BIOESTADÍSTICA

Notas de Curso

Universidad de la República

2021

Cómo utilizar estas notas.

Estas notas fueron hechas para el curso de Estadística de la Licenciatura en Recursos Naturales, dictado en el año 2012. Han sido reescritas para el curso de 2013 y ampliadas para el curso de Bioestadística de la Facultad de Ciencias, en el 2018. Para el curso de 2019 se agregaron las secciones sobre el modelo lineal simple y ANOVA, y se agregó además un apéndice que contiene algo mas del modelo lineal simple, el modelo lineal general, y algunos temas que no se dictaron en dicho curso que se dictaban antes. Han sido de gran utilidad para esta nueva edición los comentarios y recomendaciones de Pablo Inchausti y de Flavio Pazos. Se agregaron además subsecciones con comandos de R. No pretenden ser un sustituto de las clases, sino más bien una guía para que el estudiante sepa que temas se dieron en las clases, y una orientación del estilo y metodología usados en las mismas. Las definiciones se introducen de manera intuitiva, omitiendo las construcciones teóricas rigurosas que la teoría de la probabilidad requiere. Al final de cada sección hemos agregado algunos comandos básicos de R (implementados en la version 3.6.1 y con Rstudio) relacionados a lo visto en esa sección. La bibliografía y el índice figuran al final. Las erratas que hubieren, se agradece comunicarlas a acholaquidis@hotmail.com. La versión 2019 ha sido hecha usando el paquete knitr y no hubiera sido posible sin la invalorable colaboración de Gabriel Illanes, quien además mejoró sutancialmente el capítulo de modelos lineales.

Contents

Part I

Probabilidad

Chapter 1

Conteo y nociones básicas de probabilidad

1.1 Conteo

En la presente sección mencionaremos de forma breve algunos conceptos que serán muy importantes para el cálculo de probabilidades. El entendimiento de los mismos no requieren de conocimientos previos de ningún tipo. Para un estudio más completo una referencia clásica, con ejemplos y ejercicios resueltos es [G].

1.1.1 Regla del Producto

La regla del producto es un principio básico de conteo que nos dice que si tenemos n formas de realizar una tarea A y para cada una de ellas podemos realizar otra tarea B de m formas, entonces la tarea completa (realizar A y B) se puede realizar de $m \times n$ formas. Esto se generaliza de manera trivial a cualquier cantidad de tareas. Por ejemplo, si tenés 3 tipos de hamburguesas para elegir en tu restaurante de comida favorita, y para cada una de ellas tenés 5 bebidas distintas que podes elegir y 2 guarniciones distintas, en total hay $3 \times 5 \times 2 = 30$ posibles formas de armar tu menú. Veamos otro ejemplo que nos será de utilidad más adelante, supongamos que tenemos 2 dados, uno con números en rojo y otro en negro, ¿cuántos posibles resultados tenemos?. Aquí una *tarea* es por ejemplo que salga $(1,2)$, otra es $(6,3)$, y otra distinta es $(3,6)$. Para cada uno de los 6 posibles resultados de tirar el dado rojo tenemos 6 posibles resultados de tirar el dado negro, por lo tanto en total tenemos $6 \times 6 = 36$ posibles resultados. Veamos otro ejemplo, supongamos que queremos contar cuantos números pares de 3 cifras se pueden armar con los números del 0 al 6. Llamemos $C_1C_2C_3$ a un caso genérico, es claro que $C_1 \neq 0$, por lo tanto para la primera tarea (elegir C_1) tenemos 6 posibilidades (esto es, $C_1 = 1, C_1 = 2, C_1 = 3, C_1 = 4, C_1 = 5$ o $C_1 = 6$), para la segunda tarea, es decir elegir C_2 podemos usar cualquiera de los 7 números entre 0 y 6. Finalmente, para la tercer tarea tenemos una restricción ya que queremos que el número sea par, por lo tanto C_3 no puede ser ni 1, ni 3 ni 5. Esto nos da $6 \times 7 \times 4$ posibilidades.

