Introducción a la Probabilidad y la Estadística

Dr. Jaime Lincovil C.

2020-11-08

Índice general

Objetivos de la monografía

Dinámica de nuestra comunicación

Proceso (operaciones) de inferencia estadística.
Definición. 0.1. Una **operación** Op aplicada sobre un objeto A lo transforma en un único objeto B. La aplicación de Op sobre A produciendo B es simbolizado por:

$$\mathrm{Op}(A) = B.$$

Una **lista de aplicaciones operaciones** $\mathrm{Op}_1, \ldots, \mathrm{Op}_n$ aplicadas sobre A de forma secuencial e independintes produciendo $B_1, \ldots, B_n$ es simbolizado por

$$\mathrm{Op}_1(A) = B_1, \ldots, \mathrm{Op}_n(A) = B_n.$$

Formalmente, una operación Op puede ser representada por una función:

$$\begin{aligned} \mathrm{Op} : \mathrm{Dom} &\to \mathrm{Rec}, \\ A &\mapsto \mathrm{Op}(A), \end{aligned}$$

en donde Dom y Rec son el dominio y recorrido de Op, respectivamente.
Ejemplo. 0.1. Considere la operación de división por 2 simbolizada por Op_1 y exponencializacion $\{\}^2$ simbolizada por Op_2. Luego Op_1 y Op_2 aplicadas sobre el número $A = 10$ produce:

$$\mathrm{Op}_1(10) = 5/2 = 5$$

y

$$\mathrm{Op}_1(10) = 10^2 = 100.$$

Es decir, Op_1 transforma 10 en 5 y Op_1 transforma 10 en 100.

Definición. 0.2. La **aplicación compuesta de dos operaciones** Op_1 y Op_2 sobre A es la composición

$$\mathrm{Op}_2(\mathrm{Op}_1(A)).$$

Dado que Op_1 y Op_2 son funciones, formalmente, esta aplicación secuencial de estas sobre A es la composición de funciones $\mathrm{Op}_2 \circ \mathrm{Op}_1(A)$.
Ejemplo. 0.2. La aplicación secuencial de las aplicaciones /2 y y luego $\{\}^2$ sobre 10 es

$$\{\}^2 \to (/2 \to 10) = (10/2)^2 = 25.$$

En este trabajo estaremos interesados en **operaciones estadísticas**, es decir, operaciones transformen **datos** en resumenes, gráficos e inferencias. Mencionamos aquellas que serán de nuestro interes.

1. Ejecutar la planificación de un experimento de tal manera de obtener información parcial x de un fenomeno de interes. La información x puede ser registrad como un vector o una matriz.
2. Cuantificar la incertidumbre de un evento A que podría ocurrir al ejecutar un experimento, simbolizado por "Prob(A)".
3. Estructurar datos x, simbolizado por "Estr (x)".
4. Explorar los datos x, simbolizado por "Plot (x)".
5. Resumir los datos x, simbolizado por "Res (x)".
6. Estimar un parámetros poblacionales θ en base a la muesra x, simbolizado por "$\widehat{\theta}(x)$".
7. Predecir nuevas observaciones del experimento a partir de x, simbolizado por "Pred(x)".
8. Contrastar hipótesis H_0 contra H_1.

Organización de la monorafía

El Capitulo 1 presenta los elementos básicos de análisis exploratorio y resumen de datos uni y bivariado.

Capítulo 1

Operaciones de análisis exploratorio y resumen de datos

1.1. Conceptos elementales

En esta sección presento conceptos básicos para este curso.
Definición. 1.1. (a) Una **población** es un conjunto de personas, objetos o eventos, de los cuales nos interesa estudiar algunas de sus características. (b) **Medidas poblacionales** son listas de mediciones de ciertas cantidades provenientes de cada individuo o elemento de la población. (c) Un **parámetro poblacional** es una característica general de la población.

Presento tres ejemplos de población, unidades poblacionales y parámetros poblacionales.
Ejemplo. 1.1. (a) *Población*: total de autos cuyos modelos salieron a la venta en un determinado intervalo de tiempo y funcionan en una determinada país. (b) *Medidas poblacionales*: razón de millas recorrida por galón de combustible, número de cilindro, caballos de fuerza, tipo de caja de cambios. (c) *Parámetro poblacional*: media poblacional de razón de millas recorrida por galón de combustible, media poblacional de caballos de fuerza y proporción de autos con caja de cambio automático.
Ejemplo. 1.2. (a) *Población*: individuos de una cierta comuna de una ciudad que poseen problemas de insulina. (b) *Medidas poblacionales*: presencia o no de menarquía (primera hemorragia menstrual de la mujer), edad, sexo, nivel de igf1 (factor de crecimiento insulínico tipo 1,), etapa de la pubertad y nivel testicular. (c) *Parámetro poblacional*: promedio de edad poblacional, frecuencias relativas de etapas de la pubertad poblacional y media poblacional de igf1.

Ejemplo. 1.3. (a) *Población*: pacientes de una determinada comunidad que reciben tres métodos diferentes de ventilación durante la anestesia. (b) *Medidas poblacionales*: concentración de folato (microgramos por litro) y tipo de ventilación recibida por los pacientes[1]. *Parámetros poblacionales*: media de folato poblacional y proporciones de factores de niveles de ventilación.

Los conceptos de unidad de observación (o muestral), muestra y base de datos son presentados a continuación.
Definición. 1.2. (a) **Unidades de observaciones:** es un grupo de elementos de una población de la cual es posible obtener información de manera conjunta. Comunmente es conocido como una **muestra** de la población. (b) Un **experimento** es una operación bien planificada y controlada que se realiza para obtener información parcial desde unidades de observación de una determinada población. (c) La información cuantificada o codificada derivada de un experimento es llamada de **datos**. (d) Una **base de datos** es una forma ordenada organizados la muestra de tal manera de que estos puedan ser trabajados por programas computacionales.

Nota: en general, el experimento debe ser realizado por un procedimiento que evite el sesgo personal. Esto podría lograrse utilizando algún *mecanismo aleatorio*. Sin embargo, esto no siempre es posible debido a problemas ético y legales. Esto ocurre en el caso de estudios en el área de la salud donde aplicar un mecanismo aleatorio podría poner en peligro la integridad de los pacientes. En este caso, el experimento se reduce un estudio de los datos disponibles.

Presentamos algunos ejemplos.
Ejemplo. 1.4. (a) *Unidad de observación*: grupo de autos de una determinada ciudad. (b) *Experimento*: escoger autos de forma aleatoria de una determinada ciudad y medir razón de millas recorrida por galón de combustible, número de cilindro, caballos de fuerza, tipo de caja de cambios. (c) *Muestra*: lista de estos valores observados en la medición. (d) *Base de datos*: vector columna con los datos de una característica observada o matriz rectangular con toda la información codificada.

Mostraremos un ejemplo de datos en forma de vector y matriz con la base de datos *mtcars* del paquete de *R* llamado *MASS*. Esta base de datos está en formato de *data frame* y fue publicado en 1974 en la revista *Motor Trend US magazine*. La base *mtcars* contiene datos de 32 autos de modelos diferentes. Por ejemplo, la razón de millas por galón de combustible (mpg) en forma de vector columna de los primeros y últimos valores

$$\text{mpg} = (21{,}0; 21{,}0; 22{,}8; 21{,}4, \ldots, 27{,}3; 26{,}0; 30{,}4; 15{,}8; 19{,}7; 15{,}0; 21{,}4)^\top.$$

[1] (i) \$N2O + O2\$, 24h: 50 % de óxido nitroso y 50 % de oxígeno, de forma continua durante 24 horas; (ii) N2O + O2, op: 50 % de óxido nitroso y 50 % de oxígeno, solo durante el funcionamiento; (iii) $O2{,}24h$: sin óxido nitroso pero entre un 35 % y un 50 % de oxígeno durante 24 horas

En la siguiente matriz se presenta en la primera columna el modelo de los autos estudiados y en las restantes columnas las variables medidas para los 6 primeros autos estudiados. Aquí, por ejemplo, "Mazda RX4" es el modelo del primer auto, cuya *mpg* o razón de millas por galón es igual a 21 y su *cyl* o número de cilindros del motor es igual a 6.

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Te presento un ejemplo en el area de la salud.

Ejemplo. 1.5. (a) *Unidad de observación*: pacientes de un determinado hospital. (b) *Experimento*: examen de laboratorio en donde se escogen individuos para medir: presencia o no de menarquía (primera hemorragia menstrual de la mujer), edad, sexo, nivel de igf1 (factor de crecimiento insulínico tipo 1,), etapa de la pubertad y nivel testicular. (c) *Muestra*: mediciones de estos valores. (d) *Base de datos*: vector columna con los datos de una característica observada o matriz rectangular con toda la información codificada.

