum_{U_i} \sum_{U_j} \frac{\Delta_{1ij}}{\pi_{1i}\pi_{1j}} t_i t_j$$

$$V_{SSU} = \sum_{U_i} \frac{V_i}{\pi_{1i}}$$

con

$$V_i = \sum_{U_i} \sum_{kl|i} \Delta_{kl|i} \frac{x_k}{\pi_{k|i}} \frac{x_l}{\pi_{l|i}}$$

Los estimadores insesgados así como las expresiones reducidas de la varianza pueden hallarse en el formulario bietápico disponible en comunidad. **Demostación** Checar el Sarndal páginas 136-139.

7.3 Ejemplo: Disco duro

En el disco duro de una computadora hay 400 bases de datos cada una de las cuales consiste en 50 entradas (renglones). Se desea estimar el número de caracteres por entrada por lo que se hace muestreo aleatorio simple de los 80 archivos y luego dentro de cada archivo muestreo aleatorio simple para 5 entradas. En el caso del formulario de comunidad, sean $m = 80$ y $n = 5$. Después de muestrear obtenemos:

- a. La media muestral de los estimadores para el número total de caracteres por archivo dada por: $s_T^2 = 905000$
- b. La media de las m varianzas muestrales s_i^2 es igual a 805 donde s_i^2 representa la varianza del número de caracteres por entrada en el i ésimo archivo.

Estimar el número promedio de caracteres por entrada junto con su precisión de manera insesgada. Dar un intervalo de confianza al 95%.

Solución

Denotando $y_{i,k}$ al número de caracteres de la entrada k del archivo i tenemos que la cantidad de interés es:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^M \sum_{k \in U_i} y_{i,k} = \frac{1}{N} \sum_{i=1}^M \bar{N} \bar{y}_i = \frac{1}{M} \sum_{i=1}^M \bar{y}_i$$

donde

- a. $M = 400$ es el número de archivos (PSU)
- b. $\bar{N} = 50$ es el número de entradas por archivo (SSU)
- c. $N = 400 \times 50$ es el número total de entradas (n_S)
- d. \bar{y}_i es el número promedio de caracteres por entrada del archivo i .
- e. U_i es el cluster i (identificadores de las entradas del archivo i en este caso).

El estimador insesgado de la media sería

$$\hat{y} = \frac{\hat{y}}{N} = \frac{1}{N} \sum_{i \in S} \frac{\hat{t}_i}{m/M}$$

donde \hat{t}_i son los estimadores de los totales del archivo i y \mathcal{S} es la colección de índices seleccionados para la muestra. En particular:

$$\hat{t}_i = \sum_{k \in \mathcal{S}_i} \frac{y_{i,k}}{\bar{n}/\bar{N}} =$$

El estimador de $\widehat{\text{Var}}(\hat{y}) = \frac{1}{N^2} \widehat{\text{Var}}(\hat{t}) \approx 3.98$. cuando se sustituye. Finalmente:

$$\hat{y} \pm Z_{1-0.05/2} \sqrt{3.98}$$

da el intervalo de confianza.

7.4 Ejemplo: Encuesta Nacional de Salud

La Encuesta Nacional de Salud y Nutrición 2018 es una encuesta nacional estratificada bietápica. La nota metodológica completa puede hallarse en el reporte. A partir de la lectura de la nota metodológica se establece el diseño muestral:

```
library(readr)
library(survey)

## Loading required package: grid

##
## Attaching package: 'grid'

## The following object is masked from 'package:imager':
## 
##     depth

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyverse':
## 
##     expand, pack, unpack

## Loading required package: survival

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
## 
##     dotchart
```

```

ensanut <- read_csv("~/Dropbox/ITAM_CLASES/Aplicada1/Archivos_2020/ENSANUT_1.csv")

## 
## -- Column specification -----
## cols(
##   .default = col_double(),
##   UPM = col_character(),
##   VIV_SEL = col_character(),
##   NUMREN = col_character(),
##   P1_2 = col_character(),
##   P1_5 = col_character(),
##   P1_8 = col_character(),
##   P3_3V = col_character(),
##   P3_4 = col_character(),
##   P3_5D = col_character(),
##   P3_5M = col_character(),
##   P3_5A = col_character(),
##   P3_6 = col_character(),
##   P3_9M = col_character(),
##   P3_9A = col_character(),
##   P3_10M = col_character(),
##   P3_10A = col_character(),
##   P3_12 = col_character(),
##   P3_15_1 = col_character(),
##   P3_15_2 = col_character(),
##   P3_15_3 = col_character()
##   # ... with 122 more columns
## )
## i Use `spec()` for the full column specifications.

## Warning: 1342 parsing failures.
##   row     col           expected    actual
## 1281 P10_7_7  1/0/T/F/TRUE/FALSE  05  '~/Dropbox/ITAM_CLASES/Aplicada1/Archivos_2020/ENSANUT
## 2093 P10_7_7  1/0/T/F/TRUE/FALSE  09  '~/Dropbox/ITAM_CLASES/Aplicada1/Archivos_2020/ENSANUT
## 2349 P9_9_B2D 1/0/T/F/TRUE/FALSE  30  '~/Dropbox/ITAM_CLASES/Aplicada1/Archivos_2020/ENSANUT
## 2349 P9_9_B2M 1/0/T/F/TRUE/FALSE  09  '~/Dropbox/ITAM_CLASES/Aplicada1/Archivos_2020/ENSANUT
## 2349 P9_9_B2A 1/0/T/F/TRUE/FALSE 2008 '~/Dropbox/ITAM_CLASES/Aplicada1/Archivos_2020/ENSANUT
## ....
## See problems(...) for more details.

#Diseño de encuesta completa
ensanut <- ensanut %>% mutate(id = paste0(VIV_SEL, NUMREN))

#Codificar diabéticos
ensanut <- ensanut %>% mutate(Diabetes = ifelse(P3_1 == 1, 1, 0))

```

```
#Diseño muestral
```

```
diseño <- svydesign(id= ~id, strata= ~EST_DIS, weights=~F_20MAS, PSU=~UPM, data= ensa)
```

Podemos utilizar el diseño muestral y el paquete survey para estimaciones como calcular la proporción nacional de diabéticos:

```
media <- svymean(~Diabetes, diseño)
print(media)
```

```
##           mean      SE
```

```
## Diabetes 0.10321 0.0024
```

```
confint(media)
```

```
##           2.5 %    97.5 %
```

```
## Diabetes 0.09849833 0.1079285
```

Appendix A

Programación en R



Figure A.1: ‘R’ es un programa chido de estadística. FIN.

Una de las primeras cosas que necesitamos saber es que R (por más que sus más ávidos defensores digan lo contrario) no es para todo. Si tú ya conoces otro lenguaje (sea **Stata**, **Excel**, **SAS**, **Python**, **Matlab**, **Julia**, etc) sabrás utilizar muchas de sus opciones. Estoy seguro que, de conocer uno de estos, te será muchísimo más fácil seguir sacando promedios en tu lenguaje favorito que en R, realizar regresiones lineales es probablemente más sencillo en **Stata** mientras que las gráficas de barras para mí son más simples en **Excel**, **Python** excede en aplicaciones de inteligencia artificial mientras que **Matlab** es más veloz que R, **Julia** tiene muchas cosas de ecuaciones diferenciales que nadie más.

Lo que probablemente no sea más sencillo de hacer en otro lenguaje es realizar análisis estadístico, gráficas de todo tipo y modelos de simulación. Para eso, R es, indiscutiblemente, una de las mejores opciones para quienes no conocen de programación¹.

¹Modelos de simulación más avanzados suelen hacerse en **C**, **C++** o **Fortran** por su velocidad; empero, es necesario conocer más de programación.

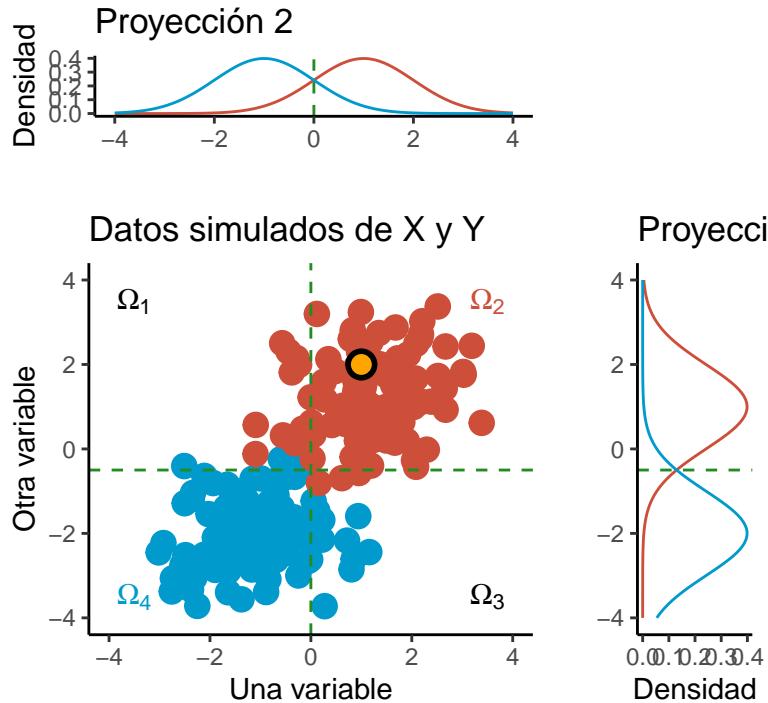
Finalmente, uno de los consejos más importantes que te puedo dar es que este curso no te va a servir si no practicas. Igual que como pasa con los idiomas uno no aprende R en una semana *sin practicarlo después*. Mi sugerencia es que, a la vez que sigues estas notas comiences a trabajar un proyecto *tuyo* específico junto con el buscador de Internet de tu preferencia a la mano y empieces a usar R en él. Practica².

A.1 Algunas ventajas de R y cosas no tan padres

A.1.1 Puntos a favor de R

- Todo el mundo lo usa. Quizá éste es el punto más a favor. Si mucha gente lo conoce y lo utiliza, hay más opciones de ayuda. Los sitios de StackOverflow en inglés y en español son excelentes para pedir apoyo en R; los grupos de usuarios de Google son otra fuente muy buena. Entre más gente usa el programa; es más fácil obtener ayuda porque seguro alguien más tuvo hace ya tiempo el mismo problema que tú.
- Todas las personas que trabajan en estadística publican sus métodos y su código en R (eso, claro, cuando publican sus métodos). Es raro encontrar *un nuevo método estadístico* en el mundo y que no se pueda usar, de alguna forma, en R.
- Dentro de los lenguajes de programación R es de los más sencillos. Quienes lo hicieron realmente se preocuparon por su público (de no especialistas) y en general desarrollan para él.
- R es gratis. Y en esta época de austeridad, cualquier ahorro es bueno. Que sea gratis no significa que no esté respaldado: existen versiones de R respaldadas por grandes compañías como Microsoft
- Todo lo que se hace en R es público. R no tiene métodos secretos ni es una caja negra. Todo lo que hace cada una de las funciones de R, cualquiera lo puede revisar, por completo.
- En R puedes hacer libros o notas ¡como este! donde guardes todo tu trabajo, reportes automatizados e incluso documentos interactivos para facilitar el análisis de datos.
- R puede hacer gráficas bonitas:

²La práctica hace al maestro



RESULTADOS DE LA SIMULACIÓN

Por supuesto, no todo es miel sobre hojuelas con R. Particularmente, algunos de los problemas con el lenguaje:

- La curva de aprendizaje es mucho más empinada que para otros programas estadísticos (como Stata, SAS o SPSS) ¡particularmente si es tu primera vez programando!
- La mayor parte de las personas que trabajan en R no son programadores de verdad. Gran parte del código que te puedes encontrar **en el mundo real** está escrito con prisa para salir del aprieto sin mucha planeación, con pocos comentarios, falta de control de versiones y pocas herramientas de revisión. ¡Internet está lleno de criaturas espantosas escritas en R!
- R de ninguna manera es veloz por lo que algunos programas (lo veremos en simulación) pueden ser extremadamente lentos.

A.2 Bienvenidx a R, Camp Pontanezen (sí, así se llama esta versión)

R es un lenguaje de cómputo y un programa estadístico libre, gratuito, de programación funcional (¿qué es eso?), orientado a objetos (*what??*) que mutó a

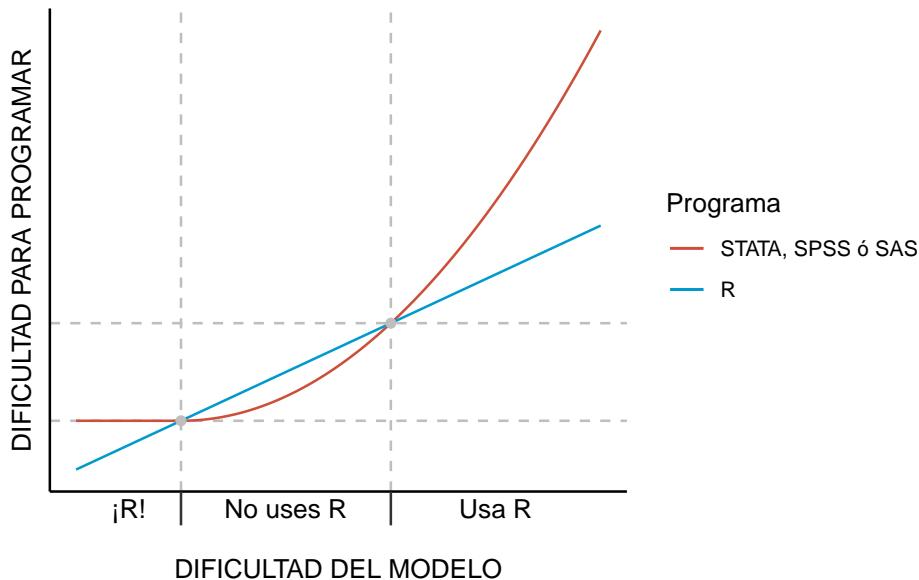


Figure A.2: La curva de aprendizaje de ‘R’ es más empinada pero después de un rato vale la pena

partir de otros dos lenguajes conocidos como **Scheme** y **S**³. El primero de estos fue desarrollado en el MIT por Sussman y Steele mientras que el segundo surgió en los laboratorios Bell⁴ creado por Becker, Wilks y Chambers. R nació en junio de 1995 a partir del trabajo de Ross Ihaka y Robert Gentleman⁵.

Desde su creación, la mayor parte del desarrollo de R ha sido trabajo completamente voluntario de la Fundación R, del equipo de R Core y de miles de usuarios que han creado funciones específicas para R conocidas como paquetes (**packages**). Actualmente el repositorio más importante de R, CRAN, contiene más de 16000 paquetes con distintas funciones para hacer ¡lo que quieras!

Como todo el trabajo en R es voluntario hace falta:

1. Una homologación en los métodos. Puedes encontrar varias funciones *que supuestamente hacen exactamente lo mismo* (como es el caso de `emojifont`, `fontemoji` y `emoGG` para graficar usando emojis).
2. Estandarizar la notación. Algunos paquetes como aquellos del **tidyverse** (veremos más adelante) utilizan **pipes** (`%>%`); estos sólo funcionan en el **tidyverse** pero no fuera del mismo.

³De ahí que se llame R porque la R es una mejor letra que la S (todos lo sabemos) -Atte. Rodrigo, el autor de este documento.

⁴Mejor conocidos ahora como AT&T, la compañía celular que nunca tiene señal.

⁵Sus nombres empiezan con la letra R ¿coincidencia?

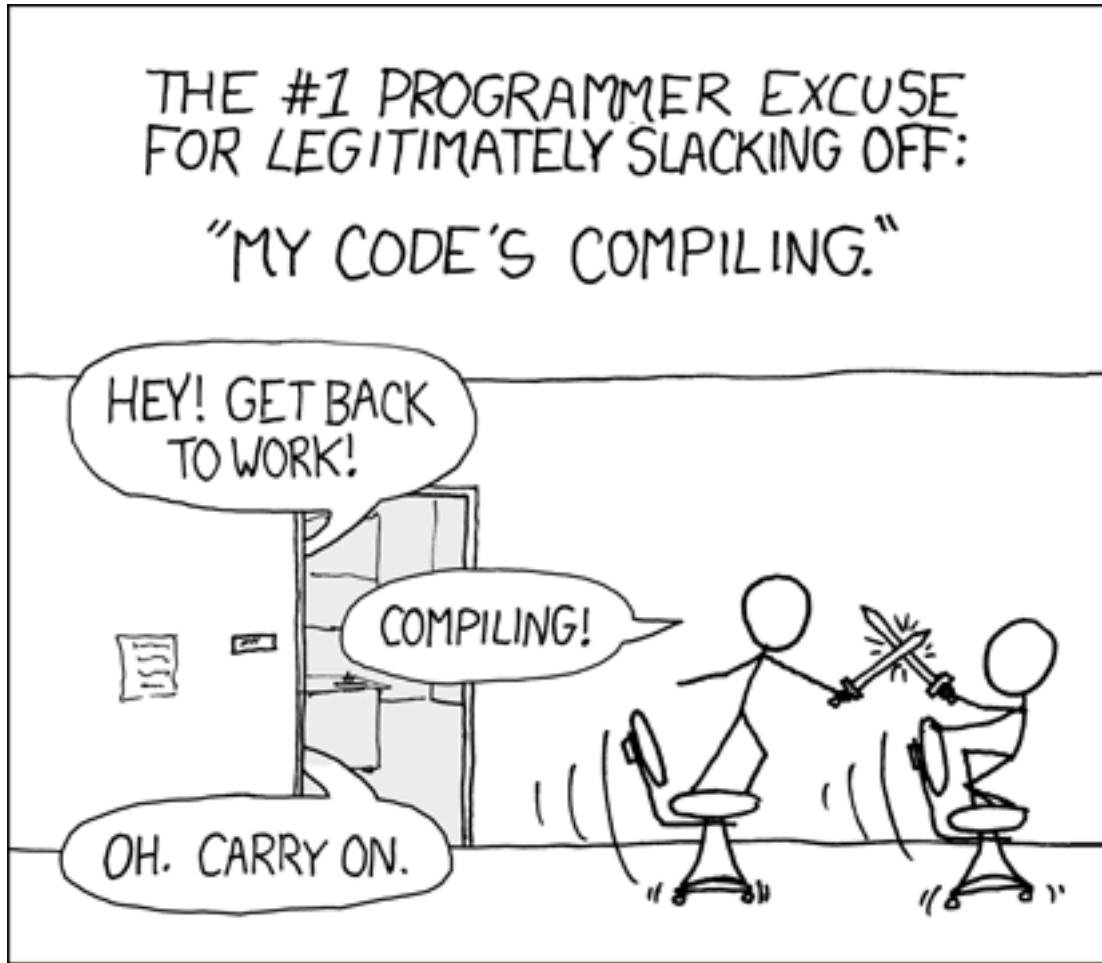


Figure A.3: 'R' puede ser muy lento pero eso te da oportunidad de hacer otras cosas ;).

Sin embargo, también es una gran ventaja que sean los usuarios de R quienes guían su desarrollo. El lenguaje va mutando según peticiones de las personas que lo usan. Si hay algo que te gustaría R tuviera y aún no existe ¡lo puedes proponer!

A.3 Instalando cosas

A.3.1 Instalación de R

A lo largo de estas notas estaré trabajando con: R version 4.1.0 (2021-05-18) *Camp Pontanezen*. La más reciente versión de R la puedes encontrar en CRAN. Para ello ve al sitio y selecciona tu plataforma.

Nota usuarios de Mac En algunas Mac, al abrir R, aparece el siguiente mensaje de advertencia: `During startup - Warning messages: 1: Setting LC_CTYPE failed [...]` para solucionarlo ve a **Aplicaciones** y abre **Terminal**. Copia y pega en ella el siguiente texto: `defaults write org.R-project.R force.LANG en_US.UTF-8` Da enter, cierra la **Terminal** y reinicia R.

- En el caso de Windows da clic en **Download R for Windows** y luego en **install R for the first time**. Finalmente, ejecuta el instalable que aparece al dar click en **Download R 4.1.0 for Windows**.