1.1.2 Permutaciones

Vamos a empezar con un ejemplo, consideremos las tres primeras letras, A, B y C del abecedario, queremos contar de cuantas maneras distintas se pueden permutar u ordenar. Es claro que las posibilidades son ABC, ACB, BCA, BAC, CBA y CAB. Es decir tenemos 6 formas distintas. Si escribimos todas las formas de ordenar las cuatro primeras letras ABCD, vemos que en este caso nos quedan 24 posibilidades, no obstante esta forma de *contar* se vuelve inapropiada cuando por ejemplo tomamos las 10 primeras letras. Veamos a continuación como contar las maneras de ordenar una cantidad cualquiera n de objetos.

Definición 1.1. Supongamos que tenemos n objetos distintos. Queremos contar la cantidad de formas distintas de ordenarlos en n lugares. En el primer lugar podemos poner cualquiera de los n elementos, en el segundo lugar, dado que ya tomamos un objeto para el primer lugar, nos quedan $n-1$ objetos. Por lo tanto para el primer y segundo lugar tenemos $n \times (n-1)$ posibilidades. Observar que aquí se usa la regla del producto donde la primera tarea A que realizamos es elegir un elemento entre los n para el primer lugar, y la segunda tarea B es elegir un segundo elemento de los $n-1$ restantes para el segundo lugar. Razonando de esta manera, llegamos a que tenemos

$$n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1,$$

posibles ordenaciones, o *permutaciones* de los n objetos.
En general se emplea la notación $n! := n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$, y se asume, por definición que $0! = 1$.

Observemos que, dado que, por definición $(n-1)! = (n-1) \times (n-2) \times \cdots \times 2 \times 1$ tenemos que $n! = n \times (n-1)!$.

Ejemplo 1.2. Supongamos que queremos contar de cuantas maneras se pueden ordenar las letras de la palabra $HOLA$. Dado que no se repiten letras (es decir pueden ser consideradas como *objetos* distintos), es un problema de conteo de permutaciones de 4 elementos distintos. En el primer lugar podemos poner la letra H, la O la L o la A, es decir tenemos 4 posibilidades. En el segundo lugar, dado que elegimos 1 de las letras, tenemos 3 posibilidades, en el tercer lugar tenemos 2 y en el último nos queda 1 sola letra sin colocar. En la Figura (1.2) se muestra el árbol de las posibilidades, si empezamos con la letra H.

1.1.3 Arreglos

Siguiendo nuestro ejemplo de las primeras tres letras del alfabeto, supongamos que queremos contar de cuantas maneras distintas se pueden extraer 2 letras de mi conjunto de 3 letras (formado por las letras A,B y C). Es decir una extracción posible es AB y otra distinta es BA o AC. En este caso las palabras que vamos a formar tienen dos letras, tomadas de las tres que tenemos, es decir las posibilidades son: AB, BA, AC, CA, BC y CB. Veamos que pasa ahora con las cuatro primeras letras del alfabeto. Aquí las posibilidades son AB, AC, AD, BA, CA, DA, BC, BD, CD, DC, CB, DB. En este caso nos da 12, es decir la mitad que el total de permutaciones de 4 elementos distintos. Veamos como contarlas cuando son más de 4 elementos.

Definición 1.3. Supongamos que tenemos n objetos distintos. Queremos contar la cantidad de formas de extraer (sin reposición) k de esos objetos (con $1 \leq k \leq n$). Es importante tener en cuenta que importa el orden