La siguiente base de datos llamada *juul2* proviene dal paquete *ISwR* del programa R. Contiene información de pacientes con problema de insulina. Aquín, por ejemplo, *age* contiene la edad de los pacientes y *menarche* presencia o no de menarquía (primera hemorragia menstrual de la mujer).

```
##     age height menarche sex igf1 tanner testvol weight
## 1    NA     NA       NA  NA   90     NA      NA     NA
## 2    NA     NA       NA  NA   88     NA      NA     NA
## 3    NA     NA       NA  NA  164     NA      NA     NA
## 4    NA     NA       NA  NA  166     NA      NA     NA
## 5    NA     NA       NA  NA  131     NA      NA     NA
## 6  0.17     NA       NA   1  101      1      NA     NA
## 7  0.17     NA       NA   1   97      1      NA     NA
## 8  0.17     NA       NA   1  106      1      NA     NA
## 9  0.17     NA       NA   1  111      1      NA     NA
## 10 0.17     NA       NA   1   79      1      NA     NA
## 11 0.17     NA       NA   1   43      1      NA     NA
## 12 0.17     NA       NA   1   64      1      NA     NA
## 13 0.25     NA       NA   1   90      1      NA     NA
## 14 0.25     NA       NA   1  141      1      NA     NA
## 15 0.42     NA       NA   1   42      1      NA     NA
```

Es normal que al aplicar la medición ocurran errores o problemas con los instrumentos de medición. Las normas éticas científicas exigen no ocultar esto sino informarlas. Por ejemplo, un símbolo **NA** indica que el valor fue perdido. El

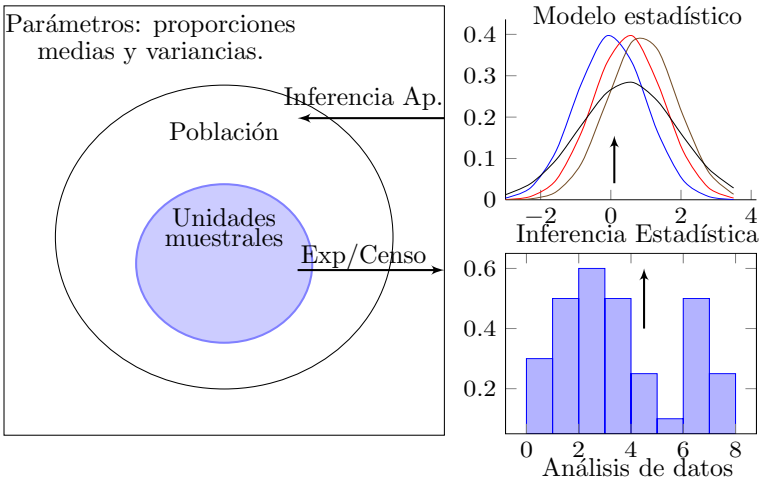
símbolo **NULL** indica que el valor obtenido fue nulo o no tiene validez. Se simboliza por **Inf** para valores que salieron fuera de los límites de los valores de medición del instrumento.

Presentamos otro ejemplo dentro del área de la salud.
Ejemplo. 1.6. *Unidad de observación*: pacientes de un determinado hospital. *Experimento*: examen de laboratorio en donde se reciben pacientes y se les mide concentración de folato (microgramos por litro) y factores con niveles de N2O+O2. *Muestra*: lista de estos valores. *Base de datos*: vector o matriz rectangular con estos datos cuantificados.

```
##    folate ventilation
## 1     243  N2O+O2,24h
## 2     251  N2O+O2,24h
## 3     275  N2O+O2,24h
## 4     291  N2O+O2,24h
## 5     347  N2O+O2,24h
## 6     354  N2O+O2,24h
## 7     380  N2O+O2,24h
## 8     392  N2O+O2,24h
## 9     206   N2O+O2,op
## 10    210   N2O+O2,op
```

Podemos resumir el proceso de análisis de datos en la siguiente figura

```
knitr::include_graphics('Gran_Cuadro_Inferencia.pdf')
```



-

Todo comienza con el interés por conocer atributos desconocidos de una determinada población. Aquí son llamados de parámetros poblacionales. Para *aprender* de estos parámetros se realiza un experimento y recopilar información. El siguiente paso es *explorar* las características generales de los datos. Luego bus-

caremos resumirlos para sacar conclusiones preliminares sobre los parámetros poblacionales.

1.2. Variables observadas en experimentos

Un elemento relevante de la planificación de un experimento científico es el conjunto de variables que será medido desde las unidades de observación.
Definición. 1.3. (a) Una **variable** X es una característica de interés que posee cada elemento de una población y que podemos medir. (b) Una variable es **cuantitativa** si sus valores son números y representan una cantidad. (c) Una variable es **cualitativa** si sus valores representan una cualidad, un atributo o una categoría. Se les llama también variables categóricas. (d) Una lista

$$x_1, x_2, \dots, x_n$$

de observaciones de una variable X obtenidas al desarrollar un experimento es llamada de **datos observados** de X extraída desde una lista de n unidades de observación. El número n será llamado de **tamaño de los datos**.

Presentamos algunos ejemplos.
Ejemplo. 1.7. - Ejemplo de variables cuantitativas:

- Razón de millas por galón de combustible.
- Número de cilindros de motores.
- Número de caballos de fuerza de motores.
- Peso en Kilogramos.
- Nivel de insulina en la sangre.
- Temperatura en grados Celsius (o Fahrenheit).
- Año de lanzamiento de un modelo de auto.

Ejemplo. 1.8. - Ejemplo de variables cualitativas

- Tipo de motor.
- Modelo del auto.
- Sexo.
- Tipo de ventilación en pacientes.

En general, las variables de experimentos son clasificadas en dos tipo.
Definición. 1.4. Tipos de variables cuantitativas. (a) **Discreta:** si el conjunto de todos sus posibles valores tiene un número finito de elementos, o bien es infinito, pero se pueden numerar uno por uno de acuerdo al conjunto de número naturales $\{0, 1, 2, 3, ...\}$. Noten que, esto implica que la variable puede asumir solamente un conjunto finito de valores dentro de un intervalo (a, b). Una variable discreta puede asumir valores con decimales. (b) **Continua:** si puede tomar todos los valores posibles dentro de un intervalo de números reales, como por ejemplo (a, b) o $[a, b]$.

Presentamos algunos ejemplos.

Ejemplo. 1.9. - Ejemplo de variables discretas:

- Número de cilindros de motores.
- Número de caballos de fuerza de motores.
- Una variable hipotética con valores: 0, 0.5, 1, 1.5, 2, 2.5, ...

.

Ejemplo. 1.10. - Ejemplo de variables continuas:

- Razón de millas por galón de combustible.
- Peso en Kilogramos.
- Nivel de insulina en la sangre.
- Temperatura en grados Celsius (o Fahrenheit).

Luego de obtener los resultados de un experimento, el investigador debe codificar los resultados para poder construir la base de datos. En el caso de variables cualitativas, esto es hecho clasificándola en dos tipos de escala de medición.

Definición. 1.5. Escalas de mediciones de variables cualitativas. (a) **Escala nominal:** si sus valores son etiquetas o atributos y no existe un orden jerárquico entre ellos. (b) **Escala ordinal:** si sus valores son etiquetas o atributos pero existe un cierto orden jerárquico entre ellos.

Ejemplo. 1.11. - Ejemplo de variables cualitativas en escala nominal:

- Tipo de motor.
- Sexo.
- Tipo de ventilación en pacientes.
- Idioma.
- Nacionalidad.

Ejemplo. 1.12. - Ejemplo de variables cualitativas ordinal:

- Grupo etario: lactante, niños, adolescente, adulto
y tercera edad.
- Clasificación de productos tecnológicos: básico,
gama media y gamma alta.
- Posición de llegada en una competencia.
- Talla de ropa: S, M, L, etc.
- Tipo de ventilación en pacientes.

Las variables cuantitativas también son clasificadas según su escala de medición.

Definición. 1.6. Escala de intervalo: existe una noción de distancia entre los valores de la variable y no existe necesariamente el valor natural cero como indicador de ausencia de algo. En este caso el cero u otro valor representa un punto de cambio de "nivel". En general, la escala una variable en escala de intervalo es establecido vía consensos científicos o de instituciones para estandarizar procesos.

Definición. 1.7. Escala de razón si los valores de la variable tienen un sentido físico de *intensidad* y existe el cero indicando ausencia de valor. En este caso el cero es el origen de los valores a ser medidos.

Ejemplo. 1.13. - Ejemplo de variables cuantitativas medidas en escala de intervalo:

```
- Escala de notas: de 0 a 7, de 0 a 10,
utilizando letras con números.
- Puntaje PSU.
- El pH,  una medida de acidez o alcalinidad de
una sustancia.
- Temperatura en grados Celsius (o Fahrenheit):
puede asumir valores positivos, negativos
y en particular ser igual a 0.
```

Ejemplo. 1.14. - Ejemplo de variables cuantitativas medidos en escala de razón:

```
- Razón de millas por galón de combustible.
- Número de cilindros  de motores.
- Número de caballos de fuerza de motores.
- Peso en Kilogramos.
- Nivel de insulina en la sangre.
```

1.3. Agrupamiento de variables

Definición. 1.8. (a) Una **clase** es un agrupamiento de *categorías* en el caso de variables cualitativas, o de *intervalos numéricos* en el caso de variables cuantitativas. (b) Una **marca de clase** es un dato que representa a una clase. En el caso de los intervalos la marcha de clase es el punto medio de los intervalos que se obtiene con el promedio de los extremos.

Ejemplo. 1.15. Consideremos los valores de las variables *mpg* de la base de datos *mtcars*. Clasificando los valores en intervalos de largo 4,4 mil millas por galón obtenemos los intervalos

$$[0; 4{,}4), [4{,}4; 8{,}8), [8{,}8; 13{,}2)[13{,}2; 17{,}6), [17{,}6; 22{,}0), [22{,}0; 26{,}4), [26{,}4; 30{,}08]$$

La marca de clase para el intervalo $[13{,}2; 17{,}6)$ es el punto 15,4. Analogamente para los otros casos.