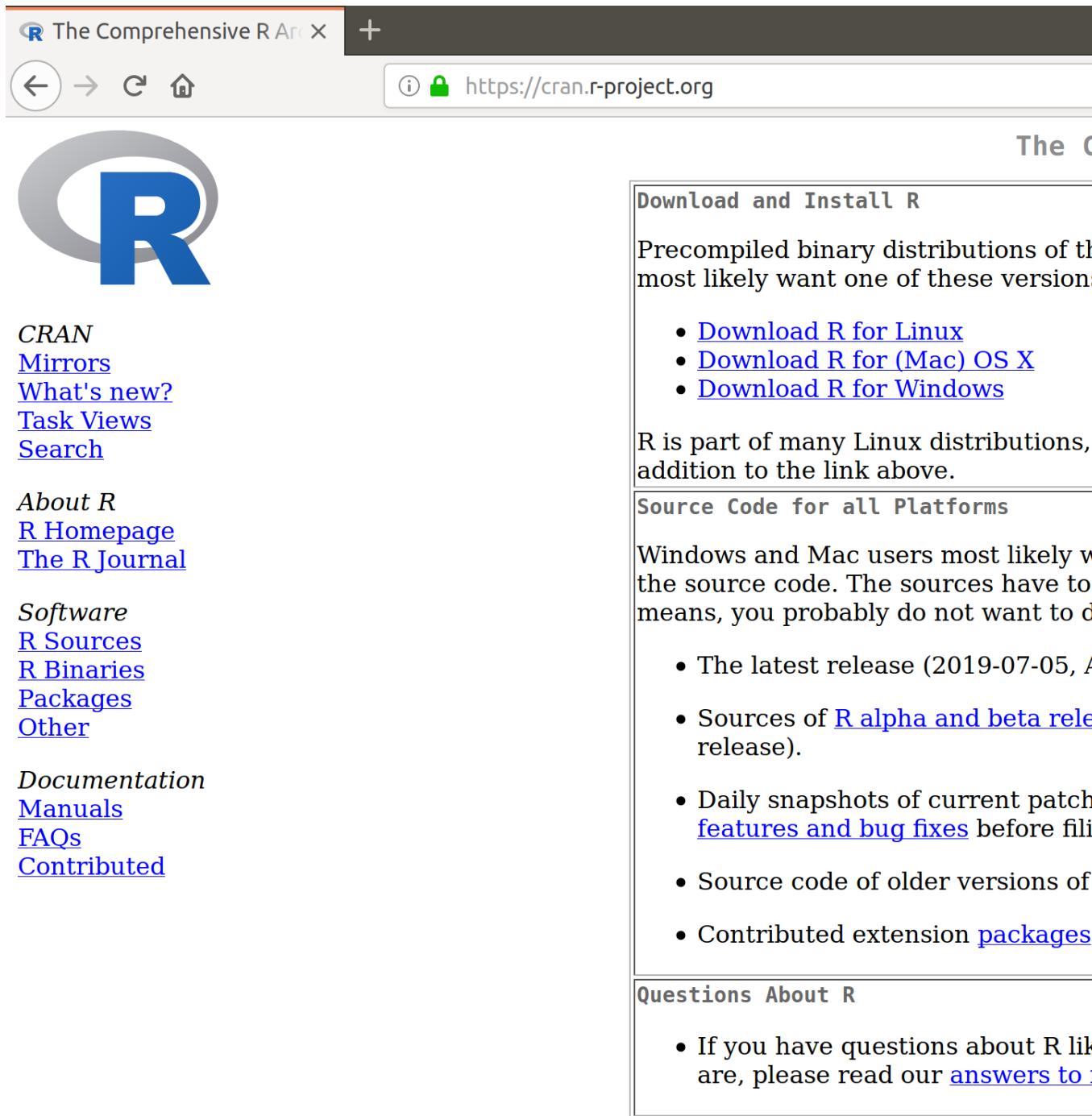
Para este curso pudiera ser que requirieras las herramientas de desarrollador Rtools.

- En el caso de Mac selecciona **Download R for (Mac) OS X** y luego elige **R-4.1.0.pkg**. En Mac puede que necesites instalar adicionalmente XQuartz (según tu versión de Mac). Si tu Mac es una versión suficientemente antigua, sigue las instrucciones específicas de CRAN.
- En el caso de Linux al elegir **Download R for Linux** tendrás la opción de buscar tu distribución específica. Al elegirla, aparecerán instrucciones para tu terminal de comandos; síguelas. En el caso de Linux, según los paquetes de R que elijamos instalar en la computadora requerirás instalar paquetería adicional para tu distribución de Linux. R te informará de la paquetería necesaria conforme la requiera.

Si tienes problemas para instalar puedes usar RStudio Cloud.

A.4 Instalación de RStudio

RStudio es una interfaz gráfica (IDE) para R. Puedes pensar a R como el *Bloc de Notas* y a RStudio como *Word*. El *Bloc* tiene todas las capacidades que necesitas para poder escribir; empero, es muchísimo mejor trabajar tus *papers* en *Word*. De la misma manera, R tiene todas las capacidades para hacer estadística *pero un formato horrible* y RStudio se ha convertido en la más popular forma de usar



The screenshot shows the official CRAN (Comprehensive R Archive Network) website. At the top, there's a navigation bar with icons for back, forward, search, and home, along with the URL <https://cran.r-project.org>. Below the header is the large R logo. The main content area is divided into several sections:

- CRAN**: Links to [Mirrors](#), [What's new?](#), [Task Views](#), and [Search](#).
- About R**: Links to [R Homepage](#) and [The R Journal](#).
- Software**: Links to [R Sources](#), [R Binaries](#), [Packages](#), and [Other](#).
- Documentation**: Links to [Manuals](#), [FAQs](#), and [Contributed](#).
- Download and Install R**: A section for precompiled binary distributions, with links to [Download R for Linux](#), [Download R for \(Mac\) OS X](#), and [Download R for Windows](#).
- Source Code for all Platforms**: A section for source code, noting it's part of many Linux distributions. It lists several bullet points about different types of releases and source code availability.
- Questions About R**: A section with a single bullet point about reading answers to questions.

Figure A.4: Oficialmente, la página de ‘R’ es de las páginas más feas del mundo.
¡No te dejes llevar por las apariencias!



Figure A.5: RStudio es una empresa que se dedica a hacer cosas para R.

R. Por supuesto que no es la única; algunas alternativas son Atom con ide-r, Eclipse con StatET y RKWard. En general es posible seguir estas notas sin que tengas RStudio pero, si es tu primera vez programando, no lo recomiendo.

Si ya tienes experiencia con lenguajes como Python, Javascript, Java ó alguno de los mil C que existen, no tendrás ningún problema usando el editor de tu preferencia.

Para descargar RStudio ve a su página y da clic en Download RStudio. Baja tu pantalla hasta donde dice **Installers for Supported Platforms** y elige tu plataforma: Windows, Mac OS X ó tu sabor de Linux preferido. Una vez descargado el archivo, ábrelo y sigue las instrucciones que aparecen en pantalla.

A.5 Primeros pasos en R usando RStudio

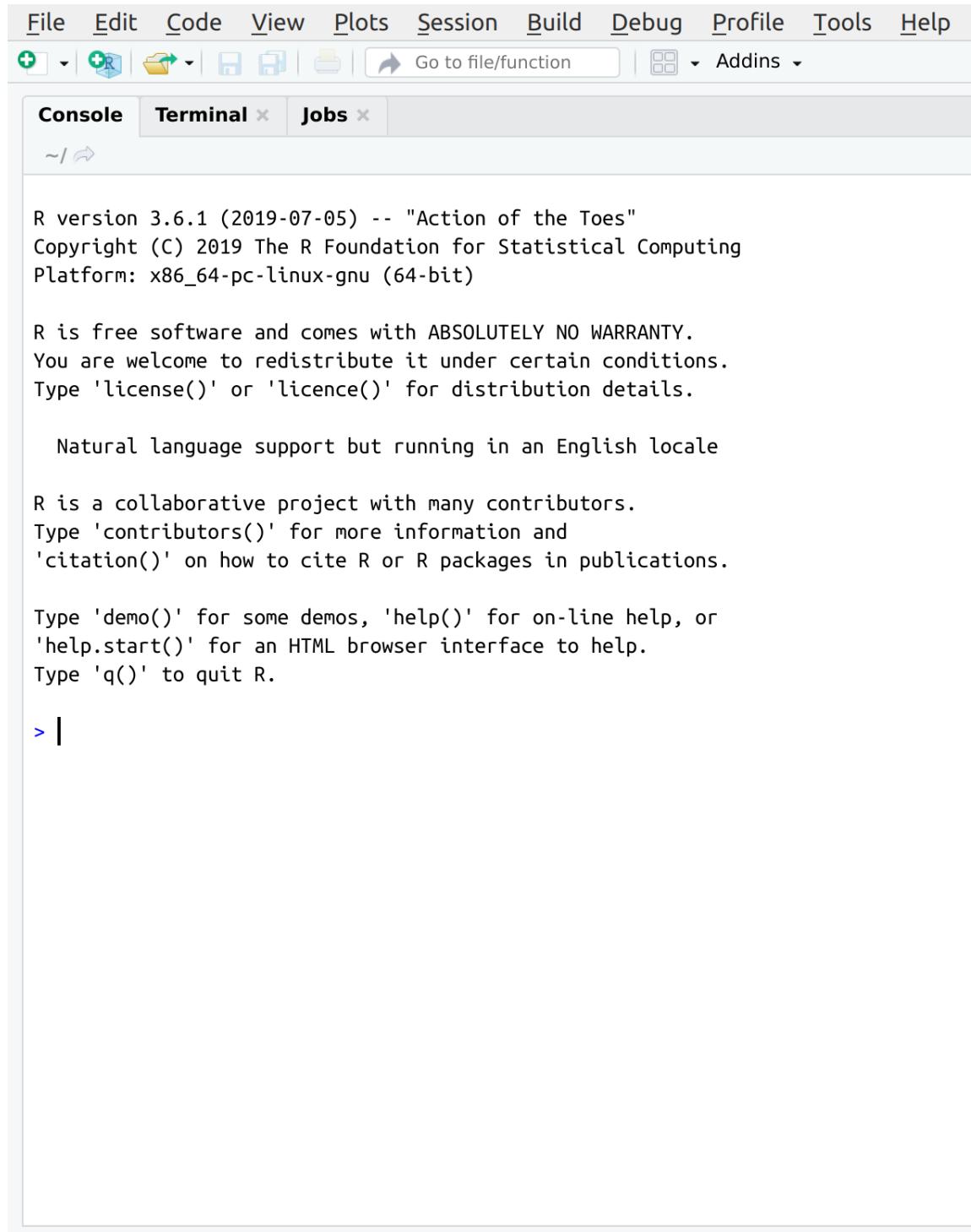
Una vez hayas instalado R y RStudio, abre RStudio⁶. Te enfrentarás a una pantalla similar a esta:

Si tu RStudio tiene sólo 3 páneles, como en mi caso, ve a la esquina superior izquierda (signo de hoja+) y elige un nuevo R Script

Tendrás, entonces, 4 páneles como se ve a continuación:

1. El primer panel (esquina inferior izquierda) es la **Consola**. Aquí es donde se ejecutan las acciones. Prueba escribir `2 + 3` en él y presiona enter. Aparece el resultado de la suma. Definitivamente, R es la calculadora que más trabajo cuesta instalar.
2. El segundo panel (esquina superior izquierda) es el panel con el **Script**. Aquí se escribe el programa pero no *se ejecuta*. Prueba escribir `10 + 9`. ¿Ves que no pasa nada? Lo que acabas de hacer es crear un programa que, cuando se ejecute, hará la suma de `10 + 9`. ¡Qué programa más aburrido! Sin embargo, no todo está perdido: presiona **CTRL+Enter** (**Cmd+Enter** en Mac) al final de la línea o bien da clic en **Run** y verás que, en la consola, aparece la instrucción y el resultado de la misma. El **Script** es una excelente fuente para tener un historial de lo que estás haciendo.
3. El tercer panel contiene el ambiente. Aquí aparecerán las variables que vayamos creando. Por ahora, para poner un ejemplo, importaremos el archivo `Example1.csv` (con valores simulados) disponible en Github dando clic en **Import Dataset** y **From Text (base)**. Selecciona el archivo y elige las opciones en la ventana de previsualización que hagan que se vea bien. Nota que una vez realizada la importación aparece en el panel derecho `Example1`. Al dar clic podrás ver la base de datos. Las bases de datos y variables que utilices durante tus análisis aparecerán en esa sección.
4. Para entender mejor lo que ocurre en el último de los páneles, lo mejor es trabajar con nuestra base. Escribe en la consola `plot(Example1)`. En el

⁶Si decidiste no instalar RStudio salta al final de esta sección.



The screenshot shows the RStudio interface with the 'Console' tab selected. The R console window displays the standard R startup message, which includes information about the version (R version 3.6.1), the date (2019-07-05), the title ("Action of the Toes"), copyright information, the platform (x86_64-pc-linux-gnu), and the license details. It also mentions natural language support and collaborative project contributors. At the bottom of the console, there is a blue cursor indicating where the user can type commands.

```
R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

Figure A.6: La primera vez que abres RStudio

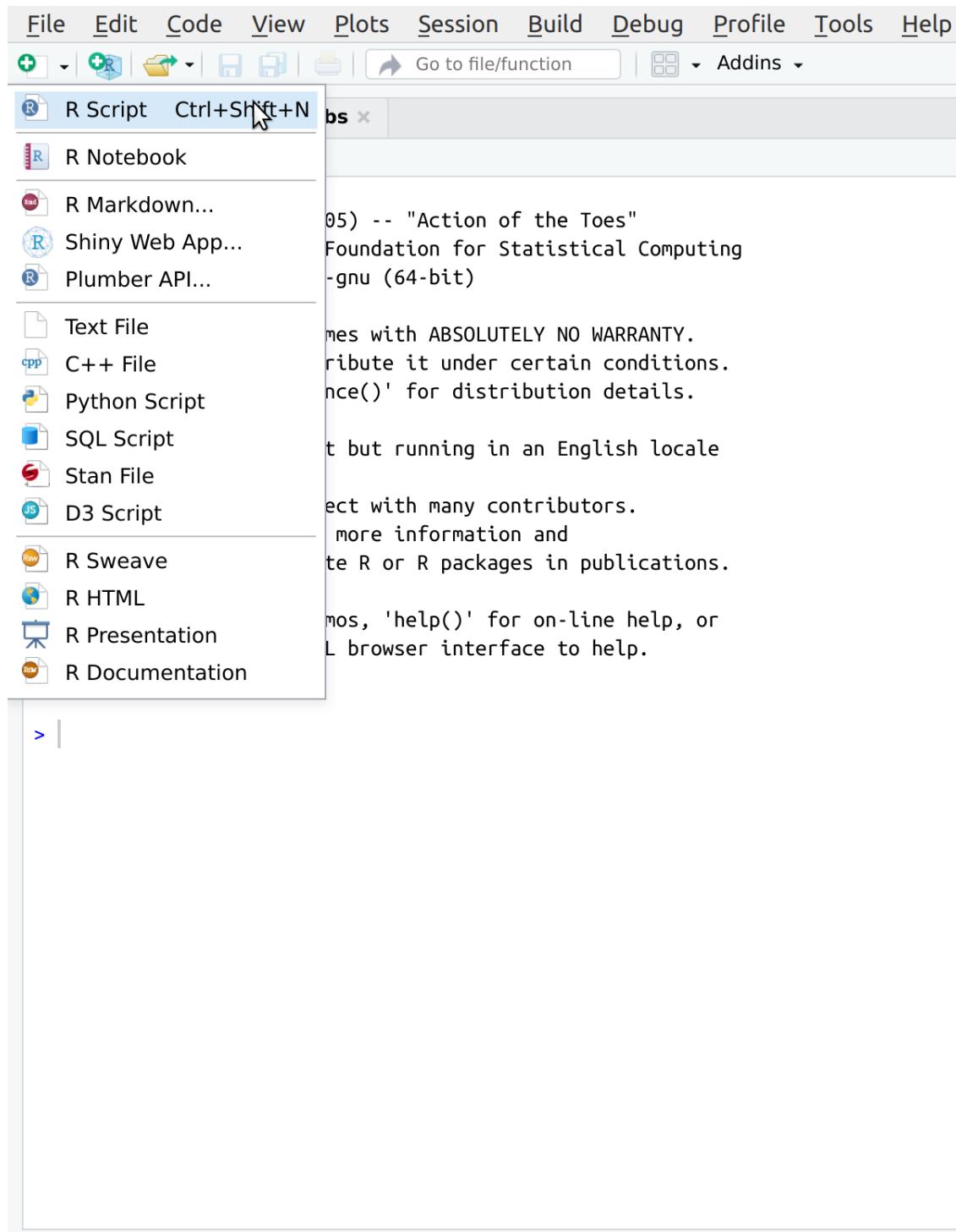


Figure A.7: Elige hoja+ para crear un nuevo archivo

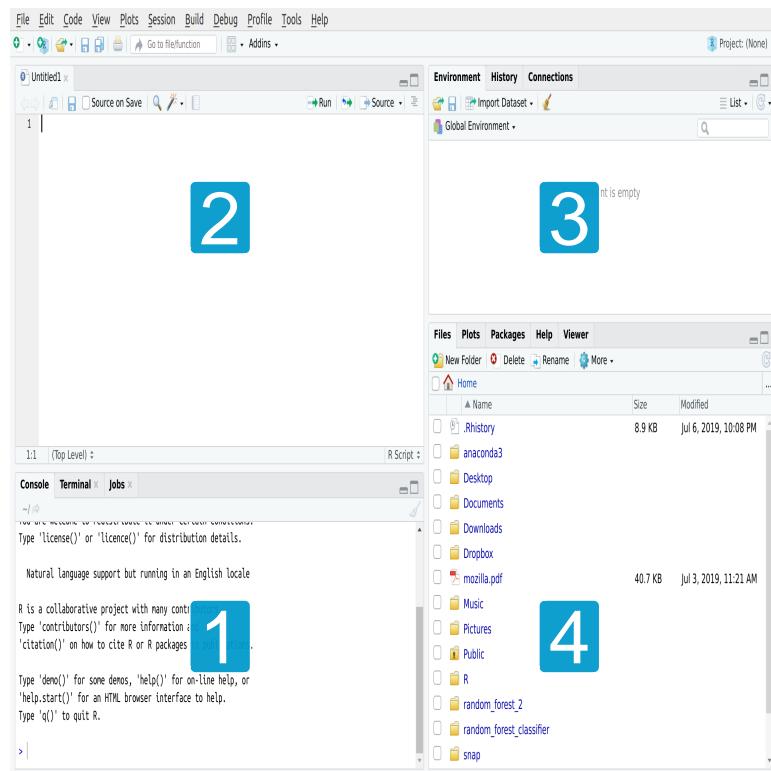


Figure A.8: RStudio <3

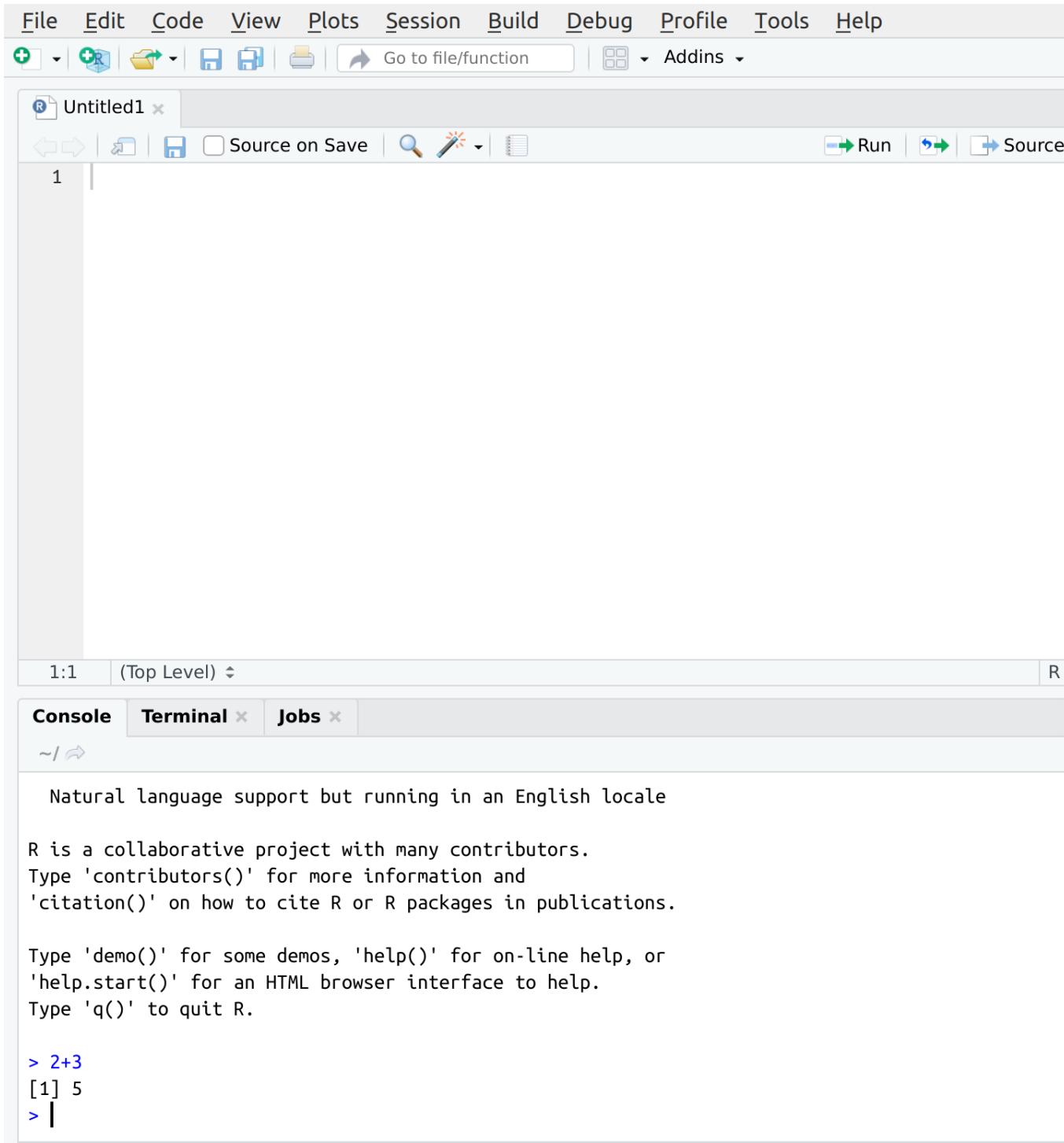


Figure A.9: La consola de ‘R’ es la calculadora más difícil de instalar que existe.

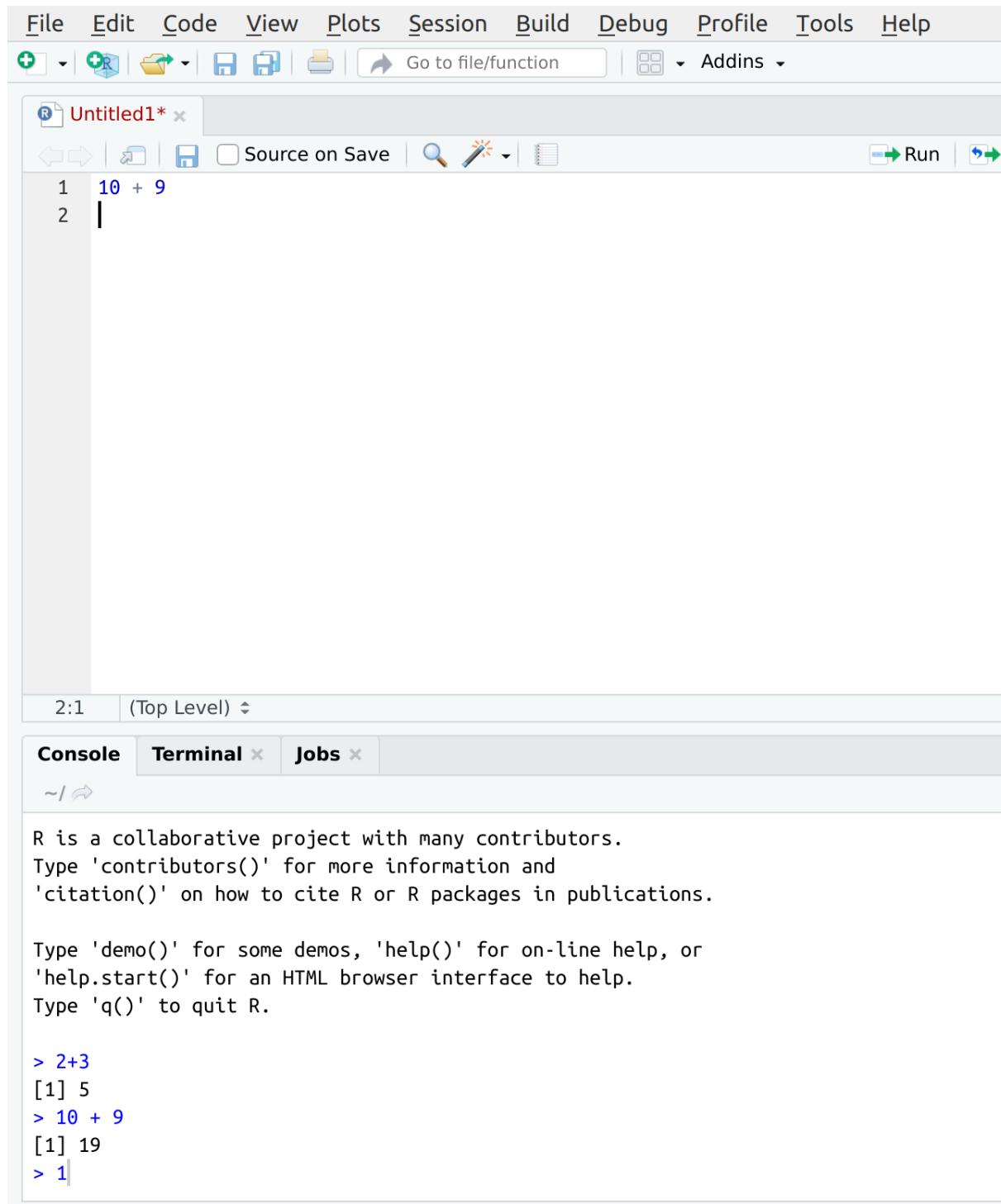


Figure A.10: El ‘Script’ sirve para salvar las instrucciones en el orden en que las vas a ejecutar.

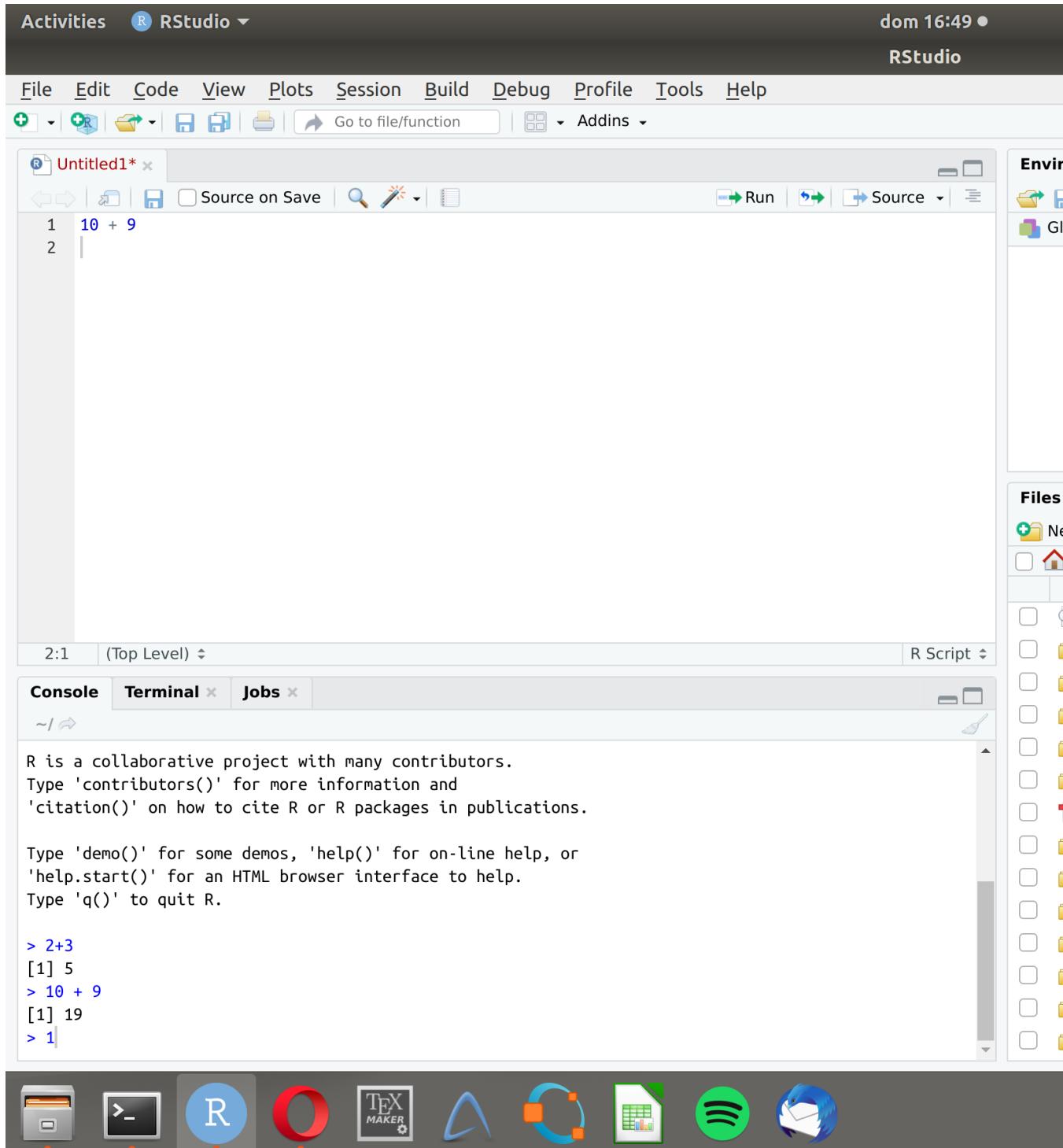


Figure A.11: El ‘Ambiente’ muestra las variables (incluyendo bases de datos) que estás utilizando en este momento. A diferencia de otros programas estadísticos (o sea ‘Stata’) en ‘R’ es posible tener múltiples bases de datos abiertas a la vez.

cuarto pánel aparecerá una gráfica. El cuarto de los pátentes para nosotros tendrá esa utilidad: mostrará las gráficas que hagamos así como la ayuda. Para ver la ayuda para las instrucciones de R puedes escribir ?. Prueba teclear ?plot en la consola. El signo de interrogación es un `help()` que muestra las instrucciones para usar una función.

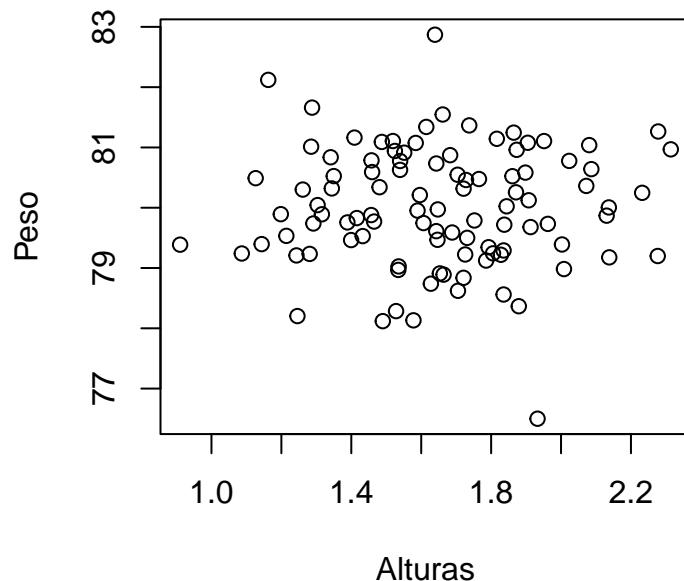


Figure A.12: La gráfica que aparece de hacer un ‘plot’ de la base de datos de ejemplo.

Mi sugerencia personal es que escribas todo lo que haces en el `Script` y que sólo utilices la consola para verificar valores. De esta manera podrás almacenar todas las instrucciones ejecutadas y volver a ellas cuando se requieran. Por último te sugiero utilizar # gatos para comentar tu código. Así, el código anterior lo podrías ver en la consola como:

```
#Aquí pruebo cómo R hace las sumas
10 + 9
```

Comenta. Comenta. Comenta, por favor. Tu ser del futuro que regrese a sus archivos de R un mes después de haberlos hecho te lo agradecerá (y tu profe también).

Finalmente y como aclaración para estas notas, el código de R aparece como:

```
#Esto es código de R
7 - 2
```

Mientras que los resultados de evaluar en R se ven con #:

```
## [1] 5
```

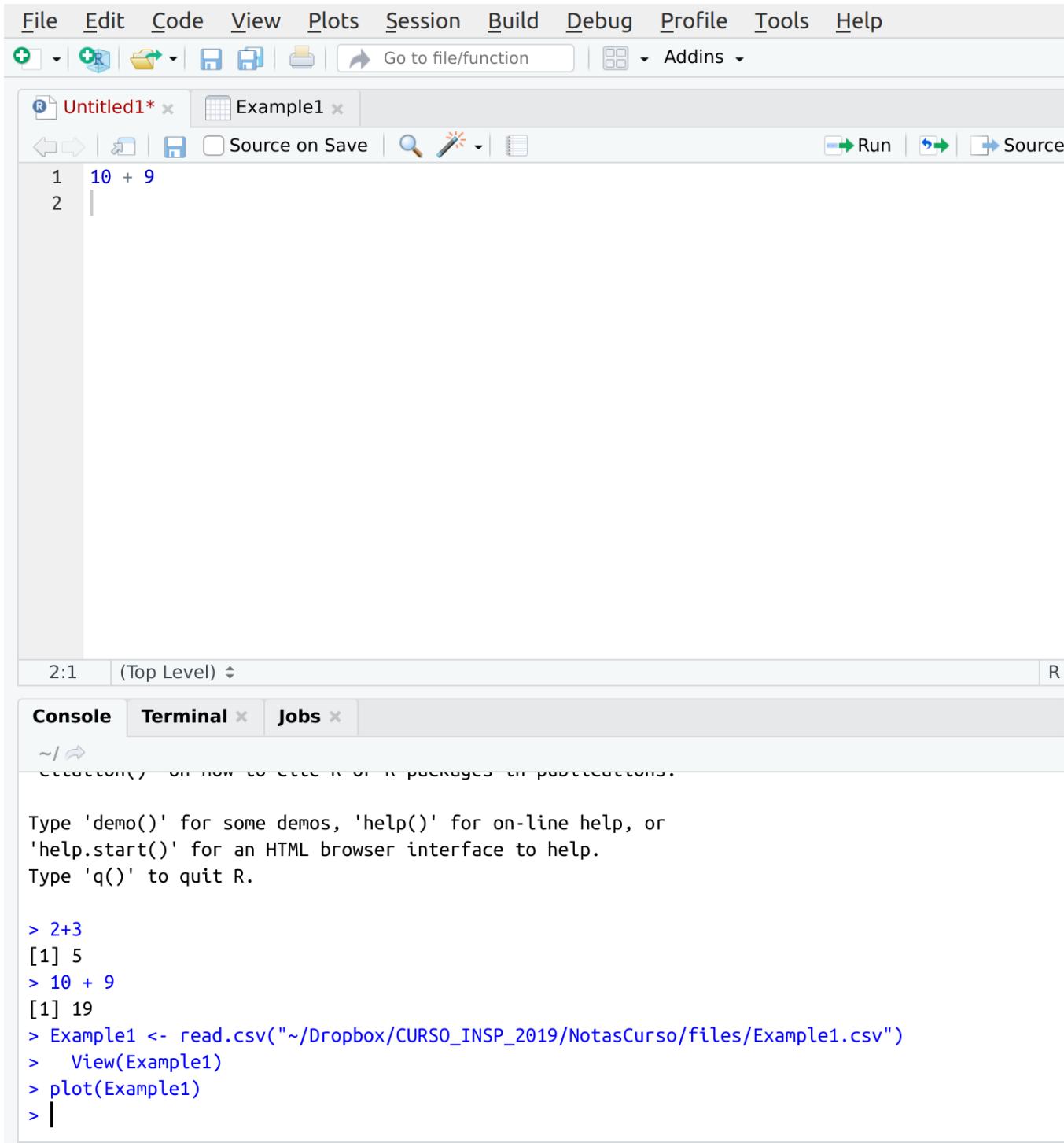


Figure A.13: El cuarto panel muestra respectivamente las gráficas y la ayuda.

Así, la evaluación con su resultado se ve de la siguiente forma:

#Esto es código de R

7 - 2

[1] 5

A.6 Cálculos numéricos

R sirve como calculadora para las operaciones usuales. En él puedes hacer sumas,

#Esto es una suma en R

12 + 31

[1] 43

restas,

#Esto es una resta en R

3 - 4

[1] -1

multiplicaciones,

#Esto es una multiplicación en R

7*8

[1] 56

divisiones,

#Esto es una división en R

4/2

[1] 2

sacar logaritmos naturales ln,

#Para sacar logaritmo usas el comando log

log(100)

[1] 4.60517

o bien logaritmos en cualquier base,⁷

#Puedes especificar la base del logaritmo con base

log(100, base = 10)

[1] 2

también puedes elevar a una potencia (por ejemplo hacer 6^3),

⁷Recuerda que un logaritmo base a te dice a qué potencia b tuve que elevar a para llegar a b . Por ejemplo $\log_{10}(100) = 2$ te dice que para llegar al 100 tuviste que hacer 10^2 .



Figure A.14: Ada Lovelace (1815-1852), la primera en diseñar un algoritmo computacional ¡y sin tener computadoras!

```
#Así se calculan potencias
6^3
```

```
## [1] 216
```

calcular la exponencial e ,

```
#Para exponentiales puedes usar exp
exp(1)
```

```
## [1] 2.718282
```

o bien exponenciar cualquier variable e^{-3} ,

```
#O bien exponentiales específicas, e^-3
exp(-3)
```

```
## [1] 0.04978707
```

también puedes usar el número π .

```
#Cálculo de pi
pi
```

```
## [1] 3.141593
```

No olvides que R usa el orden de las operaciones de matemáticas. Siempre es de izquierda a derecha con las siguientes excepciones:

1. Primero se evalúa lo que está entre paréntesis.
2. En segundo lugar se calculan potencias.
3. Lo tercero en evaluarse son multiplicaciones y divisiones.
4. Finalmente, se realizan sumas y restas.

Por ejemplo, en la siguiente ecuación

$$2 - 2 \cdot \frac{(3^4 - 9)}{(5 + 4)}$$

se resuelven primero los paréntesis $(3^4 - 9) = 81 - 9 = 72$ y $(5 + 4) = 9$; luego se resuelve la división: $\frac{72}{9} = 8$, se multiplica por el 2: $2 \cdot 8 = 16$ y finalmente se hace la resta: $2 - 16 = -14$.

A.6.1 Ejercicio

Determina, sin evaluar, los resultados de los siguientes segmentos de código:

```
#Primer ejercicio
(9 - 3)^2 * (2 - 1) - 6
```

```
#Segundo ejercicio
6 * 2 / (7 - 3) * 5

#Tercer ejercicio
2 * 3 ^ 2 * 2 / (5 - 4) * 1 / 10
```

Evalúa para comprobar tu respuesta.

A.6.2 Ejercicio

Calcula el área y el perímetro de un círculo de radio 5. Recuerda que la fórmula del área es $\pi \cdot r^2$ donde r es el radio; mientras que la del perímetro es: $\pi \cdot d$ donde d es el diámetro (= dos veces el radio).

A.6.3 Respuestas

```
## Área = 78.5398163397448
## Perímetro = 31.4159265358979
```

A.7 Variables

R es un programa orientado a objetos; esto quiere decir que R almacena la información en un conjunto de variables que pueden tener diferentes **clases** y opera con ellos según su clase. Por ejemplo, un conjunto de caracteres, entre comillas, es un **Character** (R lo piensa como texto)

```
#Un conjunto de caracteres es un char
"Hola"

## [1] "Hola"
```

Un número (por ejemplo 2 tiene clase **numeric**)⁸. Hay que tener mucho cuidado con combinar floats con **Strings**:

```
#Código que sí funciona porque ambos son números
2 + 4

## [1] 6

#Código que no funciona porque uno es carácter
2 + "4"

## Error in 2 + "4": non-numeric argument to binary operator
```

Si lo piensas, este último error ¡tiene todo el sentido! no puedes sumar un número a un texto. ¿O qué significaría 'Felices' * 4 ?

⁸Puede ser **float**, **int**, **double** pero no nos preocuparemos por eso.

Diagram for the computation by the Engine of the Numbers of Bernoulli.

Number of Operation.	Nature of Operation.	Variables acted upon.	Variables receiving results.	Indication of change in the value on any Variable.	Statement of Results.	Data.								
						1V_1	1V_2	1V_3	0V_4	0V_5	0V_6	0V_7	0V_8	0V_9
1	\times	$^1V_2 \times ^1V_3$	$^1V_4, ^1V_5, ^1V_6$	$\left\{ \begin{array}{l} ^1V_2 = ^1V_2 \\ ^1V_3 = ^1V_3 \\ ^1V_4 = ^2V_4 \\ ^1V_1 = ^1V_1 \end{array} \right\}$	$= 2n$...	2	n	2n	2n	2n			
2	-	$^1V_4 - ^1V_1$	2V_4	$\left\{ \begin{array}{l} ^1V_4 = ^2V_4 \\ ^1V_1 = ^1V_1 \end{array} \right\}$	$= 2n - 1$	1	2n - 1					
3	+	$^1V_5 + ^1V_1$	2V_5	$\left\{ \begin{array}{l} ^1V_5 = ^2V_5 \\ ^1V_1 = ^1V_1 \end{array} \right\}$	$= 2n + 1$	1	2n + 1				
4	+	$^2V_5 + ^2V_4$	$^1V_{11}$	$\left\{ \begin{array}{l} ^2V_5 = ^0V_5 \\ ^2V_4 = ^0V_4 \end{array} \right\}$	$= \frac{2n - 1}{2n + 1}$	0	0	
5	+	$^1V_{11} - ^1V_2$	$^2V_{11}$	$\left\{ \begin{array}{l} ^1V_{11} = ^2V_{11} \\ ^1V_2 = ^1V_2 \end{array} \right\}$	$= \frac{1}{2} \cdot \frac{2n - 1}{2n + 1}$...	2	
6	-	$^0V_{13} - ^2V_{11}$	$^1V_{13}$	$\left\{ \begin{array}{l} ^2V_{11} = ^0V_{11} \\ ^0V_{13} = ^1V_{13} \end{array} \right\}$	$= -\frac{1}{2} \cdot \frac{2n - 1}{2n + 1} = A_0$	
7	-	$^1V_3 - ^1V_1$	$^1V_{10}$	$\left\{ \begin{array}{l} ^1V_3 = ^1V_3 \\ ^1V_1 = ^1V_1 \end{array} \right\}$	$= n - 1 (= 3)$	1	...	n
8	+	$^1V_2 + ^0V_7$	1V_7	$\left\{ \begin{array}{l} ^1V_2 = ^1V_2 \\ ^0V_7 = ^1V_7 \end{array} \right\}$	$= 2 + 0 = 2$...	2	2	
9	+	$^1V_6 + ^1V_7$	$^3V_{11}$	$\left\{ \begin{array}{l} ^1V_6 = ^1V_6 \\ ^0V_{11} = ^3V_{11} \end{array} \right\}$	$= \frac{2n}{2} = A_1$	2n	2
10	\times	$^1V_{21} \times ^3V_{11}$	$^1V_{12}$	$\left\{ \begin{array}{l} ^1V_{21} = ^1V_{21} \\ ^3V_{11} = ^3V_{11} \end{array} \right\}$	$= B_1 \cdot \frac{2n}{2} = B_1 A_1$	
11	+	$^1V_{12} + ^1V_{13}$	$^2V_{13}$	$\left\{ \begin{array}{l} ^1V_{12} = ^0V_{12} \\ ^1V_{13} = ^2V_{13} \end{array} \right\}$	$= -\frac{1}{2} \cdot \frac{2n - 1}{2n + 1} + B_1 \cdot \frac{2n}{2}$	
12	-	$^1V_{10} - ^1V_1$	$^2V_{10}$	$\left\{ \begin{array}{l} ^1V_{10} = ^2V_{10} \\ ^1V_1 = ^1V_1 \end{array} \right\}$	$= n - 2 (= 2)$	1
13	-	$^1V_6 - ^1V_1$	2V_6	$\left\{ \begin{array}{l} ^1V_6 = ^2V_6 \\ ^1V_1 = ^1V_1 \end{array} \right\}$	$= 2n - 1$	1	2n - 1		
14	+	$^1V_1 + ^1V_7$	2V_7	$\left\{ \begin{array}{l} ^1V_1 = ^1V_1 \\ ^1V_7 = ^2V_7 \end{array} \right\}$	$= 2 + 1 = 3$	1	3		
15	+	$^2V_6 + ^2V_7$	1V_8	$\left\{ \begin{array}{l} ^2V_6 = ^2V_6 \\ ^2V_7 = ^2V_7 \end{array} \right\}$	$= \frac{2n - 1}{3}$	2n - 1	3	$2n - 1$	3
16	\times	$^1V_8 \times ^3V_{11}$	$^4V_{11}$	$\left\{ \begin{array}{l} ^1V_8 = ^0V_8 \\ ^3V_{11} = ^4V_{11} \end{array} \right\}$	$= \frac{2n}{2} \cdot \frac{2n - 1}{3}$	0	...
17	-	$^2V_6 - ^1V_1$	3V_6	$\left\{ \begin{array}{l} ^2V_6 = ^3V_6 \\ ^1V_1 = ^1V_1 \end{array} \right\}$	$= 2n - 2$	1	2n - 2			
18	+	$^1V_1 + ^2V_7$	3V_7	$\left\{ \begin{array}{l} ^1V_1 = ^1V_1 \\ ^2V_7 = ^3V_7 \end{array} \right\}$	$= 3 + 1 = 4$	1	4		
19	+	$^3V_6 + ^3V_7$	1V_9	$\left\{ \begin{array}{l} ^3V_6 = ^3V_6 \\ ^3V_7 = ^3V_7 \end{array} \right\}$	$= \frac{2n - 2}{4}$	2n - 2	4		$2n - 4$
20	\times	$^1V_9 \times ^4V_{11}$	$^6V_{11}$	$\left\{ \begin{array}{l} ^1V_9 = ^0V_9 \\ ^4V_{11} = ^6V_{11} \end{array} \right\}$	$= \frac{2n}{2} \cdot \frac{2n - 1}{3} \cdot \frac{2n - 2}{4} = A_3$	0	...
21	\times	$^1V_{22} \times ^5V_{11}$	$^0V_{12}$	$\left\{ \begin{array}{l} ^1V_{22} = ^1V_{22} \\ ^0V_{12} = ^2V_{12} \end{array} \right\}$	$= B_3 \cdot \frac{2n}{2} \cdot \frac{2n - 1}{3} \cdot \frac{2n - 2}{4} = B_3 A_3$
22	+	$^2V_{12} + ^2V_{13}$	$^3V_{13}$	$\left\{ \begin{array}{l} ^2V_{12} = ^0V_{12} \\ ^2V_{13} = ^3V_{13} \end{array} \right\}$	$= A_0 + B_1 A_1 + B_3 A_3$
23	-	$^3V_{10} - ^1V_1$	$^3V_{10}$	$\left\{ \begin{array}{l} ^3V_{10} = ^3V_{10} \\ ^1V_1 = ^1V_1 \end{array} \right\}$	$= n - 3 (= 1)$	1
Here follows a repetition of Operations thirteen.														
24	+	$^4V_{13} + ^0V_{24}$	$^1V_{24}$	$\left\{ \begin{array}{l} ^4V_{13} = ^0V_{13} \\ ^0V_{24} = ^1V_{24} \end{array} \right\}$	$= B_7$
25	+	$^1V_1 + ^1V_3$	1V_3	$\left\{ \begin{array}{l} ^1V_1 = ^1V_1 \\ ^1V_3 = ^1V_3 \end{array} \right\}$	$= n + 1 = 4 + 1 = 5$	1	...	$n + 1$	0	0	0	0

Figure A.15: El algoritmo diseñado por Ada Lovelace.

La magia de R comienza con que puedes almacenar valores en variables. Por ejemplo, podemos asignar un valor a una variable:

```
#Asignamos x = 10
x <- 10
```

La flecha de asignación funciona para ambos lados pero no se recomienda usarla al revés:

```
#Asignamos x = 10
10 -> x
```

Hay dos formas de asignar valores, una es con la flecha de asignación \leftarrow (o \rightarrow si quieres ver el mundo arder) y otra con el signo de igual:

```
#Podemos asignar valores con el signo de =
y = 6
```

Nota que, cuando realizamos operaciones, la asignación es la última que se realiza:

```
#Aquí z = 106
z <- y + x^2
```

Los valores que fueron asignados en las variables, R los recuerda y es posible calcular con ellos:

```
#Podemos realizar una suma
x + y
```

```
## [1] 16
#O bien podemos realizar una multiplicación
3*y - x
```

```
## [1] 8
```

Podemos preguntarnos por el valor de las variables numéricas mediante los operadores == (sí, son dos iguales), != (que es un \neq), >, \geq , \leq y <:

```
#Podemos preguntarnos si x vale 4
x == 4
```

```
## [1] FALSE
```

El operador de asignación también se puede utilizar al revés $2 \rightarrow x$
pero no lo hagas, por favor.

Nota que no estamos asignando el valor de x:

```
x
```

```
## [1] 10
```

Podemos preguntarnos por diferencia:

```
x != 4
```

```
## [1] TRUE
```

Así como por mayores, menores incluyendo posibles igualdades (*i.e.* los casos \geq y \leq)

```
#Nos preguntamos si x > y
```

```
x > y
```

```
## [1] TRUE
```

```
#Nos preguntamos si x >= 10
```

```
x >= 10
```

```
## [1] TRUE
```

```
#Nos preguntamos si y < 6
```

```
y < 6
```

```
## [1] FALSE
```

```
#O bien si y <= 6
```

```
y <= 6
```

```
## [1] TRUE
```

En todos los casos los resultados han sido TRUE ó FALSE. La clase de variables que toma valores TRUE ó FALSE se conoce como booleana. Hay que tener mucho cuidado con ellas porque, puedes acabar con resultados muy extraños:

```
#MALAS PRÁCTICAS, NO HAGAS ESTO
```

```
#Cuando lo usas como número TRUE vale 1
```

```
100 + TRUE
```

```
## [1] 101
```

```
#MALAS PRÁCTICAS, NO HAGAS ESTO
```

```
#Cuando lo usas como número FALSE vale 0
```

```
6*FALSE
```

```
## [1] 0
```

Aquí puedes encontrar una lista de malas prácticas en computación a evitar.

Finalmente, nota que es posible reescribir una variable y cambiar su valor:

```
#Aquí x vale 10, como antes
```

```
x
```

```
## [1] 10
```

```
#Aquí cambianos el valor de x y valdrá 0.5
x <- 0.5
x
## [1] 0.5
```

A.7.1 Ejercicios

Determina el valor que imprime R en cada caso, sin que corras los siguientes pedazos de código. Despues, verifica tu respuesta con R:

```
#Primer ejercicio
x <- 100
y <- 3
x > y
```

```
#Segundo ejercicio
z <- (4 - 2)^3
z <- z + z + z
z
```

```
#Tercer ejercicio
x <- 3
y <- 2
z <- x * y
x <- 5
y <- 10
z
```

```
#Cuarto ejercicio
variable1 <- 1000
variable2 <- 100
variable3 <- variable1/variable2 <= 10
variable3
```

```
#Quinto ejercicio
"2" - 2
```

```
#Sexto ejercicio
(0.1 + 0.1 + 0.1) == 0.3
```

A.7.2 NIVEL 3

Determina, sin correr el programa, qué regresa la consola en este caso

```
x <- 2
x <- 5 + x -> y -> x
x <- x^2
x
```

Comprueba con la consola tus resultados; puede que encuentres respuestas poco intuitivas.

A.8 Observaciones sobre la aritmética de punto flotante

Si hiciste el penúltimo ejercicio (el cual, obviamente hiciste y comprobaste con la consola) podrás haber notado una trampa. Analicemos qué ocurre; quizá hicimos mal la suma

```
#Veamos si este lado está mal
(0.1 + 0.1 + 0.1)
```

```
## [1] 0.3
#O si éste es el que tiene la trampa
0.3
```

```
## [1] 0.3
```

Aparentemente no hay nada malo ¿qué rayos le pasa a R? La respuesta está en la aritmética de punto flotante. Podemos pedirle a R que nos muestre los primeros 100 dígitos de la suma $0.1 + 0.1 + 0.1$:

```
#Veamos qué pasa con la suma
options(digits = 22) #Cambiamos dígitos
(0.1 + 0.1 + 0.1) #Sumamos
```

```
## [1] 0.3000000000000000444089
```

El comando `options(digits = 22)` especifica que R debe imprimir en la consola 22 dígitos. No más.

¡Ahí está el detalle! R no sabe sumar. En general, ningún programa de computadora sabe hacerlo. Veamos otros ejemplos:

```
4.1 - 0.1 #Debería dar 4
```

```
## [1] 3.99999999999999555911
```

```
3/10      #Debería ser 0.3
```

```
## [1] 0.299999999999999888978
```

```
log(10^(12345), base = 10) #Debería dar 12345
```

```
## [1] Inf
```

El problema está en cómo las computadoras representan los números. Ellas escriben los números en binario. Por ejemplo, 230 lo representan como 11100110 mientras que el 7 es: 111. El problema de las computadoras radica en que éstas tienen una memoria finita por lo que números muy grandes como:

A.8. OBSERVACIONES SOBRE LA ARITMÉTICA DE PUNTO FLOTANTE205

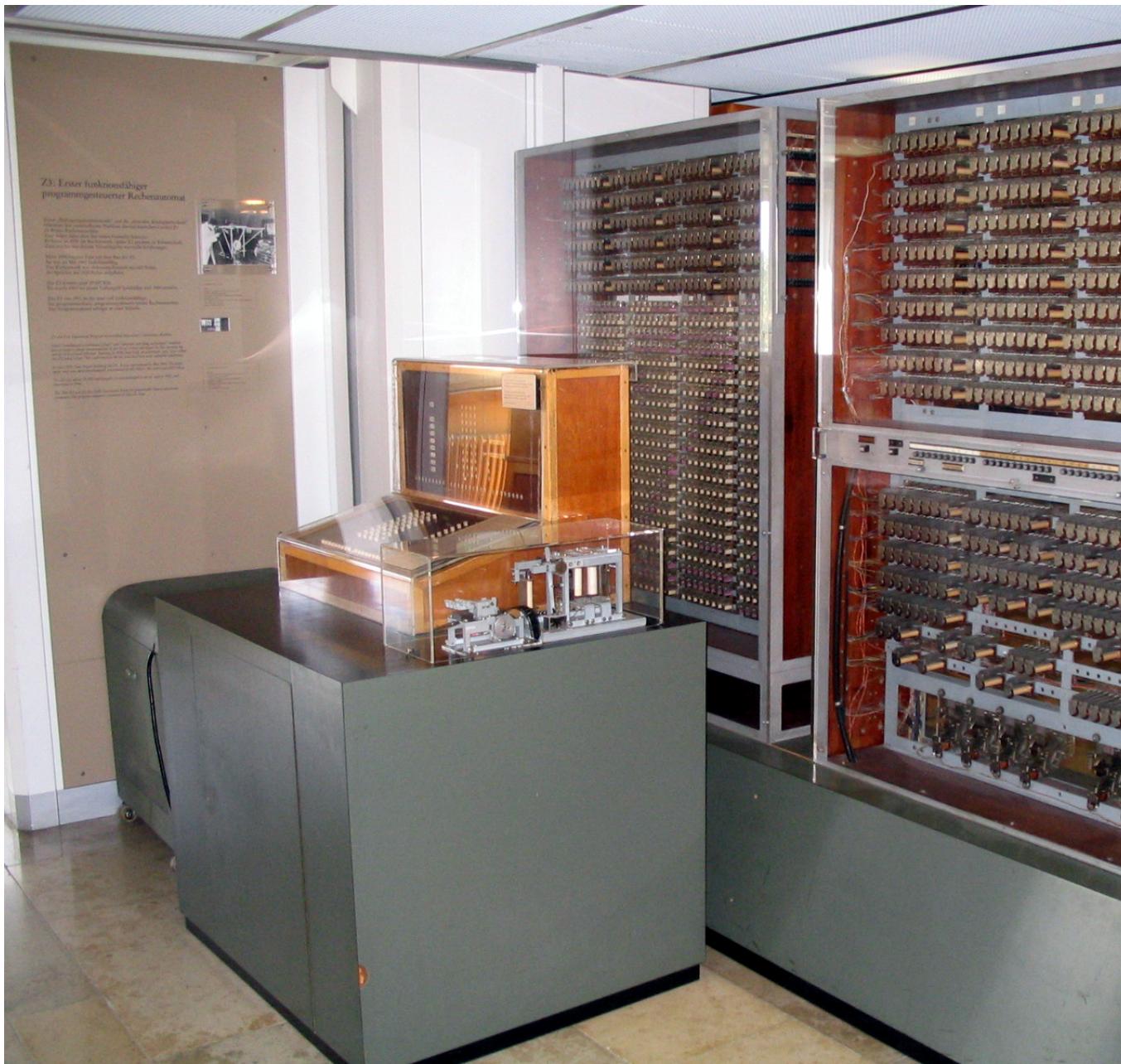


Figure A.16: Réplica de la Z3, la primer computadora con punto flotante (1941).

124765731467098372654176 la computadora hace lo mejor por representarlos eligiendo el más cercano:

```
#Nota la diferencia entre lo que le decimos a R
#y lo que resulta
x <- 124765731467098372654176
x
```

```
## [1] 124765731467098377420800
```

Un error de punto flotante en la vida real ocasionó en los años noventa, la explosión del cohete Ariane 5. Moraleja: hay que tener cuidado y respeto al punto flotante.

No olvides cambiar la cantidad de dígitos que deseas que imprima R en su consola de vuelta:

```
options(digits = 6) #Cambiamos dígitos
```

El mismo problema ocurre con números decimales cuya representación binaria es periódica; por ejemplo el $\frac{1}{10}$ en binario se representa como 0.00011001100110011.... Como es el cuento de nunca acabar con dicho número, R lo trunca y almacena sólo los primeros dígitos de ahí que, cada vez que escribes 0.1, R en realidad almacene el 0.100000000000000055511 que es *casi lo mismo* pero no es estrictamente igual. Hay que tener mucho cuidado con esta inexactitud de las computadoras (inexactitud estudiada por la rama de Análisis Numérico) pues puede generar varios resultados imprevistos.

A.8.1 ¿Cómo checar un if?

En general lo que hacen las computadoras para comparar valores es que verifican que, en valor absoluto, el error sea pequeño. Recuerda que el valor absoluto de x , $|x|$, regresa siempre el positivo:

$$|4| = 4 \quad \text{y} \quad |-8| = 8$$

Para verificar que algo es más o menos 0.3 suele usarse el valor absoluto⁹ de la siguiente manera:

```
abs( (0.1 + 0.1 + 0.1) - 0.3 ) < 1.e-6
```

```
## [1] TRUE
```

donde 1.e-6 es notación corta para 0.000001 (también escrito como 1×10^{-6}). La pregunta que nos estamos haciendo es que si el error entre sumar 0.1+0.1+0.1 y 0.3 es muy pequeño < 0.000001 :

$$|(0.1 + 0.1 + 0.1) - 0.3| < 0.000001$$

⁹En R el comando **abs** toma el valor absoluto.

A.9 Leer y almacenar variables en R

Para terminar esta sección, aprenderemos cómo guardar variables en R. Para eso, el concepto de directorio es uno de los más relevantes. En general, en computación, el directorio se refiere a la dirección en tu computadora donde estás trabajando. Por ejemplo, si estás en una carpeta en tu escritorio de nombre “Ejercicios_R” probablemente tu directorio sea ‘~/Desktop/Ejercicios_R/’ (en Mac) o bien ‘~\Desktop\Ejercicios_R\’ en Windows¹⁰. La forma de saber tu directorio (en general) es ir a la carpeta que te interesa y con clic derecho ver propiedades (o escribir `ls` en la terminal Unix).

R tiene un directorio `default` que quién sabe dónde está (depende de tu instalación, generalmente está donde tu `Usuario`). Usualmente lo mejor es elegir un directorio para cada uno de los proyectos que hagas. Para ello si estás en RStudio puedes utilizar `Shift+Ctrl+H` (`Shift+Cmd+H` en Mac) o bien ir a `Session > Set Working Directory > Choose Directory` y elegir el directorio donde deseas trabajar tu proyecto. Pensando que elegiste el escritorio (`Desktop` en mi computadora) notarás que en la consola aparece el comando `setwd("~/Desktop")` (o bien con ‘\’ si eres Windows). Mi sugerencia es que copies ese comando en tu `Script` para que, la próxima vez que lo corras ya tengas preestablecido el directorio.

```
#Si eres Mac/Linux
setwd("~/Desktop")

#Si eres Windows
setwd("C:\Users\Rodrigo\Desktop") #Rodrigo = Mi usuario
```

Podemos verificar el directorio elegido con `getwd()`:

```
getwd()
```

En general es buena práctica en R establecer, hasta arriba del `Script`, el comando de directorio. Esto con el propósito de que, cuando compartas un archivo, la persona a quien le fue compartido el archivo pueda rápidamente elegir su propio directorio en su computadora.

Probemos guardar unas variables en un archivo dentro de nuestro directorio. Para ello utilizaremos el comando `save`.

```
#Crear las variables
x <- 200
y <- 100

#Los archivos de variables de R son rda
save(x,y, file = "MisVariables.rda")
```

Si vas a tu directorio, notarás que el archivo `MisVariables.rda` acaba de ser

¹⁰Windows usa backslash. Y hay toda una historia detrás de ello

creado. De esta forma R puede almacenar objetos creados en R que sólo R puede leer (más adelante veremos cómo exportar bases de datos y gráficas). Observa que en tu ambiente (si estás en RStudio puedes verlas en el panel 3) deben aparecer las variables que hemos usado hasta ahora:

```

## [1] "vr.name"           "p.val"          "sub10"
## [4] "ybarra"            "sub11"           "sub12"
## [7] "costo"             "sub13"           "sub14"
## [10] "sub15"              "sub16"           "sub17"
## [13] "sub18"              "confianza.bajo" "sub19"
## [16] "bw"                "eps.error"       "ci"
## [19] "sub20"              "sub21"           "sub22"
## [22] "sub23"              "varianza.est"   "var.total"
## [25] "sub24"              "Uc"              "sub25"
## [28] "mediana.real"     "sub26"           "intervalos.simulados"
## [31] "sub27"              "Uf"              "sub28"
## [34] "sub29"              "sub30"           "sub31"
## [37] "sub32"              "sub33"           "sumaAh"
## [40] "sub34"              "sub35"           "sub36"
## [43] "func.opt.1"         "sub37"           "func.opt.2"
## [46] "sub38"              "sub39"           "k.val"
## [49] "datos"              "var.x"           "sub40"
## [52] "promedios.muestra" "sub41"           "sub42"
## [55] "sub43"              "base.costos"    "sub44"
## [58] "xbarra"             "sub45"           "sub46"
## [61] "sub47"              "B"               "sub48"
## [64] "sub49"              "y.val"          "sub50"
## [67] "sub51"              "sub52"           "sub53"
## [70] "nsim"               "N"               "sub54"
## [73] "sub55"              "sub56"           "sub57"
## [76] "sub58"              "sub59"           "ci_up"
## [79] "Z"                  "sub60"           "sub61"
## [82] "sub62"              "sub63"           "sub64"
## [85] "ci_low"             "sub65"           "pop"
## [88] "sub66"              "sub67"           "sub68"
## [91] "sub69"              "escuelas.seleccionadas" "varianza"
## [94] "f"                  "escuelas"        "i"
## [97] "j"                  "sub70"           "k"
## [100] "sub71"             "sub72"           "m"
## [103] "sub73"             "n"               "sub74"
## [106] "sub75"             "sub76"           "sub77"
## [109] "sub78"             "sub79"           "imagenes.muestreadas"
## [112] "x"                 "y"               "z"
## [115] "sub80"             "sub81"           "sub82"
## [118] "sub83"             "sub84"           "sub85"
## [121] "sub86"             "Loblolly"        "nombres"

```

```

## [124] "sub87"           "mediana.muestral"   "sub88"
## [127] "sub89"           "sub90"             "sub91"
## [130] "media.completa"  "sub92"             "sub93"
## [133] "sub94"           "sub95"             "sub96"
## [136] "sub97"           "alpha"              "sub98"
## [139] "sub99"           "epsilon"            "dats"
## [142] "remuestreo"       "arboles"            "media"
## [145] "Uboot"            "alpha.val"          "base.completa"
## [148] "mediana"          "img"                "pop.total"
## [151] "edad"              "total.muestra"     "base.datos"
## [154] "base.nh"          "confianza.alto"    "lado.inf"
## [157] "sub1"              "sub2"                "sub3"
## [160] "proba"             "sub4"                "sub5"
## [163] "sub6"              "lado.sup"           "sub7"
## [166] "sub8"              "sub9"                "raiz"
## [169] "make.vr"           "imagen"              "r1"
## [172] "sub100"            "r2"                  "Bi"
## [175] "r3"                 "datos.escuelas"    "lambda"
## [178] "muestra"           "g.fun"               "Ntotal"
## [181] "s2"                 "m.val"               "Example1"
## [184] "lambda.1"           "zalpha"              "lambda.2"
## [187] "ensanut"            "total"               "diseño"

```

Podemos probar sumar nuestras variables y todo funciona súper:

```
x + y #Funciona magníficamente
```

```
## [1] 300
```

Limpiemos el ambiente. El comando equivalente al `clear all` en R es un poco más complicado de memorizar:

```
#EL clear all de R
rm(list = ls())
```

Ahora, si vuelves a ver el ambiente, éste estará vacío: ¡hemos limpiado el historial! Nota que si intentamos operar con las variables, R ya no las recuerda:

```
x + y #Error
```

```
## Error in eval(expr, envir, enclos): object 'x' not found
```

Así como hay que lavarse las manos antes de comer, es buen hábito limpiar todas las variables del ambiente de R antes de usarlo.

Podemos leer la base de datos usando `load`:

```
#Leemos las variables
load("MisVariables.rda")
```

```
#Una vez leídas podemos empezar a jugar con ellas
x + y #Ya funciona
```

```
## [1] 300
```

Por último, es necesario resaltar la importancia del directorio. Para ello crea una nueva carpeta en tu escritorio de nombre Mi_curso_de_R. Mueve el archivo "MisVariables.rda" dentro de la carpeta. Borra todo e intenta leer de nuevo el archivo:

```
#Borraremos todo
rm(list = ls())

#Intentamos leer el archivo de nuevo
load("MisVariables.rda")

## Warning in readChar(con, 5L, useBytes = TRUE): cannot open compressed file
## 'MisVariables.rda', probable reason 'No such file or directory'