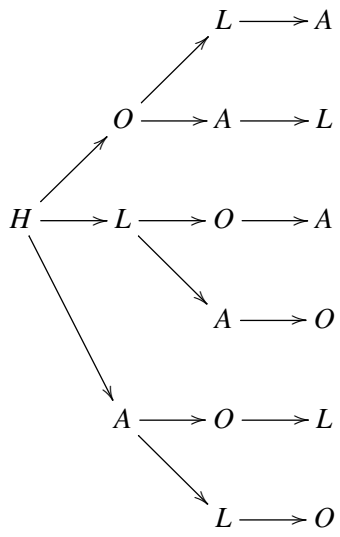


Table 1.1: Si comenzamos con H tenemos 6 posibilidades, dado que podemos empezar con cualquiera de las 4 letras (y que comenzar con letras distintas da ordenaciones distintas) tenemos $6 \times 4 = 24 = 4!$

en que se extraen los k objetos, por ejemplo, al extraer 3 letras de la palabra $HOLA$, extraer OLA es distinto que LOA. En la primera extracción tenemos 4 posibles resultados, 3 en la segunda extracción y finalmente 2 en la tercera. Por la regla del producto esto nos da $4 \times 3 \times 2$ posibilidades. En general, si tenemos n objetos distintos (razonando de la misma manera que antes) y queremos extraer k, tenemos n posibilidades para la primera extracción, $n-1$ para la segunda, $n-2$ para la tercera extracción y finalmente en el lugar k tenemos $n-(k-1) = n-k+1$ posibilidades ya que en este caso hemos extraído $k-1$ objetos de los n . Es decir en total tenemos

$$n \times (n-1) \times \cdots \times (n-k+1).$$

En general se denota $A_k^n := n \times (n-1) \times \cdots \times (n-k+1)$.

Ejemplo 1.4. ¿Cuántos números de tres cifras se pueden formar con los dígitos $0, 1, 2, 3$ y 4 si no se permite la repetición de dígitos? Observemos que en las unidades podemos poner cualquiera de los 5 dígitos, esto nos deja 4 dígitos para elegir para las decenas, y finalmente, en las centenas podemos poner 3 dígitos. En total son $A_3^5 = 5 \times 4 \times 3$. Observemos ademas que podemos escribir

$$A_3^5 = 5 \times 4 \times 3 = 5 \times 4 \times 3 \times \frac{2 \times 1}{2 \times 1} = \frac{5!}{2!} = 60.$$

Observación 1.5. *Verificar que*

$$A_k^n = \frac{n!}{(n-k)!} \tag{1.1}$$

en particular $A_n^n = n!$ *y* $A_1^n = n$ *ya que hemos definido* $0! = 1$.

1.1.4 Combinaciones

Continuando con nuestro ejemplo, supongamos que realizamos 2 extracciones (sin reposición) del conjunto formado por las primeras 3 letras del abecedario y queremos contar la cantidad de formas distintas en que se pueden realizar dichas extracciones pero ahora no nos importa el orden de los resultados. Es decir, no distinguimos AB de BA, ni AC de CA, ni CB de BC. Esto es lo mismo que agrupar de a pares de letras, o, lo que es lo mismo, construir conjuntos de dos elementos (ya que en los conjuntos no importa el orden). Tenemos entonces tres conjuntos: el que contiene las letras A y B: $\{A,B\}$, el que contiene las letras A y C: $\{A,C\}$ (observemos que como conjunto es distinto al anterior). Y el que contiene las letras B y C: $\{B,C\}$ que es distinto a los anteriores. Son entonces tres posibilidades. Nos dio exactamente la mitad que en el caso en que tenemos en cuenta el orden, ¿es esto intuitivo?. Si repetimos el razonamiento con 4 letras, y 2 extracciones nos da los conjuntos $\{A,B\}$, $\{A,C\}$, $\{A,D\}$, $\{C,B\}$, $\{C,D\}$ y $\{B,D\}$, es decir, nos da 6. Nuevamente nos dio la mitad de posibilidades que si tenemos en cuenta el orden. No obstante, si tomamos 4 letras distintas y realizamos 3 extracciones (siempre sin reposición) tenemos 4 conjuntos posibles (escribirlos). Es la misma cantidad que si hubiésemos tenido en cuenta el orden. En general no tener en cuenta el orden da lugar a *menos, o la misma cantidad,* de posibilidades que al tenerlo en cuenta ya que al no tener en cuenta el orden estamos agrupando ordenaciones.