1.4. Operación de análisis exploratorio de datos

Definición. 1.9. La **estadística descriptiva** es la area de la **Estadistica** que se preocupada de la aplicación de operaciones que ayudan a describir, mostrar y resumir, la información de un conjunto de datos.

Definición. 1.10. La parte de la Estadística Descriptiva que utiliza herramientas para **visualizar** caracteristicas generales de los datos es llamada de "Analisis exploratorio de Datos". Tales herramientas son llamadas de **operaciones exploratorias**.
Ejemplo. 1.16. Ejemplo de operaciones exploratoias son construir: tablas de frecuencias, graficos de tallos y hojas, histogramas, diagrama de caja y bigotes, diagrama de dispersión, etc.

1.4.1. Operaciones de exploración grafica de los datos: sin grupos

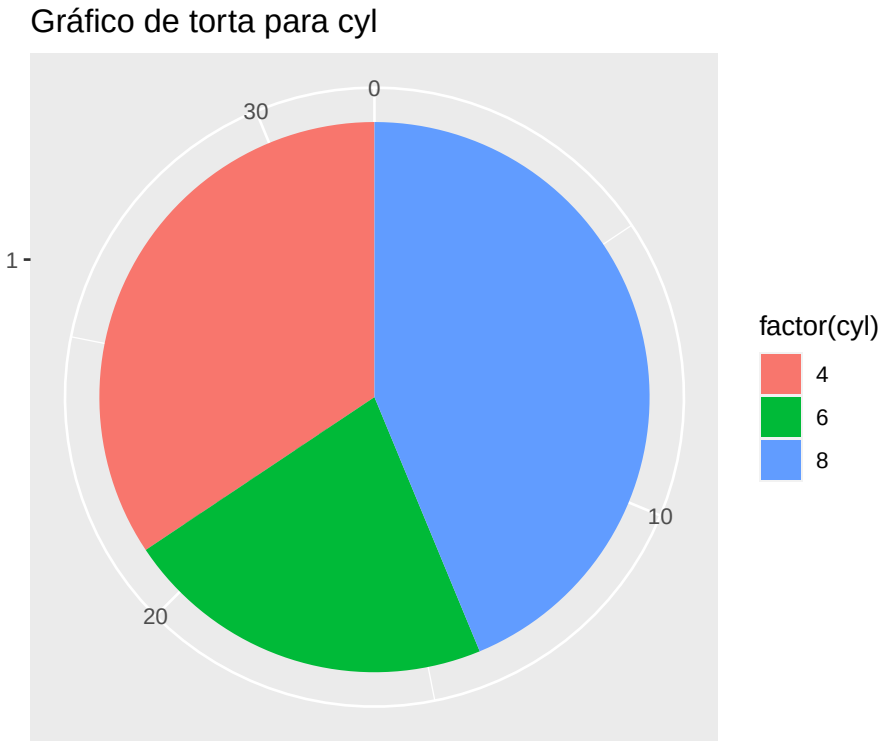
Los gráficos estadísticos nos entregarán información representada en el plano cartesiano bi o tridimensional de los valores de la variable X que acumula las mayores frecuencia y como se distribuyen los otros datos en torno de estos valores centrales, entre otros.
Definición. 1.11. Gráfico de tallos y hojas: su aspecto es muy similar al de un histograma dibujado horizontalmente. A los dígitos del primer decimal de los números que aparecen listados en la parte *derecha* de | del diagrama se les llama **hojas** y a la parte izquierda de la parte no decimal se le llama **tallo**.
Definición. 1.12. Gráfico circular o de torta: para variables cualitativas o bien para variables cuantitativas agrupadas, se pueden elaborar gráficas de pastel, también llamadas pie charts. Estas son gráficas circulares divididas en sectores que permiten comparar visualmente las frecuencias porcentuales de los valores observados de una variable.
Definición. 1.13. Gráfico de barras: las gráficas de barra ayudan a visualizar los valores de una variable que ocurren con mayor o menor frecuencia y a comparar cualitativamente estas frecuencias.
Definición. 1.14. Histograma:: un histograma es una gráfica muy similar a una gráfica de barras. Adquiere este nombre cuando existe un orden entre los valores de la variable a graficar. Nuevamente, para cada valor, categoría o clase de la variable, se asocia una barra cuya altura es la frecuencia con la que se observa la categoría.
Definición. 1.15. Boxplot: para representar los datos con base en valores que dividan los datos en partes estratégicas es adecuado usar un gráfico de tallos y hojas. Para su diseño es necesario el valor mínimo, el primer (C1), segundo (C2), tercer cuartil (C3) y el máximo de los datos. La caja es formada por C1 y C3 en cuyo centro se encuentra C2 (la mediana.)

Presento algunos ejemplos de transformaciones de los datos mpg en gráficos.

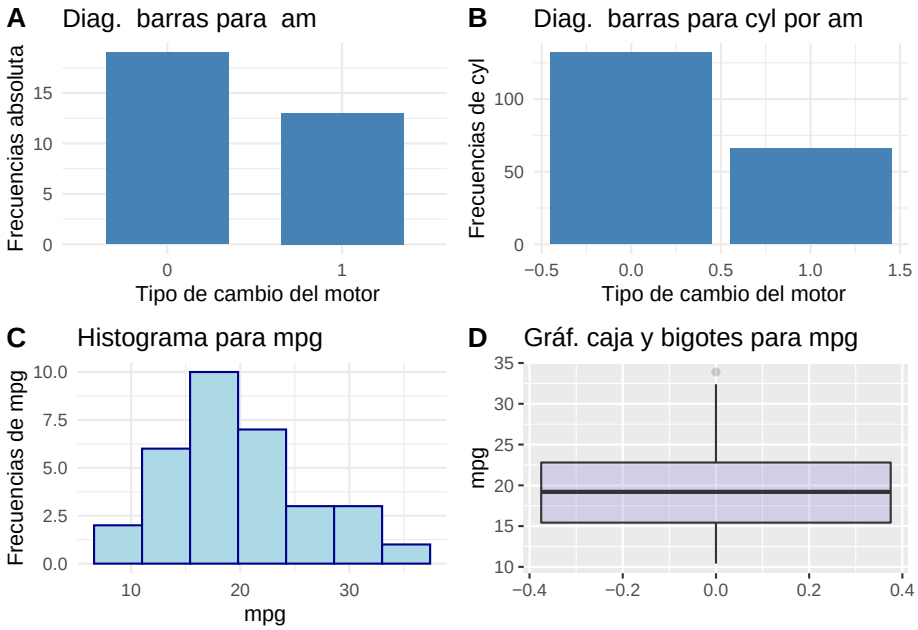
```
##
##   The decimal point is at the |
##
##   10 | 44
##   12 | 3
##   14 | 3702258
```

```
##   16 | 438
##   18 | 17227
##   20 | 00445
##   22 | 88
##   24 | 4
##   26 | 03
##   28 |
##   30 | 44
##   32 | 49
```

Interpretación: el valor más frecuente de los valores observados de *mpg* son las medidas cuya parte entera es 14. En menor cantidad, pero también frecuentes son los valores con parte entera igual a 18 y 20.



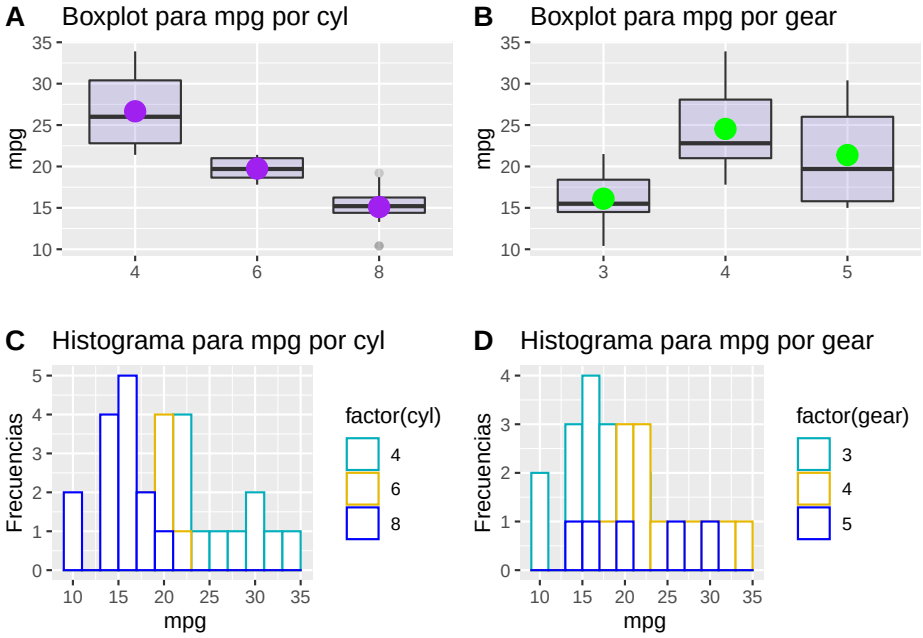
Interpretación: dentro de la base de datos mtcars, los autos con 8 y 4 ciclos son los más frecuentes.



Interpretación: en la base de datos *mtcars* los autos con cambio de velocidad automático (0) es mayor que los autos con cambios manual (1). Los autos cambio de velocidad automático suman un mayor número de cilindros que los autos con cambios manuales. Las altas frecuencias de mpg se concentran en torno a 20 millas por galón.