## Error in readChar(con, 5L, useBytes = TRUE): cannot open the connection
```

Este error es porque R sigue pensando que nuestro directorio es el escritorio y está buscando el archivo ahí sin hallarlo. Para encontrarlo hay que cambiar el directorio a través de RStudio (ya sea **Ctrl+Shift+H** o **Session >Set Working Directory > Choose Directory**) o bien a través de comandos en R:

```
#Si eres Mac/Linux
setwd("~/Desktop/Mi_curso_de_R")

#Si eres Windows
setwd("C:/Users/Rodrigo/Desktop/Mi_curso_de_R") #Rodrigo = Mi usuario

#Aquí sí se puede leer
load("MisVariables.rda")
```

A.9.1 Ejercicio

Responde a las siguientes preguntas:

1. ¿Qué es el directorio y por qué es necesario establecerlo?
2. Si R me da el error 'No such file or directory' ¿qué hice mal?
3. En RStudio, ¿qué hace Session > Restart R? ¿cuál es la diferencia con `rm(list = ls())`?
4. ¿Qué hace el comando `cat("\014")`? (*Ojo* puede que no haga nada). Si funciona, ¿cuál es la diferencia con `rm(list = ls())` y con `Restart R`?

A.10 Instalación de paquetes

Un paquete de R es un conjunto de funciones adicionales elaboradas por los usuarios, las cuales permiten hacer cosas adicionales en R. Para instalarlos requieres de una conexión a Internet (o bien puedes instalarlos a partir de un archivo, por ejemplo, mediante una USB). El comando de instalación es `install.packages` seguido del nombre del paquete. Por ejemplo (y por ocio) descarguemos el paquete `beepr` para hacer reproducir sonidos en la computadora¹¹. Para ello:

```
install.packages("beepr")
```

```
[...]
* DONE (beepr)
```

```
The downloaded source packages are in
  '/algun/lugar/downloaded_packages'
```

Esto significa que el paquete ha sido instalado. Nos interesa usar la función `beep` que emite un sonido (`??beep` para ver la ayuda). Si la llamamos así tal cual, nos da error:

```
beep(3)
```

R es incapaz de hallar la función porque aún no le hemos dicho dónde se encuentra. Para ello podemos llamar al paquete mediante la función `library` y decirle a R que incluya las funciones que se encuentran dentro de `beepr`:

```
library(beepr)
beep(3) #Este produce un sonido
```

El comando `library` le dice a R ¡hey, voy a usar unas funciones que creó alguien más y que están dentro del paquete `beepr`! De esta manera, al correr `beep(3)`, R ya sabe dónde hallar la función y por eso no arroja error.

A.10.1 Ejercicios

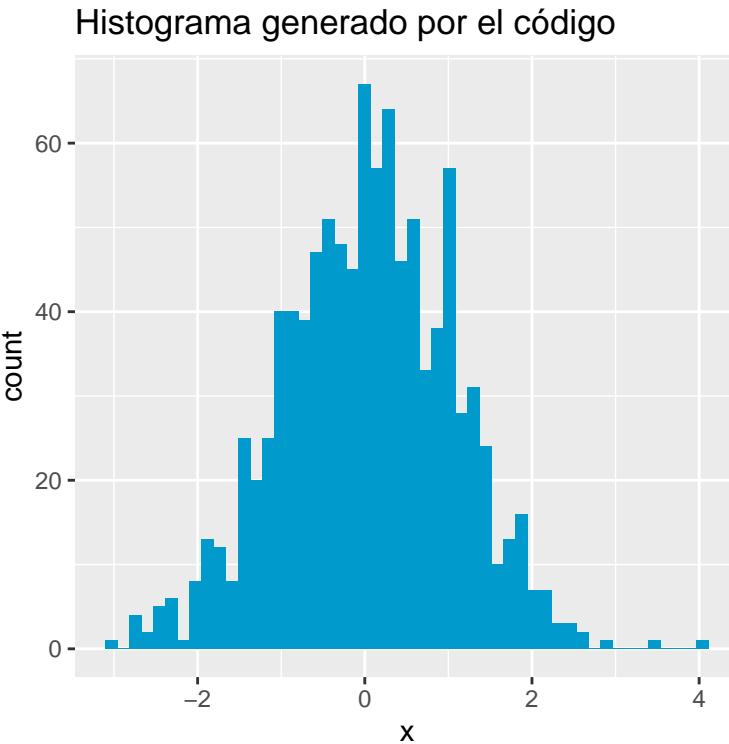
NIVEL 1

- Instala los paquetes `tidyverse` en R.
- De `tidyverse` haz lo necesario para que el siguiente bloque de código te arroje una gráfica:

```
#Aquí tienes que hacer algo
#
# RELLENA AQUÍ
#
```

¹¹En los siguientes capítulos descargaremos paquetes más interesantes; pero no desprecies la utilidad de `beepr` yo lo he usado en múltiples ocasiones para que la computadora me avise que ya terminó de correr un código.

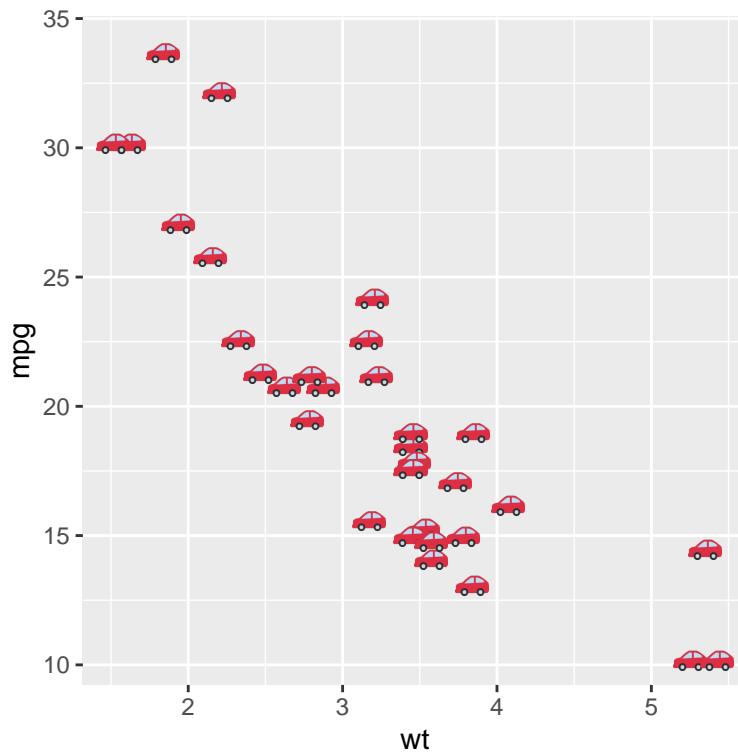
```
#Esto genera un histograma
set.seed(1364752)
mis.datos <- data.frame(x = rnorm(1000))
ggplot(mis.datos, aes(x = x)) +
  geom_histogram(bins = 50, fill = "deepskyblue3") +
  ggttitle("Histograma generado por el código")
```



NIVEL 3

1. Instala el paquete `devtools` (para hacerlo probablemente necesites instalar más cosas en tu computadora; averigua cuáles)
2. Usa `devtools` para instalar el paquete `emoGG` desde Github.
3. Verifica que tu instalación fue correcta haciendo la siguiente gráfica:

```
library(emoGG)
ggplot(mtcars, aes(wt, mpg)) + geom_emoji(emoji="1f697")
```



A.11 Comentarios adicionales sobre el formato

Así como en el español existen reglas de gramática para ponernos todos de acuerdo y entendernos entre todos, en R también existen *sugerencias* a seguir para escribir tu código. Las sugerencias que aquí aparecen fueron adaptadas de las que utiliza el equipo de Google.

1. No escribas líneas de más de 80 caracteres (si se salió de tu pantalla, mejor continúa en el siguiente renglón).
2. Coloca espacios entre operadores +, *, /, -, <-, =, <, <=, >, >=, == y usa paréntesis para agrupar:

```
#Esto no se ve muy bien
abs(3*5/(4-9)^2-60/100-888+0.1*8888-4/10*2) < 1.e-6
```

```
#Los espacios permiten distinguir el orden de las operaciones
abs( (3 * 5) / (4 - 9)^2 - 60 / 100 - 888
    + (0.1 * 8888) - (4 / 10) * 2 ) < 1.e-6
```

3. Intenta alinear la asignación de variables para legibilidad:

```
#Esto no tanto
altura <- 1.80
peso <- 80
edad <- 32

#Esto se ve bien
altura <- 1.80
peso <- 80
edad <- 32
```

4. Utiliza nombres que evoquen la variable que representas

```
#Cuando regreses a esto no sabrás ni qué
x <- 10
y <- 2
z <- 3.14
W <- z * x^y #¿Qué calculé?

#Es mejor especificar la variable
radio <- 10
potencia <- 2
pi_aprox <- 3.14
area_circulo <- pi_aprox * radio^potencia
```

5. No utilices un nombre demasiado similar para cosas diferentes.

```
#Aquí, seguro eventualmente te vas a equivocar
altura <- 10 #Altura del edificio
Altura <- 1.8 #Mi altura
ALTURA <- 2000 #La altitud de la CDMX

#Siempre elegir nombres claros, aunque largos
altura.edificio <- 10 #Altura del edificio
altura.Rodrigo <- 1.8 #Mi altura
altura.CDMX <- 2000 #La altitud de la CDMX
```

6. Comenta:

```
#¿Qué hace esto?
x <- 168
x <- x/100
y <- 71.2
print(y/x^2)

#Es mejor así
altura <- 168      #en centímetros
altura <- altura/100 #en metros
peso <- 71.2        #peso en kg
```

```
print(peso/altura^2) #índice masa corporal
```

7. Siempre pon las llamadas a los paquetes y el directorio al inicio de tu archivo para que otro usuario sepa qué necesita.

Código limpio y legible:

```
#Asumiendo aquí inicia el archivo:
setwd("Mi directorio")

#Llamamos la librería
library(beepr)
library(tidyverse)

#Analizamos una base de datos de R
data(iris) #Base de datos de flores

#Agrupamos la base por especie
iris.agrupada <- group_by(iris, Species)

#Obtenemos la media por longitud de sépalo
iris.media    <- summarise(iris.agrupada, SL.mean = mean(Sepal.Length))

#Avisa que ya terminó
beep(5)
```

es siempre preferible a código escrito *con prisas* :

```
data(iris);setwd("Mi directorio")
library(tidyverse);x<-group_by(iris,Species )
#Aquí hacemos esto
iris.means=summarise( x,SL.mean=mean(Sepal.Length));library(beepr);beep(5) #FIN
```

Siempre escribe tu código pensando que alguien más (y ese alguien más puedes ser tú) va a leerlo. ¡No olvides comentar!

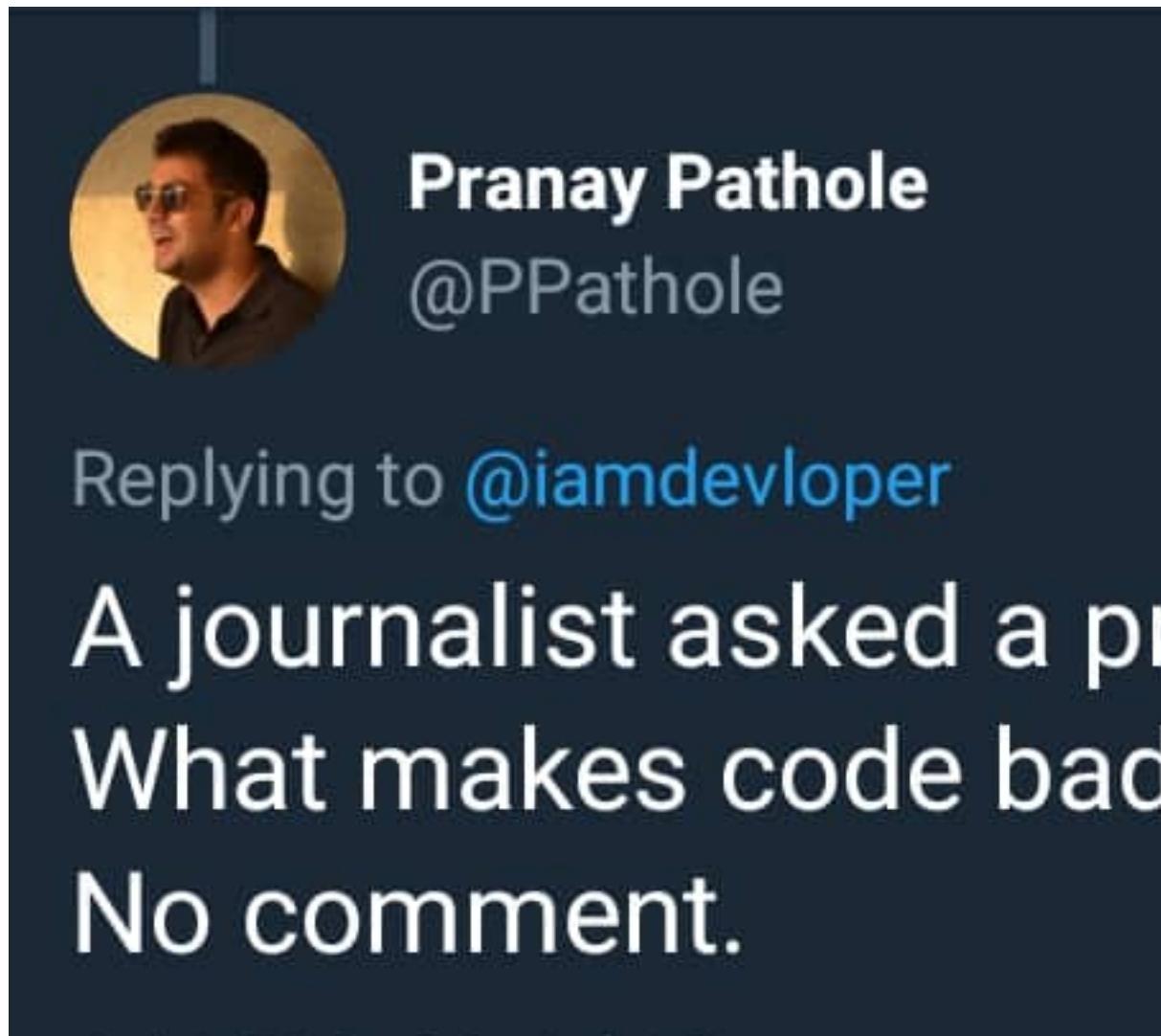


Figure A.17: Trad: Un periodista se acerca a un programador a preguntarle ¿qué hace que un código sea malo? -Sin comentarios.

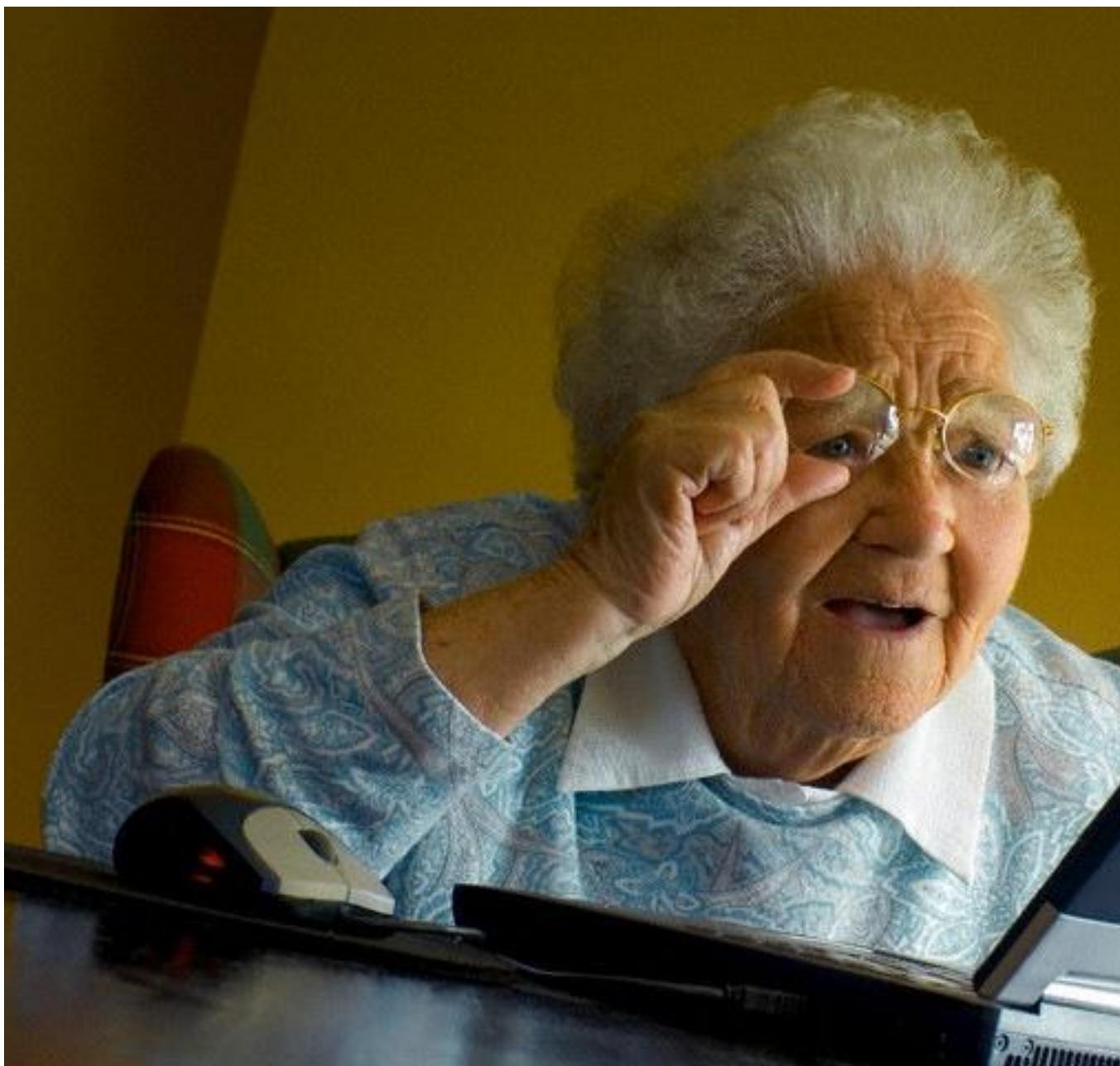


Figure A.18: Yo, leyendo mi código no comentado y con mala edición 6 meses después de haberlo hecho.

Appendix B

Repasso de Proba

B.1 Funciones indicadoras

Dado un conjunto A definimos la función indicadora de A como sigue:

$$\mathbb{I}_A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{si } x \notin A \end{cases}$$

La función indicadora cumple las siguientes propiedades:

Sean A, B conjuntos; luego:

1. $\mathbb{I}_{A \cap B}(x) = \mathbb{I}_A(x) \cdot \mathbb{I}_B(x)$
2. $\mathbb{I}_{A \cup B}(x) = \mathbb{I}_A(x) + \mathbb{I}_B(x) - \mathbb{I}_A(x) \cdot \mathbb{I}_B(x)$
3. $\mathbb{E}_X[\mathbb{I}_A(X)] = \mathbb{P}(X \in A)$

Demostración: 1. Si $x \in A \cap B$ pasa que $\mathbb{I}_{A \cap B}(x) = 1$; además, por hipótesis $x \in A$ y $x \in B$ lo que implica que $\mathbb{I}_A(x) = 1$ y $\mathbb{I}_B(x) = 1$; en caso contrario $\mathbb{I}_{A \cap B}(x) = 1$ y como no está en el conjunto al menos uno $\mathbb{I}_A(x)$ ó $\mathbb{I}_B(x)$ es cero. Esto concluye la prueba. 2. Demostración es similar 3. Para cualquier variable aleatoria X , $\mathbb{I}_A(X)$ sólo toma dos valores: 0 si $X \notin A$ y 1 si $X \in A$. Luego:

$$\mathbb{E}_X[\mathbb{I}_A(X)] = 1 \cdot \mathbb{P}(X \in A) + 0 \cdot \mathbb{P}(X \notin A) = \mathbb{P}(X \in A)$$

B.2 Conteo

Intentemos resumir todas las formas de contar que tenemos con un ejemplo de Casella and Berger (2002).

En la lotería de Nueva York se eligen 6 de 44 números para un ticket.
 ¿Cuántos boletos de lotería posibles hay?

Veamos algunas formas posibles de solución¹:

- a. **Ordenado y sin reemplazo** Si sólo importa el orden y una vez que sale un número no se vuelve a meter a los posibles entonces tenemos:

$$\frac{44!}{(44 - 6)!}$$

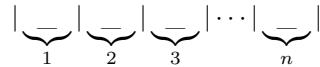
- b. **Ordenado y con reemplazo** En cada uno de los 6 lugares hay 44 números posibles:

$$44^6$$

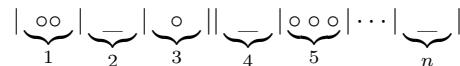
- c. **Sin orden y sin reemplazo** Esto es una combinación por lo que la forma de extraerlo es:

$$\binom{44}{6}$$

- d. **Sin orden y con reemplazo** Para resolver este caso podemos usar la técnica de las barras y los puntos. Coloquemos barras y los huecos entre ellas representan cada uno de los 44 números.



Coloquemos puntos (○) donde estén los números seleccionados. Por ejemplo la siguiente representa la combinación 113555



Tenemos entonces que el problema se reduce a colocar $n - 1 = 43$ barritas (son un total de 45 pero la primera y la última no deben cambiar de lugar) y $k = 6$ círculos por tanto colocamos 49 elementos en total. De estos, nos interesa poner 6 por lo que tenemos:

$$\binom{44 + 6 - 1}{6}$$

formas distintas. Esto nos lleva a la tabla siguiente:

Para obtener una muestra de tamaño k a partir de un conjunto de tamaño $n > 0$ éstas son las opciones:

Con Reemplazo

Sin Reemplazo

¹¿Se te ocurre alguna que no esté aquí?

Con Orden

$$n^k$$

$$(n)_k$$

Sin Orden

$${n+k-1 \choose k}$$

$${n \choose k}$$

B.3 Espacios de probabilidad

Los ingredientes para un modelo probabilístico son 3:

1. Un conjunto Ω conocido como **espacio muestral** el cual es el conjunto de los resultados de interés. Por ejemplo, en el tiro de un dado $\Omega = \{1, 2, 3, 4, 5, 6\}$, para el lanzamiento de una moneda $\Omega = \{\text{Águila}, \text{Sol}\}$ o bien en seleccionar un número uniforme entre 0 y 1 tenemos que $\Omega = [0, 1]$.
2. Una colección \mathcal{F} de subconjuntos de Ω conocida como **sigma-álgebra** o bien como **espacio de eventos** la cual cumple las siguientes características:
 - a. $\Omega \in \mathcal{F}$
 - b. Si $A \in \mathcal{F}$ entonces $A^C \in \mathcal{F}$
 - c. Si A_1, A_2, \dots es una colección finita ó numerable de elementos de \mathcal{F} entonces $\bigcup_n A_n \in \mathcal{F}$

Generalmente identificamos a la \mathcal{F} con la potencia para conjuntos Ω finitos; para casos infinitos el teorema de Vitali nos dice que las cosas son más complicadas.

3. Una función $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ que cumple que:
 - a. $\mathbb{P}(\Omega) = 1$.
 - b. $\mathbb{P}(A) \geq 0$ para todo $A \in \mathcal{F}$.
 - c. Si A_1, A_2, \dots es una colección finita ó numerable de conjuntos disjuntos ($A_i \cap A_j = \emptyset$ para $i \neq j$) entonces $\mathbb{P}(\bigcup_n A_n) = \sum_n \mathbb{P}(A_n)$.

Estos últimos tres puntos se conocen como **Axiomas de Kolmogorov**. Una vez armados con los axiomas podíamos empezar a probar cosas con ellos; por ejemplo:

Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad. Sea A evento de \mathcal{F} . Luego:

$$\mathbb{P}(A^C) = 1 - \mathbb{P}(A).$$

Para verlo, podemos escribir $\Omega = A \cup A^C$ de donde se sigue que:

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^C) = \mathbb{P}(A) + \mathbb{P}(A^C);$$

si despejamos obtenemos el resultado deseado.

También podemos probar, por ejemplo:

$$\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$$

si escribimos $A = (A \setminus B) \cup (A \cap B)$ de donde se sigue que:

$$\mathbb{P}(A) = \mathbb{P}((A \setminus B) \cup (A \cap B)) = \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B)$$

y despejamos para tener el resultado deseado.

Una última cosa de importancia es tomar A, B eventos de \mathcal{F} . Luego:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Para verlo, escribimos $A \cup B$ como $A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$ luego:

$$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A) \\ &= \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)\end{aligned}$$

B.4 Probabilidad condicional

Muchas veces la probabilidad cambia conforme obtenemos información extra. Por ejemplo, si consideramos los tiros de un dado $\Omega = \{1, 2, 3, 4, 5, 6\}$ y se sabe que cayó par $B = \{2, 4, 6\}$, la probabilidad de obtener 2 ó 4 (el evento) $A = \{2, 4\}$ cambia de probabilidad:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

En particular hay dos teoremas principales con probabilidad condicional: la ley de probabilidad total que te permite reconstruir las probabilidades originales a partir de las condicionales y el de Bayes.

El teorema de Bayes puede deducirse a partir de un simple despeje pues notamos que:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

y por otro lado:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

Si despejamos del segundo, obtenemos:

$$\mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A \cap B)$$

Podemos sustituir la definición de intersección en $\mathbb{P}(A|B)$ y así obtener:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

Por otro lado, dada una partición B_1, B_2, \dots finita o numerable de Ω podemos definir la probabilidad de A en términos de cada uno de los pedazos:

$$\mathbb{P}(A) = \sum_k \mathbb{P}(A|B_k) \cdot \mathbb{P}(B_k)$$

Esta identidad se sigue de que:

$$\mathbb{P}(A|B_k) = \frac{\mathbb{P}(A \cap B_k)}{\mathbb{P}(B_k)}$$

de donde podemos sustituir arriba y obtener:

$$\mathbb{P}(A) = \sum_k \frac{\mathbb{P}(A \cap B_k)}{\mathbb{P}(B_k)} \cdot \mathbb{P}(B_k) = \sum_k \mathbb{P}(A \cap B_k) = \mathbb{P}\left(A \cap \left(\bigcup_k B_k\right)\right) = \mathbb{P}(A \cap \Omega)$$

Tenemos entonces el teorema siguiente:

Sean B_1, B_2, \dots eventos que forman una partición de Ω ; sea A un evento cualquiera; luego:

$$\mathbb{P}(A) = \sum_k \mathbb{P}(A|B_k) \cdot \mathbb{P}(B_k)$$

Usando probabilidad condicional podemos resolver problemas como el siguiente:

Considera el conjunto $C = \{1, 2, \dots, n\}$ para $n \geq 2$. Se extraen dos números a y b (primero el a y luego el b) con probabilidad uniforme sin reemplazo. Determina la probabilidad de que $a > b$. Podemos utilizar probabilidad condicional para representar el evento:

$$\mathbb{P}(a > b) = \sum_{k=1}^n \mathbb{P}(a > b \mid a = k) \mathbb{P}(a = k)$$

Donde $\mathbb{P}(a = k) = \frac{1}{n}$ para todos los k pues es uniforme (y es el primero en salir). Luego:

$$\begin{aligned} \mathbb{P}(a > b) &= \sum_{k=1}^n \mathbb{P}(a > b \mid a = k) \mathbb{P}(a = k) \\ &= \frac{1}{n} \sum_{k=1}^n \mathbb{P}(a > b \mid a = k) \\ &= \frac{1}{n} \sum_{k=1}^n \mathbb{P}(k > b \mid a = k) \end{aligned}$$

Notamos que cuando $k = 1$ no hay forma de que $k > b$; cuando $k = 2$ hay una única forma (que b valga 1); cuando $k = 3$ hay dos formas.

En general para una k genérica hay $k - 1$ formas de seleccionar un b menor a k luego:

$$\begin{aligned}\mathbb{P}(a > b) &= \frac{1}{n} \sum_{k=1}^n \frac{k-1}{n} \\ &= \frac{1}{n^2} \sum_{k=1}^n k - 1 \\ &= \frac{1}{n^2} \sum_{k=0}^{n-1} k \\ &= \frac{1}{n^2} \frac{n(n-1)}{2} \\ &= \frac{n-1}{2n}\end{aligned}$$

B.5 Independencia

Dos eventos A, B son independientes si:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Intuitivamente esto significa que saber A no me dice nada de B pues la independencia implica que:

$$\mathbb{P}(A|B) = \mathbb{P}(A)$$

B.6 Variables aleatorias y función de distribución (acumulada)

Para hablar de probabilidad uno de los ingredientes principales eran las variables aleatorias. Éstas son funciones (**no son variables ni son aleatorias**) de tal manera que su imagen inversa pertenece a la sigma-álgebra \mathcal{F} :

Una función $X : \Omega \rightarrow \mathbb{R}$ es una variable aleatoria si:

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}$$

para todo $A \subseteq \text{Dom}_X$

En general la pregunta $\mathbb{P}(X \in A)$ la traducíamos a una pregunta sobre conjuntos:

$$\mathbb{P}(X \in A) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in A\}\right)$$

y esto nos permitía hablar de probabilidades. En particular, construímos la función de distribución acumulada como sigue:

Definimos la función de distribución acumulada de una variable aleatoria $X : \Omega \rightarrow \mathbb{R}$ como:

$$F_X(x) = \mathbb{P}(X \leq x)$$

donde X es la variable aleatoria y $x \in \mathbb{R}$ es un real.

La función de distribución acumulada cumplía varias propiedades:

1. $\lim_{x \rightarrow \infty} F_X(x) = 1$
2. $\lim_{x \rightarrow -\infty} F_X(x) = 0$
3. F_X es no decreciente.
4. F_X es continua por la derecha.
5. F_X tiene límites por la izquierda.

Los puntos 4 y 5 se resumen diciendo que la función es *càdlág*.

Tener la acumulada nos permitía calcular probabilidades de intervalos; por ejemplo:

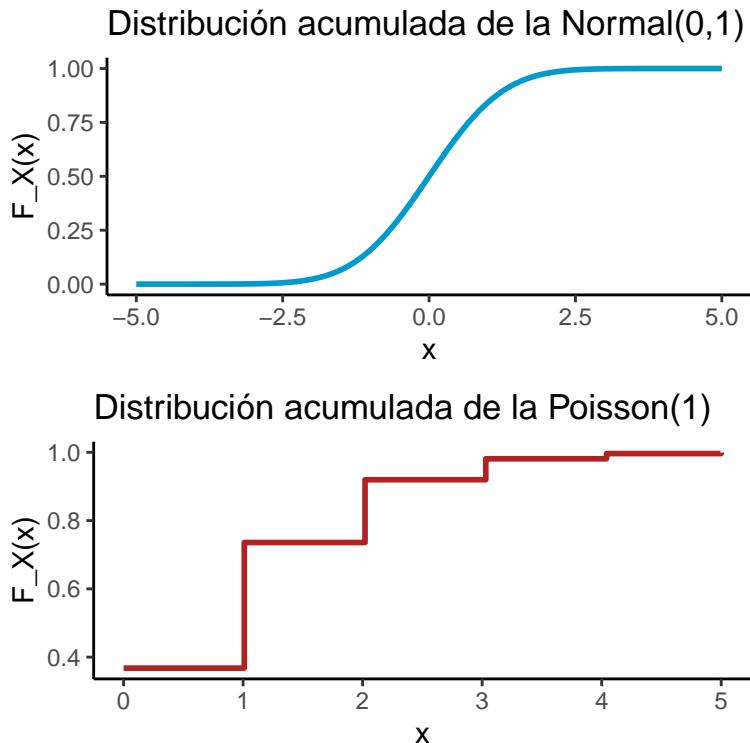
$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$$

o bien:

$$\mathbb{P}(X < x) = \lim_{z \rightarrow x^-} F(z)$$

Las funciones de distribución acumulada más comunes se veían como en la imagen:

```
x <- seq(-5,5,length.out = 100)
y <- pnorm(x)
dats <- data.frame(x = x, y = y)
plot1 <- ggplot(dats) + geom_line(aes(x = x, y = y), color = "deepskyblue3", size = 1) + theme_classic()
xlab("x") + ylab("F_X(x)")
x <- seq(0,5,length.out = 100)
y <- rpois(x, lambda = 1)
dats2 <- data.frame(x = x, y = y)
plot2 <- ggplot(dats2) + geom_step(aes(x = x, y = y), color = "firebrick", size = 1) + theme_classic()
xlab("x") + ylab("F_X(x)")
grid.arrange(plot1,plot2, ncol = 1)
```



Si una función de distribución acumulada F_X era continua entonces decíamos que la variable aleatoria asociada (X) es *continua*. En particular, la continuidad implica que:

$$\mathbb{P}(X = k) = 0 \quad \forall k$$

B.7 Funciones de masa de probabilidad

Si una variable aleatoria X tomaba una cantidad finita o numerable de valores decíamos que X es una variable aleatoria discreta. Dentro de las variables aleatorias discretas teníamos varios modelos. Una cosa importante de las variables aleatorias es la función de masa de probabilidad que se define como:

Dada una variable aleatoria discreta X definimos la función de masa de probabilidad de X como la función $p : \mathbb{R} \rightarrow \mathbb{R}$ tal que:

$$p(x) = \mathbb{P}(X = x)$$

para todo $x \in \mathbb{R}$.

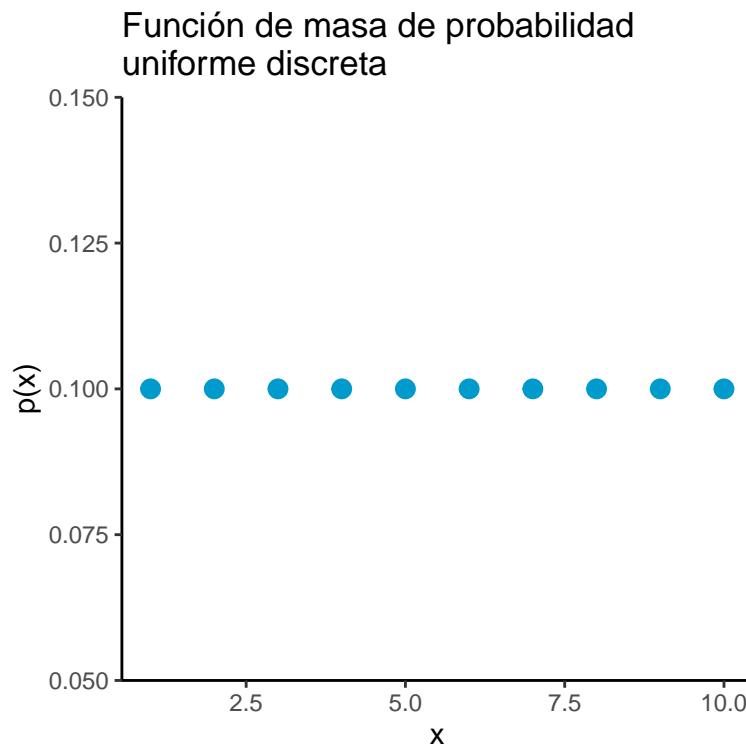
Algunos modelos importantes son:

Sea $A = \{a_1, a_2, \dots, a_n\}$ un conjunto finito de n elementos. Una variable

aleatoria X tiene una distribución uniforme discreta si:

$$\mathbb{P}(X = a_k) = \frac{1}{n} \cdot \mathbb{I}_A(a_k) \quad \forall k \in \{1, 2, \dots, n\}$$

```
x <- 1:10
y <- rep(1/length(x), length(x))
dats <- data.frame(x = x, y = y)
ggplot(dats) + geom_point(aes(x = x, y = y), color = "deepskyblue3", size = 3) + theme_classic()
  xlab("x") + ylab("p(x)")
```



Un modelo particular salía de considerar el siguiente problema:

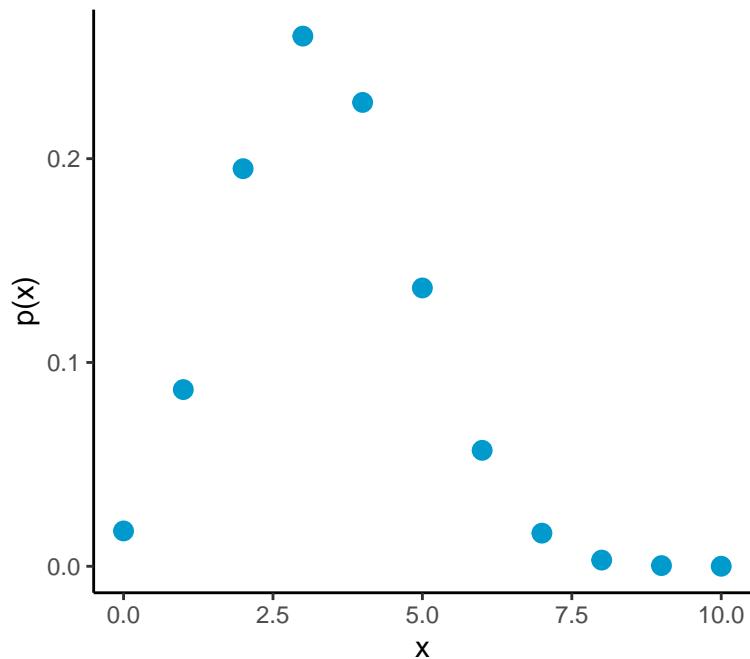
Tenemos una moneda que cae Águila con probabilidad p y Sol con probabilidad $(1 - p)$ (con $0 < p < 1$). Nos interesa saber cuál es la probabilidad de tener k Águilas en n tiros. *Solución* A fin de resolver este problema notamos que necesitamos acomodar las k águilas en los n tiros para ello hay $\binom{n}{k}$ formas de hacerlo; cada águila cae con probabilidad p y hay k ; como son independientes esto nos da p^k ; por otro lado hay $n - k$ soles cada uno cayó con probabilidad $(1 - p)$. Esta lógica da origen al modelo binomial:

Una variable aleatoria X tiene una distribución Binomial(n, p) si:

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \mathbb{I}_{\{0,1,2,\dots,n\}}(k)$$

```
x <- 0:10
y <- dbinom(x, 10, 1/3)
dats <- data.frame(x = x, y = y)
ggplot(dats) + geom_point(aes(x = x, y = y), color = "deepskyblue3", size = 3) + theme_minimal()
```

Función de masa de probabilidad
Binomial(10,1/3)



Una pregunta distinta que nos pudimos hacer fue:

Tenemos una moneda que cae Águila con probabilidad p y Sol con probabilidad $(1 - p)$ (con $0 < p < 1$). Arrojamos la moneda hasta obtener r Águilas y en ese momento nos detenemos. Determina la probabilidad de que se aviente la moneda k veces.

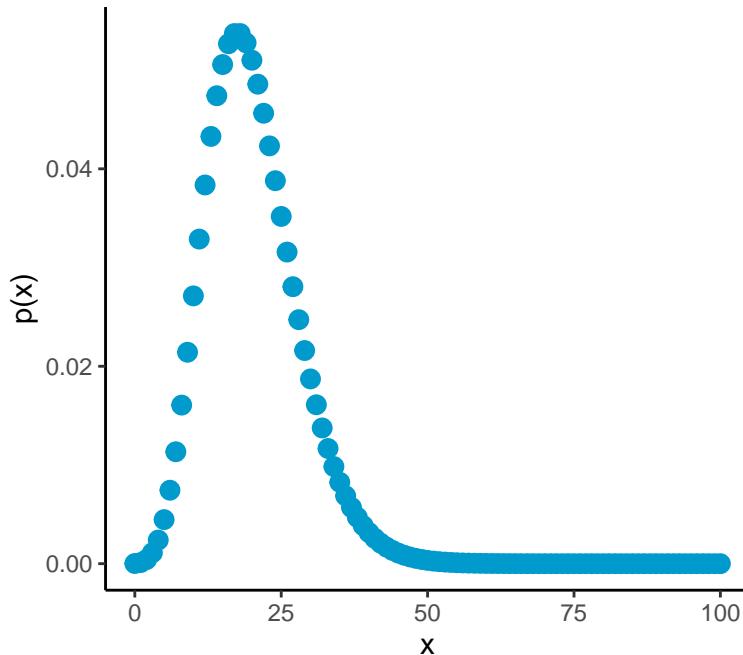
Para ello notamos que la última Águila está fija por lo que sólo debemos poner las $r - 1$ Águilas en los $k - 1$ lugares restantes, $\binom{k-1}{r-1}$. Por otro lado, cada Águila tiene probabilidad p y como son k tiros independientes entonces tenemos p^r ; para los soles tenemos $(1 - p)^{k-r}$. Esto nos genera el modelo Binomial Negativo:

Una variable aleatoria X tiene una distribución Binomial Negativa(r, p) si:

$$\mathbb{P}(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r} \mathbb{I}_{\{r, r+1, r+2, \dots\}}(k)$$

```
x <- 0:100
y <- dnbinom(x, 10, 1/3)
datas <- data.frame(x = x, y = y)
ggplot(datas) + geom_point(aes(x = x, y = y), color = "deepskyblue3", size = 3) + theme_classic()
  xlab("x") + ylab("p(x)")
```

Función de masa de probabilidad
BinomialNegativo(10,1/3)



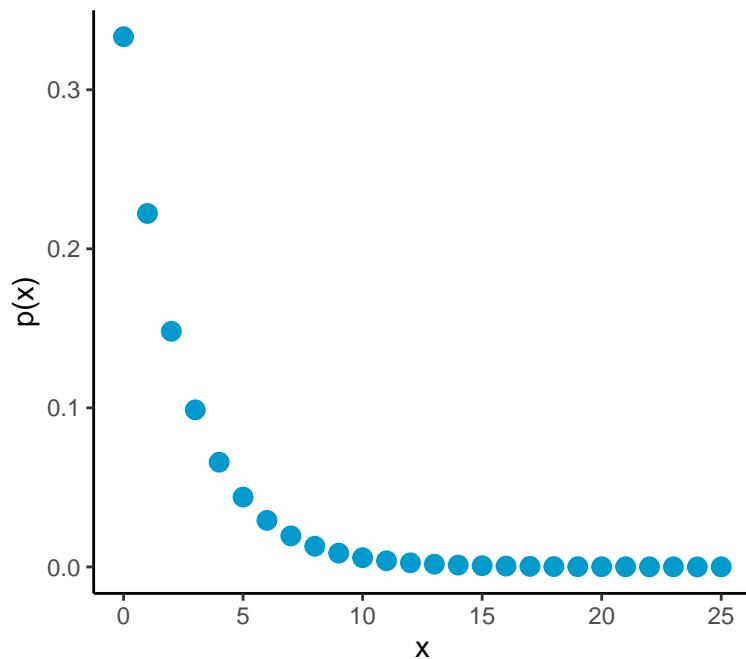
Finalmente, otro modelo que pudimos hacer con monedas es un caso específico del *Binomial Negativo*. Aquí la pregunta es, se tira una moneda que tiene probabilidad p de salir Águila hasta que se obtiene el águila. Contamos cuántos tiros ocurrieron hasta que ocurriera el primer Águila y la pregunta de interés es la probabilidad de haber realizado específicamente k tiros. Para ello necesitamos tener $(k-1)$ tiros que fueran sol: $(1-p)^{k-1}$ y un tiro que saliera águila p . Esto nos genera el modelo geométrico:

Una variable aleatoria X tiene una distribución Geométrica(p) si:

$$\mathbb{P}(X = k) = (1-p)^k p \cdot \mathbb{I}_{\mathbb{N}}(k).$$