Definición 1.6. Supongamos que tenemos n objetos distintos y que queremos contar la cantidad de subconjuntos de k elementos (con $1 \leq k \leq n$) formado con k de los n elementos. Al hablar de subconjuntos, no importa el orden. Por ejemplo: tenemos $n = 5$ objetos diferentes A,B,C, D y E y queremos contar cuántos subconjuntos de $k = 3$ elementos se pueden formar. En el subconjunto formado por los elementos A, B y C tenemos $3! = 6$ ordenaciones posibles de dichos elementos que producen el mismo subconjunto, el $\{A,B,C\}$. Por lo tanto basta contar las ordenaciones de 3 elementos (es decir los arreglos) y luego agrupar las que forman el mismo subconjunto. O, lo que es lo mismo, si tenemos el número de subconjuntos, multiplicamos por 3! y nos da el número de ordenaciones o arreglos de 3 elementos. Es decir, en base a esto último, el número que queremos (que denotaremos $\binom{n}{k}$) tiene la propiedad de que $\binom{n}{k} \times k! = A_k^n$. Si usamos la observación anterior y (1.1)

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}. \tag{1.2}$$

Ejemplo 1.7.

- Se juega a un juego del tipo 5 de Oro: hay que acertar 5 números, elegidos dentro de 44 posibilidades. ¿Cuántas jugadas posibles hay?. Lo primero y más importante que hay que observar es que el orden no importa, importa a que números jugamos, y no en que orden. Por lo tanto estamos ante un problema de combinaciones. Tenemos $n = 44$ *objetos* distintos, y queremos contar la cantidad de subconjuntos de $k = 5$ elementos, es decir la respuesta es $\binom{44}{5}$.
- De entre 8 personas debemos formar un comité de cinco miembros. ¿Cuántas posibilidades diferentes existen para formar el comité?. Nuevamente, como hablamos de un comité de personas, no importa el orden en que las elegimos. La respuesta es entonces $\binom{8}{5}$.

Ejercicio 1.8. Veamos un ejemplo que nos será de utilidad más adelante

- Supongamos que tenemos una urna con 6 bolas de color rojo (que denotaremos R_1, R_2, R_3, R_4, R_5 y R_6) y 3 de color negro (que denotaremos N_1, N_2 y N_3). Vamos a extraer 5 bolas. Contaremos la cantidad de formas de extraerlas de modo tal que 3 sean de color rojo, y 2 de color negro. Es decir una posible extracción es R_1, R_2, R_5, N_1, N_3. Esta extracción es la misma que R_2, N_1, R_1, N_2, R_5 ya que solamente nos interesa la cantidad de bolas negras y la cantidad de bolas rojas que hay en la extracción. Otra *distinta* es R_3, R_2, R_5, N_1, N_3, en este caso si bien la cantidad de bolas negras y rojas que salieron es la misma, salió la bola roja R_3 y no la R_1 por lo tanto son objetos distintos que salieron y cuentan como extracciones distintas. Es decir, como no nos interesa el orden en que salen (siempre que los objetos sean los mismos), la extracción R_1, R_2, R_5, N_1, N_3 se puede sacar de 5! formas distintas, (y lo mismo para cualquier extracción que se haga). Como no nos importa el orden, vamos a usar combinaciones. Observemos que tenemos $\binom{6}{3}$ formas *distintas* de extraer 3 bolas rojas, de las 6 bolas rojas que hay en total. Por otro lado, *para cada* extracción de 3 bolas rojas, podemos completar las 5 que tenemos que sacar en total eligiendo 2 de las 3 bolas negras, es decir tenemos $3 = \binom{3}{2}$ formas de extraer dichas bolas. En total tenemos entonces
$$\binom{6}{3} \times \binom{3}{2}.$$
Aquí hemos aplicado la regla del producto, la tarea A en este caso es elegir las bolas rojas a extraer (que como vimos eso se puede hacer de $\binom{6}{3}$ formas distintas) y la tarea B es elegir las bolas negras, que se puede hacer de $\binom{3}{2}$ formas distintas. Ambas tareas se pueden hacer entonces de $\binom{6}{3} \times \binom{3}{2}$ formas distintas. Observar que es indistinto si realizamos primero la B y luego la A, da el mismo resultado.