1.4.2. Operaciones de exploración gráfica de los datos: por grupos

Los gráficos por grupos nos ayudarán a comparar las características de los datos separados por unidades muestrales con alguna característica en común.



Interpretación: Los autos con 4 cilindros tienen un mayor nivel de *mpg* que los otros casos (ver media representado por el punto morado). Los autos con 4 engranajes tienen mayor niveles de *mpg*. Los autos con 8 cilindros concentran altas frecuencias en niveles más bajos que los de 4 y 6 cilindros.

1.4.3. Operación de analisis de frecuencias

Una forma de resumir los datos es través del conteo de número de veces en las cuales aparecen ciertas observaciones en los datos.

Definición. 1.16. Consideremos una muestra $x_1, \dots, x_n$ de n valores en donde x_i representa la i-ésima observación.

- La **frecuencia absoluta** de x_i, simbolizado por f_i, es el número de veces que su valor se repite en la muestra.
- La **frecuencia absoluta acumulada** de x_i, simbolizado por F_i , es la suma

$$F_i = \sum_{j:x_j \leq x_i} f_j,$$

es decir la suma de las frecuencias absolutas de todos los valores menores o iguales a x_i.

- La **frecuencia relativa** de x_i es el porcentaje $h_i = (f_i/n) \times 100$, es decir el porcentaje de la frecuencia absoluta en el total de la muestra.

- La **frecuencia relativa acumulada** de x_i es la suma

$$H_i = \sum_{j:x_j \leq x_i} h_j,$$

es decir la suma de los frecuencias relativas de todos los valores menores o iguales a x_i.

Las frecuencias h_i y H_i también pueden ser calculado con valores sobre el intervalo $[0, 1]$ con interpretaciones equivalentes a los porcentajes en la escala $[0\,\%; 100\,\%]$.

Una forma de resumir los valores de la variable *mpg* es mediante una operación que los transforme en tablas de frecuencia por los valores puntuales o por intervalo de clase.

En el primer caso presentamos una tabla de frecuencia de los valores puntuales.

```
freqs(mtcars$mpg)
```

```
##    data fi Fi    hi      Hi
## 1  10.4  2  2 6.250   6.250
## 2  13.3  1  3 3.125   9.375
## 3  14.3  1  4 3.125  12.500
## 4  14.7  1  5 3.125  15.625
## 5  15.0  1  6 3.125  18.750
## 6  15.2  2  8 6.250  25.000
## 7  15.5  1  9 3.125  28.125
## 8  15.8  1 10 3.125  31.250
## 9  16.4  1 11 3.125  34.375
## 10 17.3  1 12 3.125  37.500
## 11 17.8  1 13 3.125  40.625
## 12 18.1  1 14 3.125  43.750
## 13 18.7  1 15 3.125  46.875
## 14 19.2  2 17 6.250  53.125
## 15 19.7  1 18 3.125  56.250
## 16 21.0  2 20 6.250  62.500
## 17 21.4  2 22 6.250  68.750
## 18 21.5  1 23 3.125  71.875
## 19 22.8  2 25 6.250  78.125
## 20 24.4  1 26 3.125  81.250
## 21 26.0  1 27 3.125  84.375
## 22 27.3  1 28 3.125  87.500
## 23 30.4  2 30 6.250  93.750
## 24 32.4  1 31 3.125  96.875
## 25 33.9  1 32 3.125 100.000
```

Interpretación: El 25 % de los medidas de *mpg* en la muestra son menores o igual a 15,2 millas por galón. Aproximadamente el 53,13 % de los medidas de *mpg* son menores a 19,2 millas por galón. Aproximadamente el 78,13 % de los medidas de *mpg* son menores a 22,8 millas por galón.

En segundo lugar, presentamos una tabla de frecuencia de los los intervalos presentados anteriormente.

```
##   marca fi Fi     hi      Hi
## 1   2.2  0  0  0.000   0.000
## 2   6.6  0  0  0.000   0.000
## 3  11.0  2  2  6.250   6.250
## 4  15.4 10 12 31.250  37.500
## 5  19.8 11 23 34.375  71.875
## 6  24.2  4 27 12.500  84.375
## 7  30.4  5 32 15.625 100.000
```

Interpretación:

1.5. Operación de resumen de datos univariados

1.5.1. Medidas de tendencia central

Definición. 1.17. La media o **media** aritmética es simplemente el *promedio* de la muestra $x_1, \ldots, x_n$ simbolizado por $\bar{x}_n$ y dada por:

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Definición. 1.18. La **moda** es el valor que aparece con mayor frecuencia en el conjunto de datos, si lo hubiera. En general, los datos pueden ser uni, bi o multimodales (varias modas).

Definición. 1.19. Simbolisemos por $x_{(k)}$ el valor de la muestra ordenada que esta en el k-ésimo lugar. La **mediana** es el dato ordenado de en medio, esto es:

(a) Si el número de datos n es par, entonces existen dos datos ordenados de en medio y la mediana es el promedio de estos dos números, esto es $(x_{(n/2)} + x_{(n/2)+1})/2$.

(b) Si el número de datos n es impar, entonces el dato ordenado de en medio es $x_{(n-1)/2}$ y esta es la mediana.

Ejemplo. 1.17. Presentamos la media, moda y mediana de la los datos *hp* (caballos de fuerza) de la base de datos *mtcars*: $\bar{x}_n = 146{,}7$; Mediana = 123 ; Moda = $(110, 175, 180)$ los cuales con una frecuencia absoluta iguala 3.

1.5.2. Medidas de dispersión

Definición. 1.20. La **varianza** s_x^2 es un promedio de la distancia al cuadrado de cada uno de los datos x_i respecto de la media $\bar{x}_n$ y es la medida de dispersión más comúnmente usada calculada por:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

Definición. 1.21. A la raíz cuadrada positiva de la varianza se le llama **desviación estándar** o desviación típica, y se le simboliza por la letra s_x^2. Así, para su cálculo se usa la siguiente fórmula:

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}.$$

Definición. 1.22. Al promedio de los valores absolutos de las diferencias entre los datos y la media se le llama **desviación media**. Más específicamente, supongamos que $\bar{x}_n$ es la media de los datos numéricos $x1, \ldots, x_n$, entonces la desviación media se denota por dm_x y se define como el siguiente promedio

$$\mathrm{dm}_x = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}_n|.$$

Enfasis. Cuando en s_x^2, s_x y dm_x dividimos por $(n-1)$ en lugar de n, decimos que s_x^2, s_x y dm son las versiones **modificadas** de la variancia, desviación estandar y dm.

Definición. 1.23. Ahora definiremos el rango de una colección de números $x_1, \ldots, x_n$. Para calcular esta cantidad es necesario identificar el dato más pequeño $x_{(1)}$ y el dato más grande $x_{(n)}$. El rango de la colección de números dada se denota por Ran y es simplemente el dato mayor menos el dato menor:

$$\mathrm{Ran} = x_{(n)} - x_{(1)}.$$

Definición. 1.24. Sea $x_1, \ldots, x_n$ una colección de n observaciones de una variable cuantitativa. Sea $\bar{x}_n$ su media y sea s_x su desviación estándar. Al cociente de variación se le llama coeficiente de variación y se le denota por cv_x, suponiendo por supuesto que $\bar{x}_n \neq 0$. Es decir:

$$cv_x = \frac{s_x}{\bar{x}_n} \times 100.$$

Ejemplo. 1.18. Presentamos las anteriores medidas de dispersión para los datos *hp* (caballos de fuerza) de la base de datos *mtcars*: $s_x^2 = 4700{,}87$; $s_x = 68{,}56$; $dm_x = 56{,}48$, Ran= 283 y $cv_x = 46{,}74$.

1.5.3. Percentiles, cuantiles, cuartiles y deciles

Definición. 1.25. El valor númerico de los datos ordenados que deja a su izquierda (o son menores que) el $p\times 100\,\%$ de los datos es llamado de **percentiles** de orden a $p \times 100$, para una proporción $p \in (0, 1)$. El equivalente a un percentil de orden $p \times 100$ es el **cuantil** de orden p, el cual es el valores de la muestra que deja a su izquerda la proporción p de los datos menores a el. Percentiles y cuantiles son valores equivalentes. en este texto nos trabajaremos con el termino PERCENTIL.

Los percentiles de orden $0{,}25\,\%$, $0{,}50\,\%$ y $0{,}75\,\%$ son llamos de **primer cuartil** simbolizado por C_1, **segundo cuartil** simbolizado por C_2 y **tercer cuartil** simbolizdo por C_3, respectivamente. Los **cuartiles** dividen a los datos ordenados en 4 partes con aproximadamente el mismo porcentaje de datos.

Cuando $p = 0{,}1, 0{,}2, \ldots, 0{,}9$, a los percentiles correspondientes se les llama **deciles**. Los **deciles** dividen a los datos ordenados en 4 partes con aproximadamente el mismo porcentaje de datos.

En otras ocasiones se requiere dividir al conjunto de datos en cien porcentajes iguales, y entonces cuando $p = 0{,}01, 0{,}02, \ldots, 0{,}99$ a los cuantiles correspondientes se les llama percentiles.
Ejemplo. 1.19. Presentamos los valores extremos y los cuartiles de la variable *hp*. $x_{(1)} = 52$, $Q_1 = 96{,}5$, $Q_2 = 123$, $Q_3 = 180$ y $x_{(n)} = 335$.