```
x <- 0:25
y <- dgeom(x, 1/3)
datas <- data.frame(x = x, y = y)
ggplot(datas) + geom_point(aes(x = x, y = y), color = "deepskyblue3", size = 3) + theme_minimal()
```

Función de masa de probabilidad Geométrica(1/3)



Otro modelo de interés es el siguiente:

Se tiene una población de tamaño M donde N individuos pertenecen al partido político AZUL y $M - N$ pertenecen al VERDE. Se toma una submuestra de tamaño m . Determina la probabilidad de que haya n individuos del partido político AZUL.

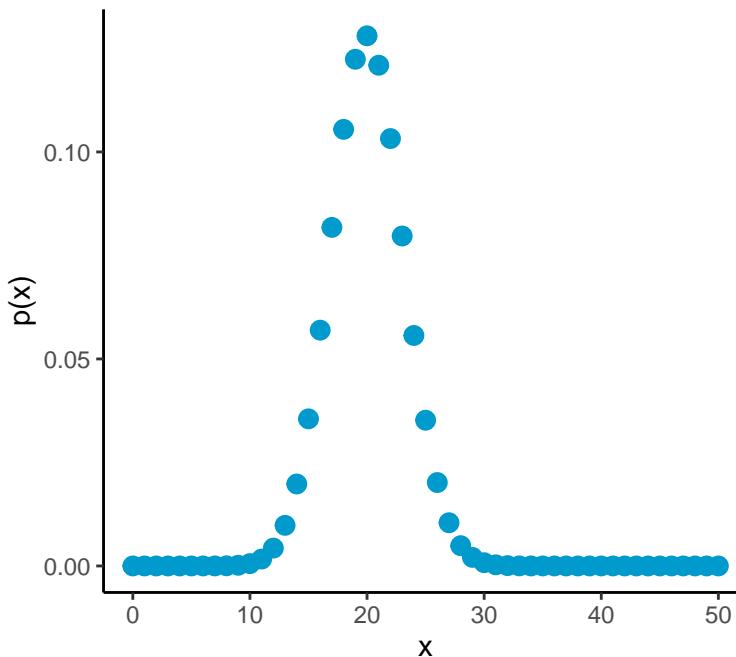
Para ello notamos que hay $\binom{M}{m}$ muestras totales. Por otro lado, necesitamos extraer de los N azules a una submuestra de n : $\binom{N}{n}$; finalmente, de los M verdes necesitamos extraer una submuestra de m , hay $\binom{M-N}{m-n}$ formas de hacerlo. Concluimos entonces con el modelo hipergeométrico:

Una variable aleatoria X tiene una distribución Hipergeométrica(M, N, m) si:

$$\mathbb{P}(X = n) = \frac{\binom{M-N}{m-n} \binom{N}{n}}{\binom{M}{m}} \cdot \mathbb{I}_{\{0, 1, \dots, \min\{m, N\}\}}(n)$$

```
x <- 0:50
y <- dhyper(x, 100, 150, 50)
datas <- data.frame(x = x, y = y)
ggplot(datas) + geom_point(aes(x = x, y = y), color = "deepskyblue3", size = 3) + theme_classic()
  xlab("x") + ylab("p(x)")
```

Función de masa de probabilidad
Hipergeométrica(250, 150, 50)



El modelo Poisson va a ser bastante útil. Para estudiarlo, consideremos un modelo. Vamos a pensar en un servidor de computación (piensa en una página de Internet) que recibe solicitudes de entrar a la página de manera independiente y aleatoria en un intervalo de tiempo entre $t = 0$ y $t = 1$. Como primera aproximación podemos dividir el intervalo en n pedazos cada uno de longitud $1/n$ y asumir que, a fuerza, sólo una conexión se puede realizar en cada uno de esos pedazos. Finalmente, asumimos que la probabilidad p de que se haga una conexión es proporcional a la longitud del intervalo y sea $p = \lambda/n$. Con estas hipótesis, la probabilidad de tener k conexiones (k entero entre 0 y n) está dada por un modelo binomial:

$$f_n(k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \frac{n!}{n^k (n-k)!} \left(1 - \frac{\lambda}{n}\right)^{-k}$$

de donde concluimos que si continuamos partiendo el intervalo en pedazos cada

vez más pequeños obtenemos:

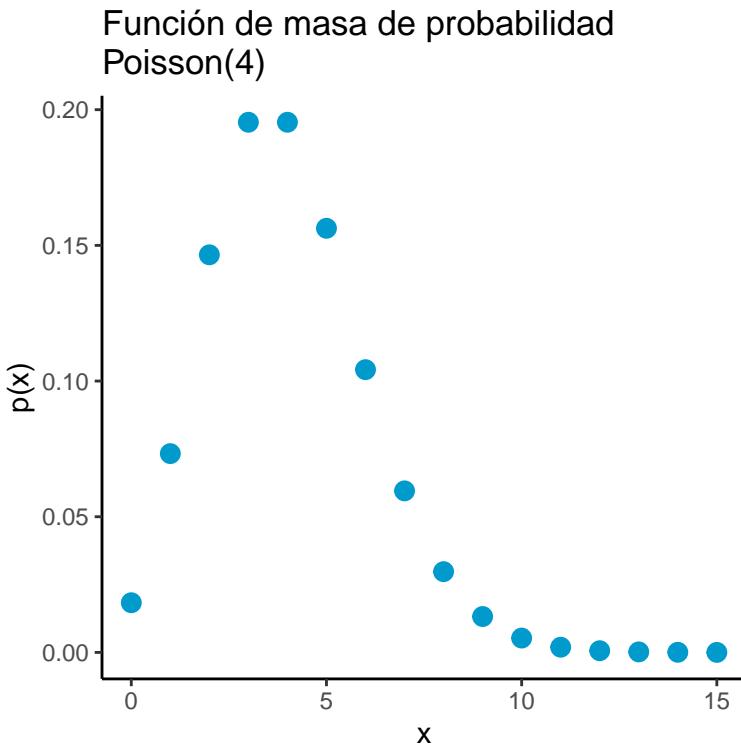
$$\lim_{n \rightarrow \infty} f_n(k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

Esto resulta en el modelo Poisson:

Una variable aleatoria X tiene una distribución Poisson(λ) si:

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \mathbb{I}_{\mathbb{N} \cup \{0\}}(k)$$

```
x <- 0:15
y <- dpois(x, 4)
datas <- data.frame(x = x, y = y)
ggplot(datas) + geom_point(aes(x = x, y = y), color = "deepskyblue3", size = 3) + theme_
xlab("x") + ylab("p(x)")
```



B.8 Funciones de densidad

Por construcción, las variables aleatorias continuas no tienen una función de masa de probabilidad (recuerda que $\mathbb{P}(X = k) = 0$ si X es continua para todo

k). Sin embargo, es posible definir, si F_X es diferenciable algo *similar*, la función de densidad.

Para una variable aleatoria X con función de distribución acumulada F_X diferenciable, definimos la función de densidad como:

$$f_X(x) = \frac{d}{dx} F_X(x)$$

Notamos que una función de densidad no es una probabilidad y no necesariamente sigue las mismas reglas; lo único que se requiere es:

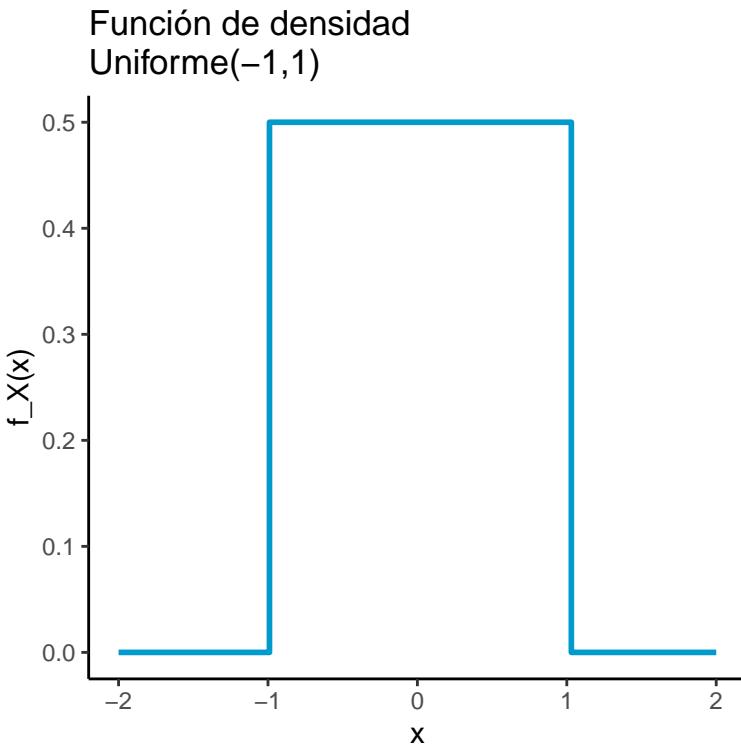
1. $f_X(x) \geq 0$ para toda x .
2. $\int_{-\infty}^{\infty} f_X(x) dx = 1$.

La primer función de densidad es la que a un intervalo $[a, b]$ (ya sea abierto, cerrado o como sea) asigna a cada subintervalo una probabilidad proporcional a su longitud. Éste es el modelo uniforme:

Una variable aleatoria X tiene una distribución Uniforme(a, b) si:

$$f_X(x) = \frac{1}{b-a} \mathbb{I}_{(a,b)}(x)$$

```
x <- seq(-2,2, length.out = 100)
y <- dunif(x, -1, 1)
datas <- data.frame(x = x, y = y)
ggplot(datas) + geom_step(aes(x = x, y = y), color = "deepskyblue3", size = 1) + theme_classic() +
  xlab("x") + ylab("f_X(x)")
```



Una generalización del modelo uniforme es el beta (eventualmente veremos de dónde sale):

Una variable aleatoria X tiene una distribución Beta(α, β) si:

$$f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \mathbb{I}_{(0,1)}(x)$$

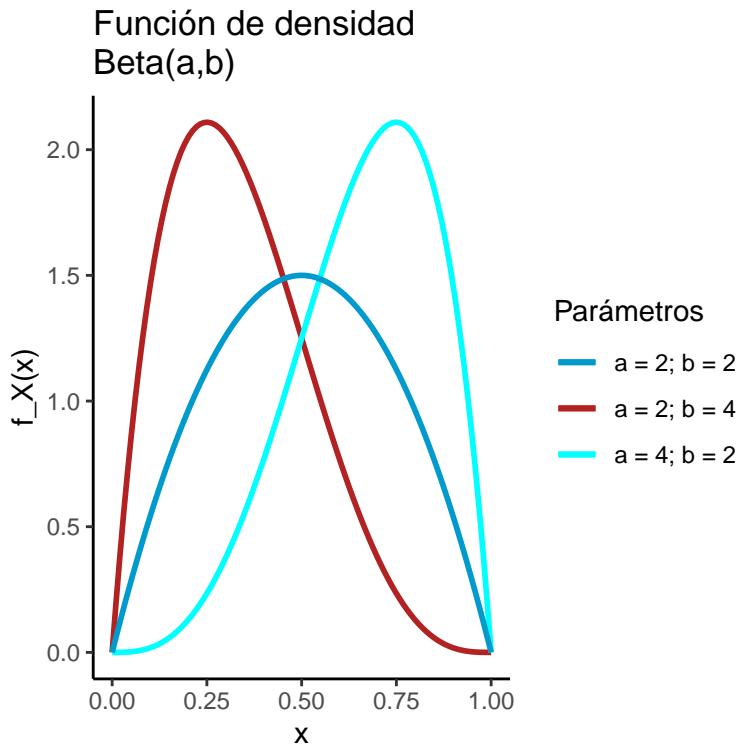
donde

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

```

x <- seq(0,1, length.out = 100)
y1 <- dbeta(x, 2, 4)
y2 <- dbeta(x, 4, 2)
y3 <- dbeta(x, 2, 2)
datas <- data.frame(x = x, y1 = y1, y2 = y2, y3 = y3)
ggplot(datas) +
  geom_line(aes(x = x, y = y1, color = "a = 2; b = 4"), size = 1) +
  geom_line(aes(x = x, y = y2, color = "a = 4; b = 2"), size = 1) +
  geom_line(aes(x = x, y = y3, color = "a = 2; b = 2"), size = 1) +
  theme_classic() + ggtitle("Función de densidad\nBeta(a,b)") +
  xlab("x") + ylab("f_X(x)") +
  scale_color_manual("Parámetros", values = c("deepskyblue3","firebrick","cyan"))

```



Podemos deducir el modelo exponencial a partir de la descripción del Poisson. Volvamos al mismo problema del Poisson(λ) donde hay computadoras conectándose a un servidor. Sea W la variable aleatoria que denota el tiempo de espera hasta el primer evento. Analicemos su distribución acumulada; notamos que

$$F_W(w) = \mathbb{P}(W \leq w) = 1 - \mathbb{P}(W > w)$$

Ahora, para que $W > w$ eso significa que ningún evento tuvo que haber ocurrido en los primeros w minutos (horas, lo que sea la unidad de tiempo). Y ese evento es equivalente a que nuestra variable aleatoria Poisson (tasa λw)² no tenga ningún arriba:

$$\mathbb{P}(X = 0) = \frac{(\lambda w)^0 e^{-\lambda w}}{0!} = e^{-\lambda w}$$

De donde se obtiene la función de distribución acumulada:

$$F_W(w) = 1 - e^{-\lambda w}$$

De donde, al derivar respecto a w , se obtiene el modelo exponencial:

Una variable aleatoria X tiene una distribución Exponencial(λ) si:

$$f_X(x) = \lambda e^{-\lambda x} \mathbb{I}_{(0, \infty)}(x)$$

²Recuerda que λ era para un tiempo entre 0 y 1; λw es para un escalamiento del tiempo entre 0 y w .

Para deducir la distribución gamma, vamos a preguntarnos por exactamente el mismo proceso pero esta vez, en lugar de preguntarnos por el tiempo para la primer conexión nos preguntaremos por el tiempo para la α -ésima conexión. Para ello, sea W_α el tiempo hasta la α -ésima conexión. Usamos el mismo truco del complemento que la vez pasada:

$$F_{W_\alpha}(w) = \mathbb{P}(W_\alpha \leq w) = 1 - \mathbb{P}(W_\alpha > w)$$

Y notamos que para que $W_\alpha > w$ entonces a lo más debieron haber $\alpha - 1$ conexiones. Podemos reescribir:

$$F_{W_\alpha}(w) = 1 - \mathbb{P}(W_\alpha > w) = 1 - \sum_{k=0}^{\alpha-1} \frac{(\lambda w)^k e^{-\lambda w}}{k!} = 1 - e^{-\lambda w} - \sum_{k=1}^{\alpha-1} \frac{(\lambda w)^k e^{-\lambda w}}{k!}$$

Derivamos:

$$\frac{d}{dw} F_{W_\alpha}(w) = -\lambda e^{-\lambda w} - \sum_{k=1}^{\alpha-1} \frac{k\lambda(\lambda w)^{k-1}e^{-\lambda w} - \lambda(\lambda w)^k e^{-\lambda w}}{k!} = -\lambda e^{-\lambda w} - \lambda e^{-\lambda w} \underbrace{\sum_{k=1}^{\alpha-1} \frac{(\lambda w)^{k-1}}{(k-1)!}}_{\text{Telescopica}} - \frac{(\lambda w)^\alpha e^{-\lambda w}}{\alpha!}$$

donde tomamos $\beta = \frac{1}{\lambda}$. Esto sugiere el modelo gamma:

Una variable aleatoria W tiene una distribución Gamma(α, β) si:

$$f_W(w) = \frac{\beta^\alpha}{\Gamma(\alpha)} w^{\alpha-1} e^{-\frac{w}{\beta}} \mathbb{I}_{(0, \infty)}$$

para $\alpha, \beta > 0$.

Para deducir el modelo normal consideremos lo siguiente. Pensemos que estamos midiendo la posición de las estrellas en el cielo. Para ello hay dos formas. Bajo coordenadas cartesianas (x, y) pensemos que el error de medición es independiente; es decir, si $f(x, y)$ es la densidad de los errores entonces:

$$\rho(x, y) = f(x)f(y)$$

Por otro lado, asumamos que existe también una representación en coordenadas polares de la posición de la estrella:

$$g(r, \theta) = g(r)$$

donde el error de medición depende sólo del radio (no del ángulo). Notamos entonces que:

$$f(x)f(y) = g\left(\sqrt{x^2 + y^2}\right)$$

Si tomamos $y = 0$ tenemos que $f(x)f(0) = g(x)$ (asumo $x > 0$; los otros casos son similares). Podemos entonces sustituir:

$$\frac{f(x)f(y)}{f(0)^2} = \frac{f(\sqrt{x^2 + y^2})}{f(0)}$$

Tomamos logaritmo:

$$\ln \frac{f(x)}{f(0)} + \ln \frac{f(y)}{f(0)} = \ln \frac{f(\sqrt{x^2 + y^2})}{f(0)}$$

Notamos que una solución es que:

$$\ln \frac{f(x)}{f(0)} = \alpha x^2$$

de donde despejamos y obtenemos:

$$f(x) = \frac{1}{f(0)} e^{\alpha x^2}$$

Finalmente sabemos que debe integrar a 1 y por tanto esto fuerza a α a ser negativo. En particular tomaremos $\alpha = -\frac{1}{2}$

$$f(x) = \frac{1}{f(0)} e^{-\frac{1}{2}x^2}$$

Y para que integre a 1:

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

Por último, notamos que si $Z \sim \text{Normal}(0, 1)$ entonces $X = \sigma Z + \mu$ tiene la densidad dada por³:

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

Una variable aleatoria X tiene una distribución $\text{Normal}(\mu, \sigma)$ si:

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

B.9 Teorema de cambio de variable unidimensional

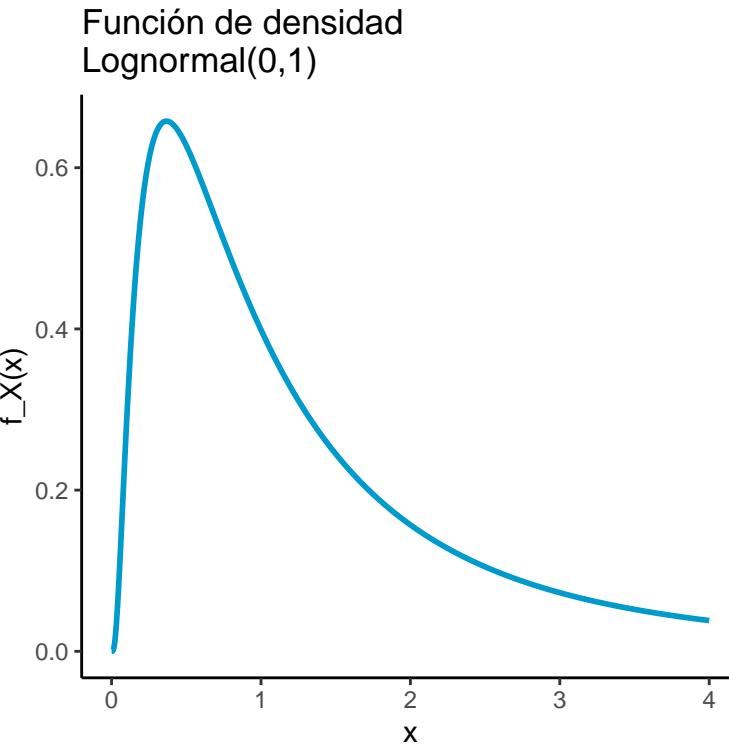
Supongamos que tenemos una variable aleatoria X y nos interesa ver cómo se ve la X después de aplicarle una función ϕ . Por ejemplo, si $X \sim \text{Normal}(0, 1)$ la función de densidad de e^X está dada por:

³Por teorema de cambio de variable.

$$f_X(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-(\ln(x)-\mu)^2/2\sigma^2} \mathbb{I}_{(0,\infty)}(x).$$

Lo cual cambia mucho la forma de la distribución:

```
x <- seq(0,4, length.out = 1000)
y <- dlnorm(x, 0, 1)
data <- data.frame(x = x, y = y)
ggplot(data) + geom_line(aes(x = x, y = y), color = "deepskyblue3", size = 1) + theme_c
```



La pregunta es, cómo obtener la función de densidad de X si se conoce la función ϕ ; el teorema de cambio de variable nos da una respuesta cuando ϕ es monótona estrictamente creciente o bien estrictamente decreciente y diferenciable.

Sea X una variable aleatoria continua y ϕ una función estrictamente creciente ó estrictamente decreciente y diferenciable. Entonces:

$$f_{\phi(X)}(t) = f_X(\phi^{-1}(t)) \cdot \left| \frac{d}{dt} \phi^{-1}(t) \right|$$

DEM: Caso estrictamente decreciente Como ϕ es estrictamente decreciente

es invertible y por tanto:

$$\begin{aligned} F_{\phi(X)}(t) &= \mathbb{P}(\phi(X) \leq t) \\ &= \mathbb{P}(X \geq \phi^{-1}(t)) \\ &= 1 - \mathbb{P}(X \leq \phi^{-1}(t)) \\ &= 1 - F_X(\phi^{-1}(t)) \end{aligned}$$

luego derivamos respecto a t :

$$\begin{aligned} f_{\phi(X)}(t) &= \frac{d}{dt} F_{\phi(X)}(t) \\ &= -\frac{d}{dt} F_X(\phi^{-1}(t)) \\ &= -f_X(\phi^{-1}(t)) \cdot \frac{d}{dt} \phi^{-1}(t) \\ &= f_X(\phi^{-1}(t)) \cdot \left| \frac{d}{dt} \phi^{-1}(t) \right| \end{aligned}$$

Donde el valor absoluto sale de que $\phi^{-1}(t) < 0$ por ser estrictamente decreciente la ϕ .

B.10 Probabilidad Multivariada

De la misma manera que hablamos de una sola variable aleatoria podemos hablar de muchas como múltiples funciones de $\Omega \in \mathbb{R}$. Para una colección finita $\{X_i\}_{i=1}^n$ de variables aleatorias podemos hablar de su función de distribución acumulada conjunta como:

$$F_{\vec{X}}(x_1, x_2, \dots, x_n) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$$

donde suponemos que $\vec{X} = (X_1, X_2, \dots, X_n)^T$ es un vector aleatorio cuyas entradas son las variables de la colección anterior. En el caso de que las n variables sean discretas la función de masa conjunta está dada por:

$$p_{\vec{X}}(x_1, x_2, \dots, x_n) = \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

En el caso de que sean continuas ($F_{\vec{X}}$ diferenciable en sus n entradas) entonces la densidad está dada por:

$$f_{\vec{X}}(x_1, x_2, \dots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \dots \partial x_n} F_{\vec{X}} \Big|_{(x_1, x_2, \dots, x_n)}$$

En general la función de probabilidad conjunta siempre va a esta dada por:

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n) = \mathbb{P}\left(\{\omega \in \Omega : X_1(\omega) \in A_1 \text{ y } X_2(\omega) \in A_2 \text{ y } \dots \text{ y } X_n(\omega) \in A_n\}\right)$$

para A_1, A_2, \dots, A_n medibles (bajo X_1, X_2, \dots, X_n respectivamente).

Dos variables aleatorias X_i y X_j ($i \neq j$) son independientes si:

$$\mathbb{P}(X_i \in A, X_j \in B) = \mathbb{P}(X_i \in A) \cdot \mathbb{P}(X_j \in B)$$

para A, B medibles. Una colección $\{X_i\}_i$ de variables aleatorias es **completamente independiente** si para cualquier subcolección finita $\{X_{i_k}\}_{i_k}$ se tiene que:

$$\mathbb{P}(X_{i_1} \in A_{i_1}, X_{i_2} \in A_{i_2}, \dots, X_{i_n} \in A_{i_n}) = \prod_{k=1}^n \mathbb{P}(X_{i_k} \in A)$$

en el contexto de estas notas, a menos que se indique lo contrario, las variables aleatorias que utilicemos serán **completamente independientes**.

Un aspecto interesante de la independencia es que permite partir las funciones de masa, densidad y distribución acumulada en dos funciones independientes. Así, si X, Y son independientes con masa conjunta p :

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y) = p_X(x) \cdot p_Y(y)$$

El resultado se mantiene para distribuciones:

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq y) = F_X(x) \cdot F_Y(y)$$

y si derivamos (en caso de F diferenciable), se mantiene para densidades:

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y} \Big|_{(x,y)} = \frac{\partial^2}{\partial x \partial y} F_X(x) \cdot F_Y(y) \Big|_{(x,y)} = f_X(x) f_Y(y)$$

B.11 Esperanza, varianza y covarianza

Para una función medible g de una variable aleatoria X definimos su valor esperado (si existe) como:

$$\mathbb{E}[g(X)] = \begin{cases} \sum_{x \in \text{Supp}(X)} g(x) \cdot \mathbb{P}(X = x) & \text{si } X \text{ discreta.} \\ \int_{-\infty}^{\infty} g(x) \cdot f_X(x) dx & \text{si } X \text{ continua} \end{cases}$$

donde f_X es la densidad de X en el caso continuo y $\text{Supp}(X)$ es el conjunto imagen de X (el soporte):

$$\text{Supp}(X) = \{x : X(\omega) = x \text{ para } \omega \in \Omega\}$$

En el caso de conjuntos finitos de variables aleatorias la definición es similar:

Para una función $g : \mathbb{R}^n \rightarrow \mathbb{R}$ multivariada de n variables aleatorias (sobre los reales) X_1, X_2, \dots, X_n definimos su valor esperado (si existe y sin pérdida de

generalidad suponiendo las primeras j son discretas y las últimas $n - (j + 1)$ continuas) como:

$$\mathbb{E}[g(X_1, X_2, \dots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{x_j \in \text{Supp}(X_j)} \cdots \sum_{x_1 \in \text{Supp}(X_1)} g(x_1, x_2, \dots, x_n) p(x_1) \cdots p(x_j) f_{X_{j+1}}(x_{j+1}) \cdots f_{X_n}(x_n) dx_{j+1} \cdots dx_n$$

donde $p(x_j)$ es la masa de X_j (es decir $p(x_j) = \mathbb{P}(X_j = x_j)$). En el caso particular de dos variables aleatorias X_1 y X_2 podemos escribir la expresión de manera más sencilla:

$$\mathbb{E}[g(X_1, X_2)] = \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x_1, x_2) f_{X_1}(x_1) f_{X_2}(x_2) dx_1 dx_2 & \text{ambas continuas,} \\ \sum_{x \in \text{Supp}(X_1)} \sum_{x \in \text{Supp}(X_2)} g(x_1, x_2) p(x_1) p(x_2) & \text{ambas discretas,} \\ \int_{-\infty}^{\infty} \sum_{x \in \text{Supp}(X_1)} g(x_1, x_2) p(x_1) f(x_2) dx_2 & X_1 \text{ discreta, } X_2 \text{ continua.} \end{cases}$$

En particular, en el espacio de las variables aleatorias definimos un producto interno, la **covarianza** la cual está dada por:

$$\text{Cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}[X_1]) \cdot (X_2 - \mathbb{E}[X_2])]$$

La **varianza** es un caso particular de la covarianza: cuando $X_1 = X_2$:

$$\text{Cov}(X_1, X_1) = \mathbb{E}[(X_1 - \mathbb{E}[X_1])^2]$$

B.11.1 Propiedades de valor esperado, varianza y covarianza

El valor esperado al ser representable mediante sumas ó integrales cumple todas las propiedades de las sumas (resp integrales) en particular la linealidad:

$$\mathbb{E}[aX + Y] = a\mathbb{E}[X] + \mathbb{E}[Y]$$

La demostración se hace exactamente igual en el caso de variables discretas, continuas (ó mezcla de una y una). Aquí muestro la de continuas con densidades f_X y f_Y :

$$\begin{aligned} \mathbb{E}[aX + Y] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (ax + y) f_{X,Y}(x, y) dx dy \\ &= a \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf_{X,Y}(x, y) dx dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf_{X,Y}(x, y) dx dy \quad (\text{B.1}) \\ &= a \left[\int_{-\infty}^{\infty} xf_X(x) dx \right] + \int_{-\infty}^{\infty} yf_Y(y) dy \\ &= a\mathbb{E}[X] + \mathbb{E}[Y] \end{aligned}$$

Otro resultado importante es que si dos variables aleatorias X, Y son independientes entonces el valor esperado del producto se parte:

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

La demostración se hace de manera idéntica en todos los casos. Aquí mostramos el caso de X, Y discretas:

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{y \in \text{Sup}(Y)} \sum_{x \in \text{Sup}(X)} xy \mathbb{P}(X = x, Y = y) \\
&= \sum_{y \in \text{Sup}(Y)} \sum_{x \in \text{Sup}(X)} xy \mathbb{P}(X = x) \mathbb{P}(Y = y) \\
&= \left[\sum_{y \in \text{Sup}(Y)} y \mathbb{P}(Y = y) \right] \left[\sum_{x \in \text{Sup}(X)} x \mathbb{P}(X = x) \right] \\
&= \mathbb{E}[X] \cdot \mathbb{E}[Y]
\end{aligned} \tag{B.2}$$

La linealidad nos permite reescribir la covarianza:

$$\begin{aligned}
\text{Cov}(X_1, X_2) &= \mathbb{E}[(X_1 - \mathbb{E}[X_1]) \cdot (X_2 - \mathbb{E}[X_2])] \\
&= \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1] \mathbb{E}[X_2] - \mathbb{E}[X_1] \mathbb{E}[X_2] + \mathbb{E}[X_1] \mathbb{E}[X_2] \\
&= \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1] \mathbb{E}[X_2]
\end{aligned} \tag{B.3}$$

de tal forma que es claro que si X_1 y X_2 son independientes entonces $\text{Cov}(X_1, X_2) = 0$ por la propiedad anterior del valor esperado. **OJO** De manera general covarianza 0 no implica que las variables sean independientes como puede verse con las variables aleatorias siguientes:

$$f_{X,Y}(x, y) = \begin{cases} 1/8 & \text{si } (x, y) \in \{(-1, -1), (-1, 1), (1, -1), (1, 1)\} \\ 1/2 & \text{si } (x, y) = (0, 0), \\ 0 & \text{en otro caso} \end{cases}$$

las cuales no son independientes pues $\mathbb{P}(X = 0, Y = 0) = 1/2 \neq 1/4 = \mathbb{P}(X = 0) \cdot \mathbb{P}(Y = 0)$; sin embargo (ejercicio sugerido) la covarianza es 0.

Una segunda propiedad de interés de la covarianza es que actúa como el producto interno (de hecho es uno):

$$\text{Cov}(aX + bY, cW + dV) = ac\text{Cov}(X, W) + ad\text{Cov}(X, V) + bc\text{Cov}(Y, W) + bd\text{Cov}(Y, V)$$

la cual se demuestra igual mediante la linealidad:

$$\begin{aligned}
\text{Cov}(aX + bY, cW + dV) &= \mathbb{E}[(aX + bY)(cW + dV)] - \mathbb{E}[aX + bY] \mathbb{E}[cW + dV] \\
&= \mathbb{E}[acXW + bcYW + adXV + bdYV] - (a\mathbb{E}[X] + b\mathbb{E}[Y])(c\mathbb{E}[W] + d\mathbb{E}[V]) \\
&= ac\text{Cov}(X, W) + ad\text{Cov}(X, V) + bc\text{Cov}(Y, W) + bd\text{Cov}(Y, V)
\end{aligned} \tag{B.4}$$

donde la última igualdad se sigue de agrupar los términos idénticos tras sus constantes.

B.12 Condicionamiento por otra variable aleatoria

A rellenarse pronto

B.13 Funciones características

A rellenarse pronto

B.14 Convergencias

A rellenarse pronto

B.14.1 Teorema de continuidad de Lévy

A rellenarse pronto

B.15 Ley de los grandes números

A rellenarse pronto

B.16 Teorema del límite central

A rellenarse pronto

B.16.1 Programación en R del teorema del límite central con variables aleatorias independientes idénticamente distribuidas

Lo que programaremos (por facilidad) en esta sección corresponde a ejemplos del teorema de proba 2: dadas variables aleatorias independientes idénticamente $\{X_i\}$ distribuidas con media μ y varianza finita σ^2 tenemos que:

$$Z = \lim_{n \rightarrow \infty} \sqrt{\frac{n}{\sigma^2}} \cdot \left(\frac{1}{n} \sum_{i=1}^n X_i - \mu \right) \sim \text{Normal}(0, 1)$$

donde el símbolo \sim se lee “se distribuye.” En este caso la interpretación va a ser que para n muy grande tendremos que

$$\sqrt{\frac{n}{\sigma^2}} \cdot \left(\frac{1}{n} \sum_{i=1}^n X_i - \mu \right) \sim \text{Normal}(0, 1)$$

donde \sim se lee como “se distribuye aproximadamente.” Programaremos una función en R que para n grande muestre eso:

```
TeoremaCentralLimite <- function(numero_simulaciones = 1000,
                                    n = c(10,100,1000,10000),
                                    distribucion = rpois, mu = 1, sigma = 1,
                                    bins = 50,
                                    ncol = 2, distname = "Poisson(1)",
                                    rcolor = sample(rainbow(100),1), ...){

  #Creamos
  plot_list <- list()

  for (k in n){

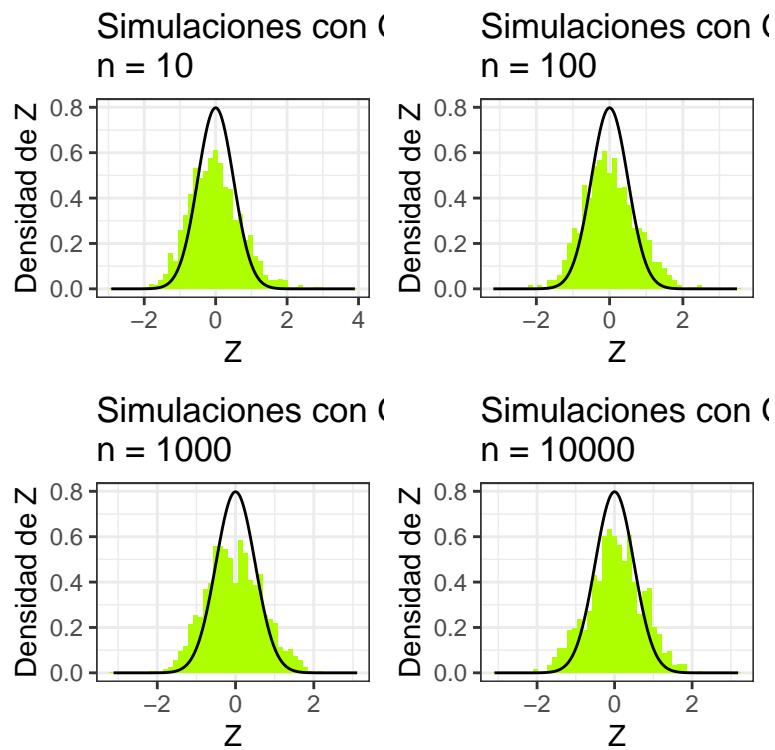
    #Guardamos las Zi en un vector
    Z <- rep(NA, numero_simulaciones)

    #Simulamos todas las simulaciones
    for (i in 1:numero_simulaciones){
      simulaciones_X <- distribucion(n = k, ...)
      Z[i]           <- sqrt(k)*(sum(simulaciones_X/k) - mu)
    }

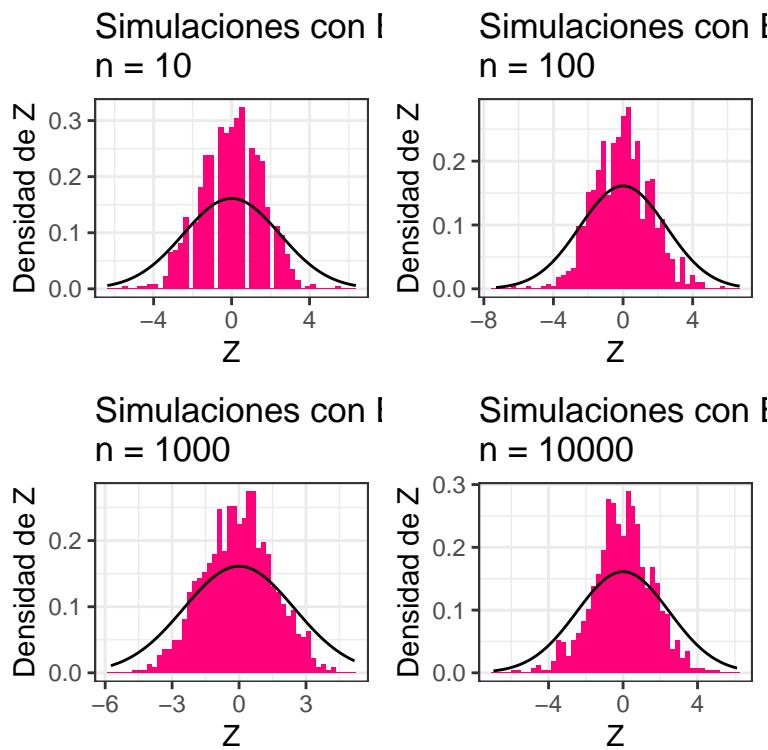
    #Gráficación
    x <- seq(min(Z)-1, max(Z) + 1, length.out = 1000)
    y <- dnorm(x, sd = sigma)
    plot_list <- list.append(
      plot_list,
      ggplot() +
        geom_histogram(aes(x = Z, y = ..density..), bins = bins, fill = rcolor,
                      data = data.frame(Z = Z)) +
        geom_line(aes_string(x = x, y = y), color = "black", data = data.frame(x = x, y = y)) +
        ggtitle(paste0("Simulaciones con ", distname, "\nn = ", k)) +
        xlab("Z") + ylab("Densidad de Z") +
        theme_bw()
    )
  }
  do.call("grid.arrange", c(plot_list, ncol = ncol))
}
```

donde podemos ver la aproximación normal si tomamos, por ejemplo, las X_i siguen una distribución Gamma:

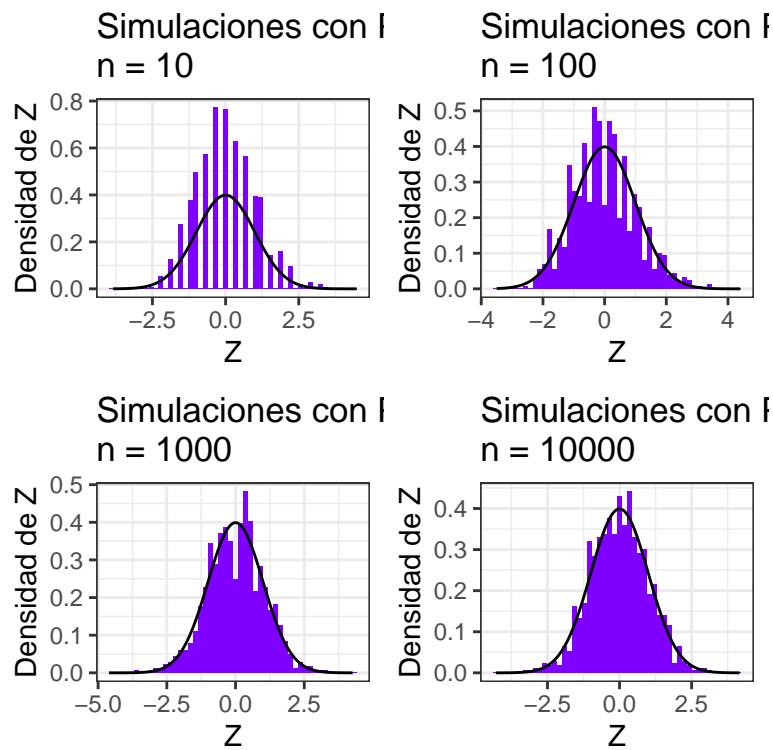
```
TeoremaCentralLimite(distribucion = rgamma, mu = 1, sigma = 0.5, shape = 2,
                      scale = 0.5, distname = "Gama(2,1/2)")
```



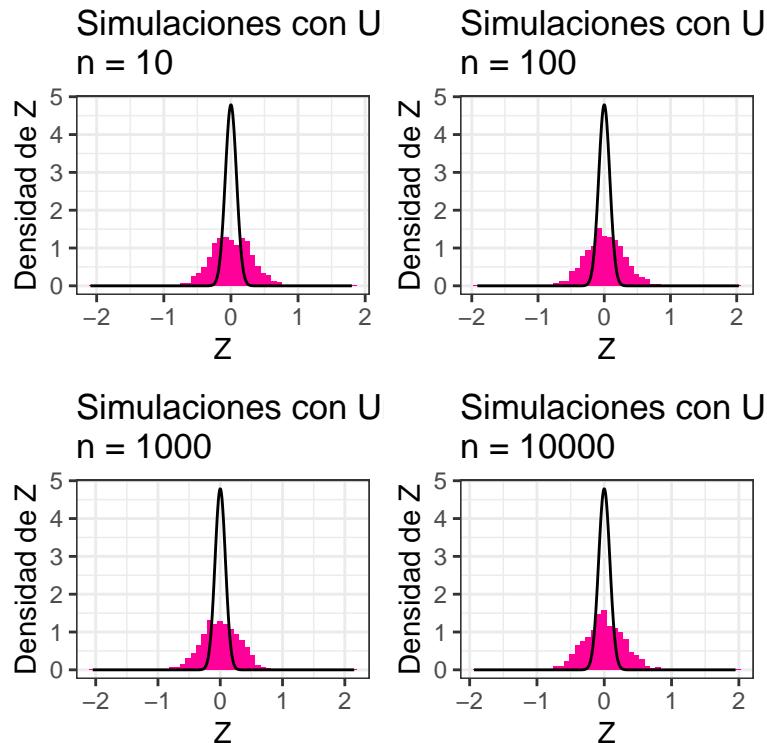
La binomial se ve así:
TeoremaCentralLimitete(distribucion = rbinom, mu = 4.5, sigma = 2.475, size = 10,
prob = 0.45, distname = "Binomial(10,0.45)")



```
Poisson:  
TeoremaCentralLimites(lambda = 1, distname = "Poisson(1)")
```



```
E inclusive uniformes:  
TeoremaCentralLimitte(distribucion = runif, mu = 1/2, sigma = 1/12,  
distname = "Uniforme(0,1)")
```



Experimenta con otras distribuciones ¿puedes encontrar alguna para la que no funcione?

B.16.2 Ejercicio

Repite la programación del teorema del límite central pero ahora tomando las X_k con distintas distribuciones siempre y cuando X_k tenga media μ_k finita y las variables aleatorias satisfagan la condición de Lindberg (una forma de hacerlo es teniendo varianzas finitas que no incrementan con la k).

Casella, George, and Roger L Berger. 2002. *Statistical Inference*. Vol. 2. Duxbury Pacific Grove, CA.

Fisher, Nicholas I. 1995. *Statistical Analysis of Circular Data*. cambridge university press.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis*. CRC press.

Hyndman, Rob J, and Yanan Fan. 1996. "Sample Quantiles in Statistical Packages." *The American Statistician* 50 (4): 361–65.

Mardia, Kanti V, and Peter E Jupp. 2009. *Directional Statistics*. Vol. 494. John Wiley & Sons.

Myatt, Glenn J, and Wayne P Johnson. 2007. *Making Sense of Data*. Wiley Online Library.

Panaretos, Victor M. 2016. "Statistics for Mathematicians." *Compact Textbook in Mathematics*. Birkhäuser/Springer 142: 9–15.

Peck, Roxy, Chris Olsen, and Jay L Devore. 2015. *Introduction to Statistics and Data Analysis*. Cengage Learning.

Pewsey, Arthur, Markus Neuhauser, and Graeme D Ruxton. 2013. *Circular Statistics in r*. Oxford University Press.

Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 2003. *Model Assisted Survey Sampling*. Springer Science & Business Media.

SURI, NNR MURTY RANGA, M Narasimha Murty, and G Athithan. 2019. *Outlier Detection: Techniques and Applications*. Springer.

Wolfe, Douglas A, and Grant Schneider. 2017. *Intuitive Introductory Statistics*. Springer.