- Si lo que queremos es contar de cuantas maneras ordenar 5 bolas donde 3 son de color rojo (y asumimos que son indistinguibles entre si) y dos son de color negro (y también las asumimos indistinguibles), esto da $\binom{5}{3}$ ya que únicamente tenemos que elegir donde colocaremos las bolas rojas. Usamos combinaciones porque por ejemplo R_1, N_2, R_2, R_3, N_1 es igual a R_3, N_1, R_2, R_1, N_2 (y es distinto que por ejemplo $R_3 R_1 R_2 N_1 N_2$) ya que las bolas son indistinguibles entre si. Es importante entender que este ejemplo es distinto al anterior, aquí hemos asumido que las bolas rojas son indistinguibles entre si (y lo mismo las negras).

- Supongamos que se tiene una urna con 15 bolas de color rojo y 7 de color negro.¿De cuántas formas se pueden extraer 9 de modo tal que 6 sean de color rojo y 3 de color negro?

Observación 1.9. *Verificar que*

- $\binom{n}{n} = 1 \quad \binom{n}{1} = n$. *Nótese que* $A_1^n = \binom{n}{1}$. *Sin usar las fórmulas respectivas, explicar por qué es razonable ese resultado.*

- $\binom{n}{k} = \binom{n}{n-k}$. *Esta fórmula, que se demuestra fácilmente a partir de* (1.2) *nos dice que, elegir subconjuntos de k elementos, es lo mismo que elegir* $n-k$ *elementos que no vamos a incluir en el subconjunto.*

En R

Si queremos calcular el número de permutaciones de $n = 10$ objetos distintos tenemos la función

```
n=10
factorial(n)

## [1] 3628800
```

Si queremos que nos de cuales son esas permutaciones (suponiendo que los n objetos son los números $1,\ldots,n$ tenemos el paquete combinat y la función permn(n), por ejemplo para $n = 2$,

```
library(combinat)

##
## Attaching package:  'combinat'
## The following object is masked from 'package:utils':
##
##     combn

permn(2)

## [[1]]
## [1] 1 2
##
## [[2]]
## [1] 2 1
```

1.2 Probabilidad: casos favorables sobre casos posibles

Aquí mencionaremos de forma breve, y sin el rigor que la teoría de la probabilidad requiere, los conceptos de probabilidad, independencia, probabilidad condicional, y algunas de las propiedades de la misma. Un libro recomendable para profundizar en estos temas es [FPP] o [S].
El objeto de estudio de la probabilidad son los experimentos cuyos resultados no se pueden *determinar* de antemano. Es decir la repetición del experimento no nos conduce necesariamente al mismo resultado (en este sentido es que decimos que el resultado es aleatorio o depende del azar). Por ejemplo arrojar un dado, una moneda. etc. El ejemplo clásico donde esto no sucede, son los experimentos de la química donde siempre que se combinan, en iguales condiciones, determinadas sustancias, se obtiene el mismo resultado. Por otro lado, en probabilidad supondremos que el conjunto de resultados posibles es conocido de antemano. En el caso del dado, sabemos que solo puede salir 1,2,3,4,5 o 6 en la cara superior. El conjunto de resultados posibles se llama espacio muestral, y suele denotarse con la letra griega Ω, los subconjuntos de Ω se llaman sucesos y son a los que le queremos asignar una probabilidad (es decir un número en $[0,1]$) Finalmente, si

bien no sabemos con anterioridad el resultado del experimento, supondremos que no pueden haber múltiples resultados al mismo tiempo. Es decir, por ejemplo, al arrojar un dado no puede salir 3 y 5 en la cara superior al mismo tiempo.