1.5.4. Medidas de asimetría y curtosis

Definición. 1.26. La cantidad que llamaremos **coeficiente de asimetría** κ (en inglés *skewness*) es una medida de la asimetría (falta de simetría) de un conjunto de datos dados por:

$$\kappa = \frac{1}{n \times s_x^3} \sum_{i=1}^{n} (x_i - \bar{x}_m)^3.$$

Si $\kappa < 0$, la distribución de frecuencias es **asimétrica a la izquerda**. Si $\kappa > 0$, la distribución de frecuencias es **asimétrica a la derecha**. Si $\kappa = 0$, la distribución de frecuencias es **simétrica**.
Definición. 1.27. Sea $x_1, \ldots, x_n$ una colección de datos numéricos con media $\bar{x}_n$ y desviación estándar s_x. La **curtosis**, que denotaremos por la letra ν, es un número que se define de la siguiente manera:

$$\nu = \frac{1}{n \times s_x^4} \sum_{i=1}^{n} (x_i - \bar{x}_m)^4.$$

Si $\nu < 0$, la distribución de frecuencias es **platicúrtica**, decaimiento lento, colas largas. Si $\nu > 0$, la distribución de frecuencias **leptocúrtica**, decaimiento

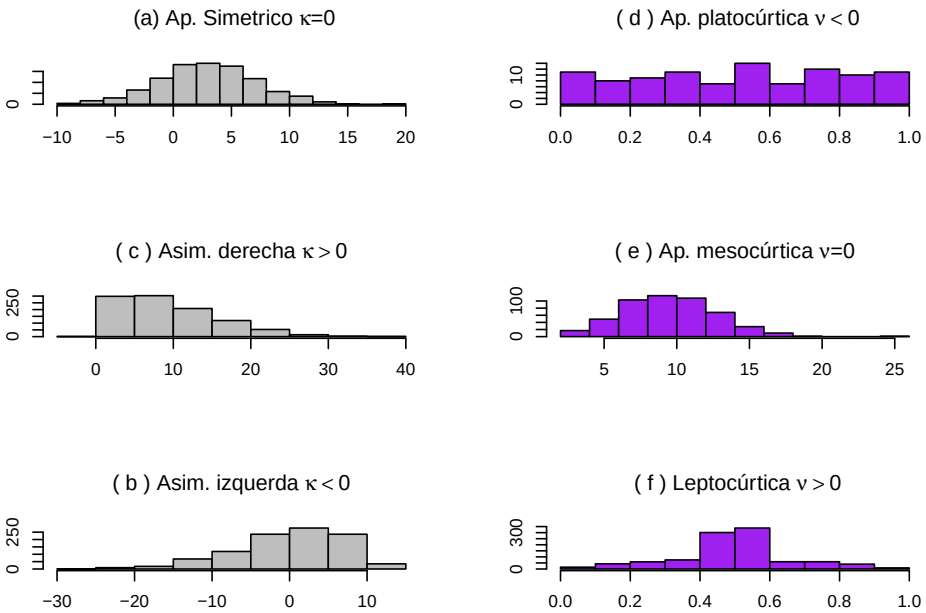
rápido, colas cortas. Si $\nu = 0$, la distribución de frecuencias es **mesocúrtica**, frecuencia aproximada por la densidad Gaussiana.

La siguiente figura muestra seis gráficos demostrativos de los posibles casos de formas de asimetria y curtosis que podrian tomar los datos representados histogramas.

```
## Loading required package: stats4
```

```
##
## Attaching package: 'sn'
```

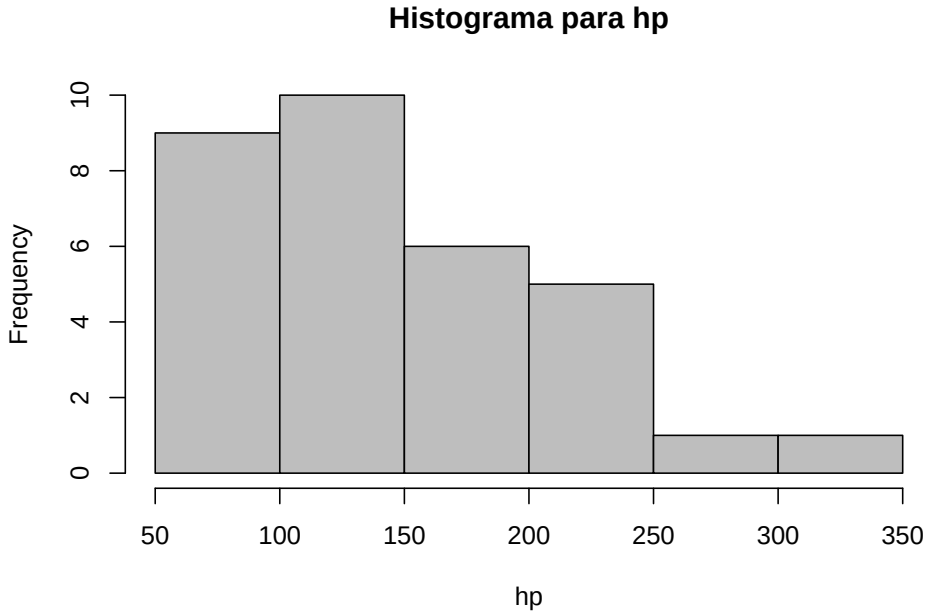
```
## The following object is masked from 'package:stats':
##
##     sd
```



Ejemplo. 1.20. Coeficiente de asimetría y curtosis para los datos *hp* de la base *mtcars* son $\kappa = 0{,}73$ (asimetría a la derecha) y $\nu = -0{,}135$ (platocúrtica), respectivamente.

Vemos esto en un histograma

```
hist(mtcars$hp, main = "Histograma para hp", xlab =
       "hp", col = "gray" )
```



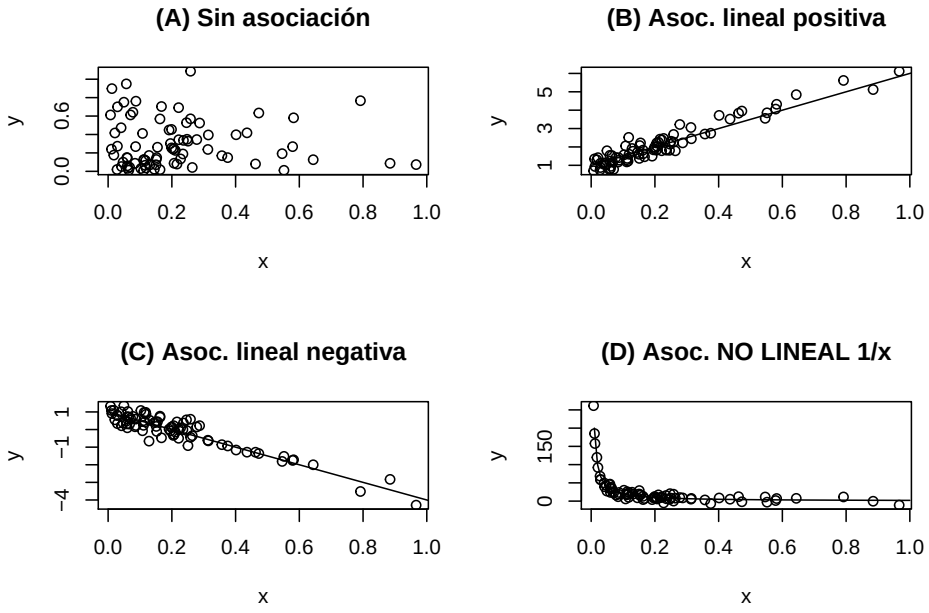
1.6. Operación de resumen de datos bivariados

En esta sección presentaron un análisis de asociación de dos tipos: correlación lineal entre dos variables continuas y analisis de asociación entre dos variables categóricas.

1.7. Gráficos de dispersión

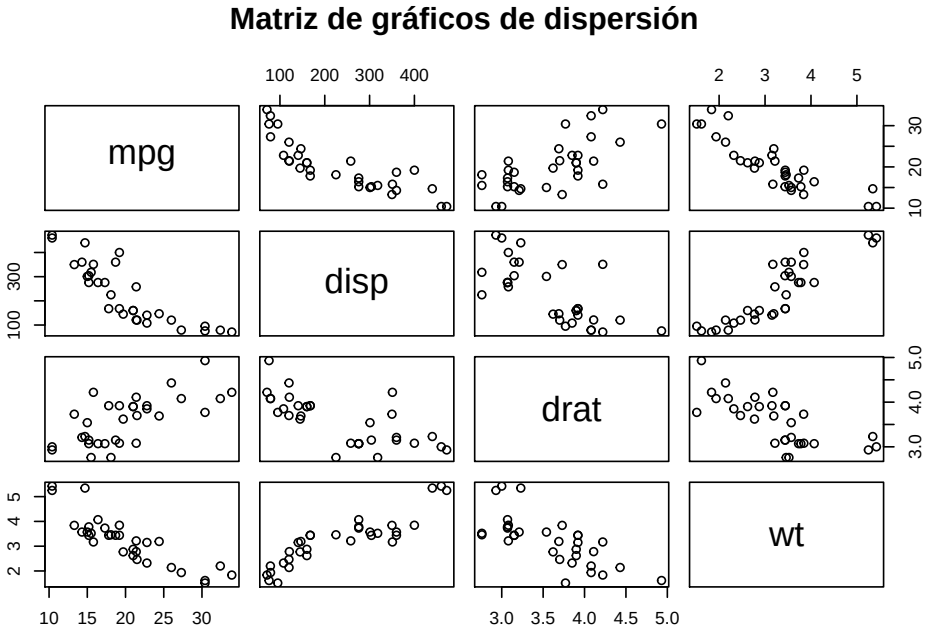
Definición. 1.28. Sea $(x_1, y_1), \ldots, (x_n, y_n)$ una muestra de tamaño n de datos pareados en que (x_i, y_i) fueron observados desde una misma unidad o individuo para las variables X e Y, respectivamente. Un **gráfico de dispersión** para esta muestra es la representación de estos puntos en el plano Cartesiano. En general, el objetivo es determinar si existe alguna relación funcional $y = f(x)$ evidenciada por los datos.

Presentamos ejemplos de posibles tendencia de asociaciones entre variables.



Interpretación: La figura (A) muestra un caso donde no existe una clara asociación entre las variables. La figura (B) muestra una fuerte asociación lineal (directamente proporcional). La figura (C) muestra una fuerta asociación linea negativa (inversamente proporcional). La figura (D) muestra un caso de asociación NO LINEAL del tipo inversa.

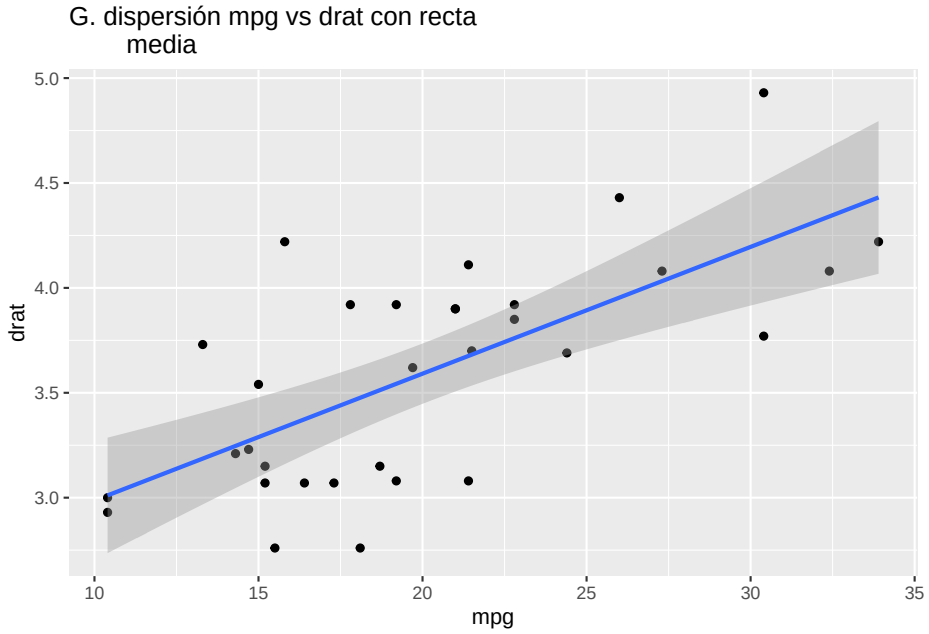
Presentamos una matriz de correlaciones para las variables mpg, disp, drat y wt la base de datos *mtcars*. Este tipo de gráfico es útil para identificar posibles relaciones lineales o no lineales entre variables de una base de datos.



Notamos una asociación lineal entre mpg (millas recorridas por galón lleno) y drat[2] (revoluciones del eje)la cual exploramos de forma especifica a seguir.

```
## `geom_smooth()` using formula 'y ~ x'
```

[2]La relación del eje es el número de revoluciones que el eje de salida o el eje de transmisión necesita hacer para hacer girar el eje una vuelta completa.



Unos de los objetivos consiste en determinar la fórmula de la recta que pasa por "en medio de los puntos".

1.8. Medidas de asociación entre dos variables

Definición. 1.29. Sea $\bar{x}_n$ y $\bar{y}_n$ la media de los datos observados de X e Y, respectivamente. La covarianza entre estas dos variables es un número que se denota por $\text{cov}(x, y)$ calculado por:

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y}) .$$

Definición. 1.30. El **coeficiente de correlación muestral** entre estas dos variables X e Y es un número que se denota por $r(x, y) = r$. Este coeficiente se define de la siguiente forma.

$$r(x, y) = r = \frac{\text{cov}(x, y)}{\sqrt{s_x^2 \times s_y^2}} \text{ donde } -1 < r < 1.$$

El coeficiente r mide la asociación lineal entre los datos observados de X e Y. Si $r = 0$, decimos que **no existe** asociación lineal entre los datos. Si $r < 0$, decimos que existe una **correlaciónlineal negativa** o inversamente proporcional entre los datos. Si, $r > 0$, decimos que existe una **correlación lineal positiva** entre los datos o directamente proporcional.

Ejemplo. 1.21. Para los datos de mpg y drat, la covariancia es 2,20 y el coeficiente $r = 0{,}68$ es decir existe una asociación lineal positiva (o directamente proporcional) en mpg y drat. Interpretación: si los valores de mpg crecen, también lo hace drat. Si los valores de mpg decrecen, también lo hacen drat.

1.9. Recta media estimada

Consideremos un par de variables, una de las cuales será denominada variable de entrada, y la otra, variable de respuesta. Supongamos que para un valor dado, x, de la variable de entrada, la variable de respuesta, y, se puede expresar en la forma

$$Y = \beta_0 + \beta_1 x + e.$$

Los coeficientes β_0 y β_1 son llamados los parámetros del modelo. Se asume que la variable e, denominada error aleatorio, es una variable aleatoria con media 0, es decir, asume valores muy cercano a 0.

Definición. 1.31. La relación entre la variable de ***respuesta (dependiente,output)***, Y, y la variable de ***entrada (independiente, input)***, x, especificadas ambas en la anterior ecuación, se denomina regresión lineal simple.

Para los pares de datos dados $(x_i, y_i), i = 1, ..., n$, los estimadores de (por) mínimos cuadrados son los valores de β_0 y β_1 que hacen lo más pequeño posible.

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

lo más pequeño posible.

Se puede demostrar que los estimadores de mínimos cuadrados de β_0 y β_1 dados por

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

Definición. 1.32. La recta

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

se denomina **recta de regresión media estimada**: $\hat{\beta}_1$ es la pendiente y $\hat{\beta}_0$ es el intercepto constante (o término independiente) de la recta.

Interpretación: Para valores cercanos a cero de x, observamos un valore medio cercano a $\widehat{\beta_0}$ de y. Por cada incremento (o decrecimo) de x, observamos un incremento (o decrecimo) medio de $\widehat{\beta_1}$ unidades en y.

Preveer una nueva observación de las variables respuesta y_{n+1} dado una variable de entrada x_{n+1} es posible directamente simplemente evaluando la ecuación de la recta, es decir:

$$\widehat{y}_{n+1} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{n+1}.$$

1.10. Diagnostico de la recta lineal

1.11. Análisis de asociación entre variables categóricas

Definición. 1.33. Una **tabla de contingencia:** contiene las frecuencias conjuntas observadas en la muestra de pares de atributos expresadas por las unidades muestrales.

Presentamos ejemplos de dos formas que podria tomar una tabla de contingencia de dimensión 2×3.

Cuadro 1.1: Tabla en frecuencia absoluta

s/r				Suma
	N_{11}	N_{12}	N_{13}	N_{1+}
	N_{21}	N_{22}	N_{23}	N_{2+}
Suma	N_{+1}	N_{+2}	N_{+3}	n

Cuadro 1.2: Tabla en frecuencia relativa

s/r				Suma
	p_{11}	p_{12}	p_{13}	p_{1+}
	p_{21}	p_{22}	p_{23}	p_{2+}
Suma	p_{+1}	p_{+2}	p_{+3}	1

Relación entre categorias de variables nominales: la probabilidad de que una categria A ocurra es condicional (depende) o no depende (independiente) de otra categoria B.

Definición. 1.34. Estadística **chi-cuadrado** es dada por para una tabla de congencia de 2×3 es dada por:

$$Q = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(N_{ij} - np_{i+}p_{+j})^2}{np_{i+}p_{+j}}$$

en que $Q > 0$.

Coeficiente de contingencia de Pearson es dado por:

$$C = \sqrt{\frac{Q}{Q+n}} \in \left[0, \sqrt{\frac{k-1}{k}}\right] \text{ en que } k = \text{mín}\{s, r\}.$$

Definición. 1.35. Coeficiente de contingencia de Pearson CORREGIDO es dado por:

$$C' = \sqrt{\frac{k}{k-1}} \times C \in [0, 1].$$

Un $C^{'}$ cercano a 0 indica características independientes. C' cerca de 1 señala una mayor medida de dependencia entre las características.
Ejemplo. 1.22. Consideremos la base de datos *mtcars* y las variables nominales vs (tipo de motor en forma de v o recto) y am (cambio automatico o manual). Para este caso obtenemos un $C' = 0{,}235$. Es decir, la probabilidad de que un auto escogido al azar sea automatico o mecánico no depende de la forma de su motor.

1.12. Ejercicios

1. texto.
2. texto.
3. Texto.