De manera intuitiva y a partir de la experiencia cotidiana (que motivó y dio origen a la formalización teórica posterior) la probabilidad de un suceso es la frecuencia con que este se da. Así por ejemplo, la probabilidad de que salga cara al arrojar una moneda es $1/2$ porque la mitad (aproximadamente) de las veces que se arroja la moneda, sale cara, es decir la frecuencia con que sale cara es $1/2$. De la misma manera, la probabilidad de que al tirar un dado salga 2, es $1/6$ ya que si arrojamos *muchas veces* el dado, 1/6 de las mismas saldrá 2 (y es igual a la probabilidad de que salga cualquiera de los 6 números). Observemos que, lo que estamos haciendo es asignar *el mismo* número entre 0 y 1, que llamamos probabilidad, a cada uno de los posibles resultados, y dado que son 6, solamente podemos asignar $1/6$.

Es importante aclarar que la probabilidad de un suceso *no es un porcentaje*, sino un número mayor o igual que cero, y menor o igual que uno, aunque son conceptos relacionados, no son lo mismo. Podemos decir que el 50% de las veces que tiramos una moneda sale cara, o que al tirar un dado aproximadamente el 16,6% de las veces sale 1 (o cualquiera de los 6 posibles números), pero la probabilidad de que salga cara *no* es 50%, ni de que salga 1 al tirar el dado es 16.6% (observar que 16.6% es una *aproximación* al porcentaje, que sería 100/6).

En el caso en que Ω tiene k elementos, con $1 \leq k < \infty$ de elementos y le asignamos a cada uno de ellos la misma probabilidad (es decir $1/k$), estamos ante el modelo de casos favorables *casos favorables sobre casos posibles*. Esto nos da además una forma de calcularle la probabilidad a cualquier suceso, ya que si $A = \{\omega_1, \ldots, \omega_l\}$ con $0 \leq l \leq k$ entonces $P(A) = l/k$.

Nuestra intuición nos dice que en general no es cierto que los sucesos tengan la misma probabilidad siempre, pensemos por ejemplo que queremos asignar una probabilidad al color en que está la luz del semáforo cuando llegamos a una esquina, intuitivamente es *menos probable* que la luz este en color amarillo, ya que el tiempo de duración del color amarillo es menor. Si pensamos que el tiempo en que está la luz en amarillo es $1/4$ del tiempo en que está en rojo, y que el tiempo en que está en rojo y en verde son iguales, tenemos que, al asignar probabilidades la probabilidad de que sea amarillo es $1/9$ (¿por qué?).

Como mencionamos antes, los sucesos a los que vamos a asignar probabilidades (es decir números entre 0 y 1) serán subconjuntos de un cierto conjunto *conocido* Ω. Por ejemplo, en el caso de los resultados de tirar una moneda $\Omega = \{0,1\}$ donde 0 representa cara, y 1 número. En el caso de los resultados de tirar un dado, $\Omega = \{1,2,3,4,5,6\}$ y un suceso es, por ejemplo $A = \{1,2\}$. En este sentido la probabilidad de A, que denotaremos $P(A)$ es la probabilidad de que al tirar un dado salga 1 o 2, que, intuitivamente es $2/6$. Si tiramos dos dados tenemos que Ω es el conjunto de las 36 parejas (i,j) con $i,j = 1,\ldots,6$ (ver tabla (1.2) más adelante) y A puede ser por ejemplo el suceso $\{$la suma de los resultados es 7$\}$.

En la tabla (1.2) se muestran los posibles resultados de tirar dos dados, vamos a suponer que podemos distinguirlos entre sí, por ejemplo uno tiene los números en rojo y el otro en negro. Para simplificar la explicación, cuando decimos, por ejemplo, que el dado rojo es mayor que el dado negro, nos referimos a que el resultado que sale en la cara superior del dado cuyos números son rojos es mayor que el resultado que sale en la cara superior del dado cuyos números son negros.