Capítulo 2

Calculo de probabilidades

2.1. Operaciones con eventos posibles de experimentos

Definición. 2.1. Experimento científico: es un experimento que se ejecuta controlando las circunstancias del entorno de manera de que este sea **replicable** y que las propiedades de los resultados posibles se mantengan constantes.
Definición. 2.2. El **espacio muestral** de un experimento es el conjunto (propuesto por el investigador) que incluye todos los posibles resultados de un experimento. Un caso que forma parte del espacio muestral es llamado de **evento posible** del experimento. Un **evento elemental** es un evento mínimo que puede de ocurrir.
Ejemplo. 2.1. Considere el experimento que consiste en lanzar un dado de seis caras al aire sobre una superficie plana y observar la cara superior para registrar el número que muestre. Incertidumbre: el dado podría mostrar un número del uno al seis. El espacio muestral son los números del 1 al 6. Ejemplo de eventos obtenemos un número par, obtenemos un número impar, obtenemos un 5 y obtenemos un múltiplo de 3. Un evento imposible seria obtenemos un número mayor a 7. Un evento elemental es obtenemos el número 5.

Una herramienta para representar eventos de un experimento es la teoría ingenua de conjuntos. Usaremos la letra S para representar el espacio muestral. Un evento elemental es simbolizado por $\omega \in S$. Un evento A del experimento será un subconjunto de S, es decir $A \subseteq S$. Los eventos se pueden describir por *extensión* o por *comprensión*.

Presentamos un ejemplo:
Ejemplo. 2.2. - Por compresión: $\omega \in S$ tal que $\omega = 1, ..., 6$. Un evento $\omega \in A$ siempre que $\omega = 2, 4, 6$ (números pares). Evento elemental $\omega \in B$ siempre que $\omega = 5$. Otra forma de escribir S por comprensión es

$$S = \{\omega : \omega = 1, \dots, 6\}.$$

- Por extensión: $S = \{1, 2, 3, 4, 5, 6\}$, $A = \{2, 4, 6\}$ y $B = \{5\}$.

Operaciones con eventos (conjuntos). Las operaciones de eventos de un experimento son transformaciones de uno o más eventos en otros eventos del experimento. Consideramos las siguientes:

- La union de A y B, simbolizado por $A \cup B$, transforma A y B en el evento $A \cup B$ que contiene todos los eventos elementales de A y B juntos. Es decir,

$$A \cup B = \{\omega \in S : \omega \in A \text{ o también } \omega \in B\}.$$

- La intersección de A y B, simbolizado por $A \cap B$, transforma A y B en el evento $A \cap B$ que contiene todos los eventos elementales en común de A y B.

$$A \cap B = \{\omega \in S : \omega \in A \text{ y a su vez } \omega \in B\}.$$

- El complemento de A, simbolizado por A^c, transforma A en el evento A^c que contiene todos los eventos elementales que no están en A, pero si están en S.

$$A^c = \{\omega \in S : \omega \notin A\}.$$

Definición. 2.3. Una lista $A_1, \dots, A_k$ de eventos de S , simbolizada por $\mathcal{A}$ es llamada de una **clase de eventos** de S. La clase $\mathcal{A}$ es una álgebra de sucesos si cumple las siguientes condiciones:

- $S \in \mathcal{A}$.
- Todas las uniones finitas de eventos de S estan también en S.
- El complemento de todos los eventos de S también están en S.

Definición. 2.4. Sigma álgebra de Borel (pendiente).

2.2. Medida de probabilidad y experimentos aleatorios

Definición. 2.5. Sea $\mathcal{A}$ una álgebra de eventos de S. Una función $\Pr : \mathcal{A} \to [0, 1]$ que le asigna a cada evento de $\mathcal{A}$ un número dentro del intervalo $[0, 1]$ es llamada de **medida de probabilidad finita** siempre que satisfaga las siguientes condiciones:

- 1 $\Pr(A) \geq 0$.

- 2 $\Pr(S) = 1$.
- 3 $A_1, \ldots, A_n \in \mathcal{A}$ disjuntos (2 a 2), entonces

$$\Pr\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} \Pr(A_k).$$

Una medida de probabilidad es una operación que mide la incertidumbre del evento A en un número en $[0, 1]$. Esta medida le da el máximo valor al espacio muestral. La suma de probabilidades de eventos disjuntos es la probabilidad de las uniones.

Propiedades de la medida de probabilidad. Una medida de probabilidad que satisface 1 a 3 satisface las siguientes propiedades:

- $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ - $\Pr(A \cap B)$.
- $\Pr(A^c) = 1 - \Pr(A)$.

Definición. 2.6. Sea A un evento de un experimento con espacio muestral S de con un número de eventos elementales finito. Se define la **probabilidad clásica** del evento A como el cociente

$$\Pr(A) = \frac{\text{Número de casos a favor de } A}{\text{Número de casos totales (a favor de } S)}.$$

Este cálculo de probabilidad también es conocida como la **regla de Laplace**.

Principio de razón insuficiente: sino tenemos información previa o no queremos asumir un sesgo personal sobre la incertidumbre de una propiedad muestral, entonces debemos asignar la *misma probabilidad* a cada evento puntual de Ω. La medida de probabilidad de un evento A es la razón entre el número de casos a favor de A y el número tal de casos posibles.

Definición. 2.7. Un **modelo de probabilidad** para un experimento es formado por el espacio muestral S, una álgebra $\mathcal{A}$ de eventos de S y una medida de probabilidad de Pr. Esto puede ser representado por la tripla $(S, \mathcal{A}, \Pr)$

Un **experimento aleatorio** es un *experimento* si y solmente si en cuya construcción tiene asociado una medida de probabilidad para sus eventos.

2.2.1. Interpretación frecuentista de una medida de probabilidad

Supongamos que repetimos un mismo experimento secuencialmente. Los primeros n resultados posibles son representados por $\omega_1, \omega_2, \ldots, \omega_n$ y que $\mathbb{I}_A(\omega_i) = 1$ cuando $\omega_i \in A$. Luego, según la interpretación frecuentista, para un máximo de diferencia absoluta ϵ siempre va a existir un momento de repetición n_0 del experimento tal que:

$$\left|1/n \sum_{i=1}^{n} \mathbb{I}_A(\omega_i) - \Pr(A)\right| < \epsilon, \forall n \geq n_0$$

.

Es decir, a partir de una repetición finita del experimento, la diferencia entre $\Pr(A)$ y $1/n \sum_{i=1}^{n} \mathbb{I}_A(\omega_i)$ es tan pequeña como querramos. Es decir $\Pr(A)$ es el **límite de las frecuencias relativas** de observar eventos elementales de A en la replicación secuencial de un mismo experimento.

2.3. Propiedades básicas de la medida de probabilidad

Definición. 2.8. Probabilidad conjunta, marginal y condicional.
Definición. 2.9. Eventos estocasticamente independientes.
Definición. 2.10. Probabilidad total.
Definición. 2.11. Teorema de Bayes.

2.4. Resultados básicos de combinatoria

Definición. 2.12. Principio básico del conteo.
Definición. 2.13. Principio básico del conteo generalizado.

Capítulo 3

Variables aleatorias y sus distribuciones

3.1. Elementos básicos

Definición. 3.1. Sea S el espacio muestral de un experimento. Una función X con valores reales que le otorga a cada elemento de S un valor numérico es **variable aleatoria**. El conjunto de todos los valores posibles que puede asumir X es llamado de espacio muestral de X y es denotado por $\mathcal{X}$. Usaremos la notación $X : S \to \mathcal{X}$.
Ejemplo. 3.1. - X_1: número de caras que resultan al lanzar 10 veces un monedas al aire.

- X_2: el número de seis al lanzar un dado 15 veces.
- X_3: el número de accidentes por semana que ocurren una empresa.
- X_4: el consumo de leña por casas en una ciudad.

Definición. 3.2. La función de densidad f para una variable aleatoria continua X cuyo espacio muestral es $\mathcal{X}$ es una función que modela la variabilidad de los posibles valores de X y nos ayuda a calcular la probabilidad de que ella asuma determinados valores. Esta función satisface las siguientes propiedades.

- $f(x) \geq 0$ para todo x.
- $\int_{\mathcal{X}} f(x)dx = 1$.

La probabilidad de que $X \in (a, b)$ es el área bajo la curva de f limitado al intervalo (a, b).
Definición. 3.3. Por otro lado, la función de probabilidad p para una variable aleatoria discreta X cuyo espacio muestral es $\mathcal{X}$ es una función que modela la variabilidad de los posibles valores de X y nos ayuda a calcular la probabilidad

de que ella asuma determinados valores. Esta función satisface las siguientes propiedades.

- $p(x) \geq 0$ para todo x.
- $\sum_{x \in \mathcal{X}} p(x) = 1$.

La probabilidad de que $X \in \{a_1, \ldots, a_k\}$ es la suma $\sum_{x \in \{a_1, \ldots, a_k\}} p(x)$.

Ejemplo. 3.2. Consideremos la función de densidad f dada por

$$f(x) = \frac{1}{\pi(1 + x^2)}, \quad x \in R,$$

en que π es la constante pi.

La función de densidad p dada por

$$p(x) = (1 - 1/6)^{x-1} 1/6, \quad x = 1, 2, \ldots .$$

Cuando f o p esta totalmente especificada, ambas son llamados de **modelo de probabilidad** para X. Cuando f o p depende una constante θ asumida como desconocida, entonces ellas son llamadas de **modelo estadístico** para X. El objetivo de los métodos de inferencia es determinar valores plausibles de θ desde una muestra observada.