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Table 1.2: Todos los posibles resultados de tirar 2 dados donde uno tiene los números en rojo y otro en negro.

Tenemos 36 posibles resultados (casos posibles). Es razonable pensar que la frecuencia con que cualquiera de ellos aparece es la misma, y por lo tanto $1/36$. Así, por ejemplo, la probabilidad de que salga $(1,1)$ es igual a la de que salga $(1,3)$ es igual a la de que salga $(3,1)$ y es $1/36$. Si queremos calcular la probabilidad de que salga un 1 y un 3, tenemos 2 *casos favorables* a considerar, el $(1,3)$ y el $(3,1)$ y por lo tanto la probabilidad es $1/36+1/36=1/18$, lo que hacemos es *sumar* las probabilidades de los sucesos. Si queremos calcular la probabilidad de que los dos números sean iguales, tenemos 6 *casos favorables*, es decir los 6 que forman la diagonal. Por lo tanto la probabilidad es $6/36=1/6$. Lo que hacemos nuevamente es *sumar* las probabilidades de los sucesos $(1,1)$ hasta $(6,6)$. Supongamos que queremos calcular la probabilidad de que el número que sale en uno de los dados sea estrictamente mayor que el que sale en el otro. Si observamos la tabla, tenemos que contar todos los elementos, excepto los de la diagonal. Esto nos da $36-6=30$ *casos favorables*. Es decir la probabilidad es $30/36$. Veamos que este número lo podemos calcular de otra manera: si contamos la cantidad de posibles resultados para los cuales el dado rojo es mayor que el dado negro esto nos da 15 casos (correspondientes a las entradas de la tabla cuya fila es estrictamente menor que la columna), es decir la probabilidad de que el dado rojo sea mayor que el negro es $15/36$. Por otro lado, si contamos la cantidad de casos en los que el dado rojo es menor que el dado negro obtenemos nuevamente 15 casos (observar que corresponden a las entradas de la tabla cuya fila es estrictamente mayor que la columna). Es decir la probabilidad de que el dado rojo sea menor que el negro es $15/36$. Es claro que la probabilidad de que un dado sea menor que el otro, es la *suma* de estos dos sucesos, y es, como habíamos calculado antes $15/36+15/36=30/36$.

1.3 Propiedades de la probabilidad: uniones y complementos

El ejemplo anterior nos conduce a una propiedad de la probabilidad, (que será cierta en general, no solo en el modelo de casos favorables sobre casos posibles). Si tenemos A y B dos posibles resultados o conjunto de resultados de un experimento (por ejemplo $A=\{$el dado rojo es menor que el dado negro $\}$ y $B=\{$el dado rojo es mayor que el dado negro$\}$), de modo tal que si siempre que pasa A no sucede B (es decir son disjuntos), entonces

$$P(A\cup B)=P(A)+P(B). \tag{1.3}$$

Esto quiere decir que la probabilidad de la unión de *dos* sucesos disjuntos (es decir, la probabilidad de que *alguno* de ellos ocurra) es igual a la suma de las probabilidades de los sucesos.

Análogamente si tenemos A, B, C tres sucesos, tal que si sucede uno, no sucede *ninguno* de los otros 2, entonces

$$P(A \cup B \cup C) = P(A) + P(B) + P(C). \tag{1.4}$$

Por ejemplo, $A = \{$el dado rojo es menor que el dado negro$\}$, $B = \{$el dado rojo es mayor que el dado negro$\}$ y $C = \{$ambos dados son iguales$\}$ (es decir el número que sale es el mismo). Es claro que $A \cup B \cup C$ incluye los 36 casos. Por lo tanto $P(A \cup B \cup C) = 36/36 = 1$, y por lo que calculamos antes esto coincide con $P(A) + P(B) + P(C)$.