En el último caso $f(x)$ es representada por $f(x; \theta)$ indicando que f es una función de x para un valor de θ que debe ser especificado. En el caso discreto, $p(x)$ es representado por $p(x; \theta)$.

Definición. 3.4. La función de distribución acumulada de una variable aleatoria X es la funcion $F(x) = \text{Prob}(X \leq x)$, en el caso continuo, es definida por

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

y en el caso discreto

$$F(x) = \sum_{t=-\infty}^{x} p(t; \theta).$$

Propiedades de $F(x)$:

- $\Pr(a < X < b) = F(b) - F(a)$,
- $\Pr(X > a) = 1 - F(a)$,
- $\lim_{x \to +\infty} F(x) = 1$,
- $\lim_{x \to -\infty} F(x) = 0$.

Ejemplo. 3.3. Para la función f y P antes presentadas, las respectivas funciones de probabilidad acumulada son,

$$F(x) = \int_{-\infty}^{x} \frac{1}{\pi(1 + t^2)} dt$$

y

$$F(x) = \sum_{t=0}^{x} (1-\pi)^{t-1}\pi$$

respectivamente.

Definición. 3.5. Una variable aleatoria X es **discreta** si el conjunto de sus posibles resultados $\mathcal{X}$ es finito o puede ser enumerados en la forma $\{a_1, a_2, \dots, a_n, \dots\}$. Una variable aleatoria X es **continua** si existe una función de densidad para ella.

3.2. Distribuciones de probabilidad de variables aleatorias discreta

Definición. 3.6. Una variable aleatoria **Bernoulli** es una indicadora de la ocurrencia de evento $A \subseteq S$. Es decir, $X = 1$ cuando $\omega \in A$ (ocurre A) con probabilidad p y $X = 0$ cuando $\omega \notin A$ (no ocurre A) con probabilidad p. Esto se representa por $X \sim \text{Bernoulli}(p)$.

La función de probabilidad de X es:

$$p(x;\theta) = \theta^x(1-\theta)^{x-1} \text{ para } x = 0, 1 \text{ y } \theta \in (0, 1).$$

Ejemplo. 3.4. Consideremos el experimento del lanzamiento de una moneda al aire. Sea $X = 1$ cuando obtenemos una cara con probabilidad $p = 1/2$ y $X = 0$ cuando obtenemos un sello con probabilidad $1 - 1/2$.

Consideremos el experimento del lanzamiento de un dado al aire. Sea $X = 1$ cuando obtenemos una seis con probabilidad $p = 1/6$ y $X = 0$ cuando no con probabilidad $1 - 1/6$.

Definición. 3.7. Considere una repetición de mismo experimento n y sea $X_1, X_2, \dots, X_n$ una secuencia independiente de variables Bernoulli, en que $X_i = 1$ cuando $\omega \in A$ con probabilidad θ para $i = 1, \dots, n$. Una variable **Binomial** Z es el número de caso a favor de A al repetir el experimento n veces, es decir:

$$Z = \sum_{i=1}^{n} X_i.$$

Esto se representa por $Z \sim \text{Binomial}(n, \theta)$. Su función de probabilidades es:

$$p(z;n,\theta) = \frac{n!}{z!(n-z)!}\theta^z(1-\theta)^{n-z}.$$

Ejemplo. 3.5. En primer caso, considere el experimento de lanzar una moneda $n = 15$ veces y observar los lanzamientos de cara. Sea $X_1 = x_1, \dots, X_n = x_n$

las indicadoras de observar este evento en cada repetición. Luego $Z = \sum_{i=1}^{15} x_i$ es el número de caras observar al lanzar la moneda 15 veces. En este caso $Z \sim \text{Binomial}(15, 1/2)$

En segundo caso, considere el experimento de lanzar un dado $n = 20$ veces y observar los números un seis. Sea $X_1 = x_1, ..., X_n = x_n$ las indicadoras de observar este evento en cada repetición. Luego $Z = \sum_{i=1}^{20} x_i$ es el número de seis que observamos al lanzar el dado 20 veces. En este caso $Z \sim \text{Binomial}(20, 1/6)$.
Ejemplo. 3.6. Distribución multinomial (pendiente).
Ejemplo. 3.7. Cuando X es una variable que mide el número eventos que ocurren en un determinado intervalo de tiempo $[c, c+t]$ de longitud t. Sobre ciertas suposiciones, una de las opciones es el modelo de **Poisson** con un promedio de ocurrencias por intervalo igual a λ.

Una variable Poisson tiene función de probabilidad dada por:

$$p(x; \lambda) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} \text{ para } x \in \mathbb{N} \text{ y } \lambda > 0.$$

y función de distribución acumulada dada por:

$$F(x; \lambda) = \sum_{k=0}^{x} \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}.$$

Ejemplo. 3.8. - X: número de accidentes por semana que ocurren en una industria papelera. Para $\lambda = 3$, si observamos repetidas observaciones de X, luego observaríamos en promedio 3 accidentes por semana.

- X: número de accidentes en una carretera por día. Para $\lambda = 4$, si observamos repetidas observaciones de X, luego observaríamos 4 accidentes en la carretera por día.
- X: número de partículas por minuto que emite una sustancia radiactiva. Para $\lambda = 17$, si observamos repetidas observaciones de X, luego observaríamos en promedio 17 partículas por minuto.

Ejemplo. 3.9. Distribución geométrica (pendiente).
Ejemplo. 3.10. Distribución binomial negativa (pendiente).
Ejemplo. 3.11. Distribución hipergeométrica (pendiente).

3.3. Distribuciones de probabilidad de variables aleatorias continuas.

En esta sección presentamos algunos modelos para variables aleatorias continuas.

Ejemplo. 3.12. Una variable $X > 0$ sigue un modelo **Exponencial** con taza de concentración $1/\lambda$ cuando su función de densidad es dada por:

$$f(x;\lambda) = \lambda \mathrm{e}^{-x\lambda}, \text{ para } x > 0, \ \lambda > 0.$$

La taza de concentración $1/\lambda$ indica el promedio de tiempo medio de X. El parámetro de decaimiento λ nos indica la forma de la densidad en el sentido de que indica que tan rápido decae (va hacia cero) la cola izquierda de $f(x;\lambda)$.

Su función de probabilidad acumulada es dada por:

$$F(x;\lambda) = \int_{-\infty}^{x} \lambda \mathrm{e}^{-t\lambda} dt = 1 - \mathrm{e}^{-x\lambda}.$$

Ejemplo. 3.13. Una variable aleatoria se suele considerar Exponencial cuando esta es contínua y mide el tiempo de "vida" o de "duración" de vida útil de objetos. Por ejemplo:

- X: tiempo de funcionamiento de una máquina procesadora de leche en miles de horas. Para $\lambda = 1/10$, si observásemos repetidos tiempos de funcionamiento de muchas máquinas del mismo tipo, observaríamos un promedio de 10 mil horas.
- X: tiempo de vida de circuitos eléctricos en miles de horas. Para $\lambda = 1/5$, si observásemos repetidos tiempos de vida de circuitos, observaríamos un promedio de 5 mil horas.
- X: tiempos de llegada en minutos de clientes a una fila de banco. Para $\lambda = 1/4$, si observásemos una gran cantidad de observaciones de esta, observaríamos un promedio de 4 minutos.

Definición. 3.8. Una variable normal X se usa para modelar variables límites o aproximadamente simetricas. Su función de densidad es dada por

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],$$

donde $\mu \in \mathbb{R}$ es el parámetro posición, $\sigma > 0$ es la varianza y $x \in \mathbb{R}$.

Ejemplo. 3.14. - X: puntajes de pruebas estandarizadas.

- X: promedios de variables con más de 30 datos observados.

Ejemplo. 3.15. Distribución Beta.

Ejemplo. 3.16. Distribución Gumbel.

Ejemplo. 3.17. Distribución Weibull.

Ejemplo. 3.18. Distribución uniforme.

Capítulo 4

Distribuciones bidimensionales

4.1. Uno

4.2. Dos

Capítulo 5

Valor esperado

Definición. 5.1. Media y variancia poblacional.
Ejemplo. 5.1. Ejemplos media y variancia poblacional.
Definición. 5.2. Función generadora de probabilidad y momentos.

5.1. Valor esperado de modelos discretos

5.2. Valor esperado de modelos discretos

Capítulo 6

Resultados asintóticos.

Capítulo 7

Funciones de muestras aleatorias

Definición. 7.1. Muestras aleatorias.
Ejemplo. 7.1. Muestras aleatorias ejemplo.
Definición. 7.2. Distribuciones de muestreo de estadísticas.
Ejemplo. 7.2. La distribución de muestreo de χ^2_n.
Ejemplo. 7.3. La distribución de muestreo de S2.
Ejemplo. 7.4. La distribución t de Student.
Ejemplo. 7.5. La distribución F.

Capítulo 8

Estimadores Puntuales

8.1. Propuesta de informe

Algo para definir

Capítulo 9

Estimadres intervalares

Capítulo 10

Prueba/contraste de hipótesis

Capítulo 11

Regresión lineal

Capítulo 12

ANOVA de un factor.

Apendices: formulas y calculos algebraicos

Referencias