Observación 1.10. *Ni* (1.3) *ni* (1.4) *son ciertos en general ya que los sucesos tienen que ser disjuntos dos a dos. Pensemos por ejemplo que* $A = \{$*el dado rojo es menor o igual que el dado negro*$\}$*, mientras que* B *es el suceso* $\{$*el dado rojo es mayor o igual que el dado negro* $\}$*. En este caso* $P(A \cup B) = 1$ *mientras que* $P(A) = 21/36 = P(B)$ *es decir* $P(A) + P(B) = 42/36 > 1$.
Tampoco es cierto en general que si se cumple (1.4) *entonces los sucesos son disjuntos dos a dos.*

Observación 1.11. *En general se cumple que* $P(A \cup B) \leq P(A) + P(B)$ *y* $P(A \cup B \cup C) \leq P(A) + P(B) + P(C)$. *Usando la tabla de tirada de dos dados, encontrar ejemplos de conjuntos* A *y* B *donde se cumple el menor estricto.*

Observación 1.12. *Observar que si* A *y* B *son sucesos tal que* $A \subset B$*, entonces* $B = (B \setminus A) \cup A$ *y además esta unión es disjunta por lo tanto*

$$P(B) = P(B \setminus A) + P(A) \geq P(A),$$

donde en la desigualdad hemos usado que $P(B \setminus A) \geq 0$*. La propiedad anterior se llama monotonía de la probabilidad y en particular implica que si* $P(A) = 0$ *entonces cualquier subconjunto de* A *tiene probabilidad 0.*

Ejemplo 1.13. Consideremos el suceso $A = \{$el dado rojo es menor que el dado negro$\}$, y el suceso complementario, que denotaremos $A^c = \{$el dado rojo *no* es menor que el dado negro$\}$, es decir es mayor o igual. Observemos que dichos sucesos son disjuntos y que la probabilidad de la unión de los mismos es 1. Por lo tanto si usamos (1.3) obtenemos que

$$1 = P(A \cup A^c) = P(A) + P(A^c).$$

Este razonamiento, que hicimos para un A particular *es cierto en general* ya que A y su complementario A^c son disjuntos. De dicho resultado se sigue que, para cualquier suceso A

$$P(A) = 1 - P(A^c). \tag{1.5}$$

Ejemplo 1.14. Consideremos el suceso $A = \{$la suma de los resultados es 7$\}$, y $B = \{$ el dado rojo es menor que el dado negro$\}$. Observando la tabla vemos que $P(A) = 6/36 = 1/6$ y $P(B) = 15/36$. Consideremos ahora el suceso: $\{$el dado rojo es menor que el dado negro *y* la suma de los resultados es 7 $\}$. Este suceso, que

se denota $A \cap B$, esta formado por los resultados $(6,1),(5,2),(4,3)$, por lo tanto $P(A \cap B) = 3/36 = 1/12$. Por otro lado $P(A) \times P(B) = 1/6 \times 15/36 < 1/12$. Es decir, *no es cierto en general* que para todo A y B $P(A \cap B) = P(A) \times P(B)$, el hecho de que esta última igualdad ocurra es importante y es el objeto de estudio de la siguiente sección.

Chapter 2

Probabildiad

2.1 Probabilidad Condicional

Definición 2.1. Probabilidad Condicional Cuando saber que pasó un suceso A nos aporta información sobre la probabilidad de que pase otro suceso B tiene interés calcular la probabilidad de que pase B *dado que* pasó A, esto se denota $P(B|A)$. Si, por ejemplo, siempre que pasa A pasa B, es intuitivo que $P(B|A)$ tiene que ser 1. Por otra parte, si saber que pasó A no me aporta información y por lo tanto no hace a B más o menos probable, razonable también que $P(B|A) = P(B)$. Veamos con un ejemplo como se calcula. Consideremos el suceso $A = \{$ la suma de los dados es 7 $\}$ y $B = \{$ el producto es 10 $\}$. En este caso $P(A) = 1/6$ y $P(B) = 1/18$. Observemos que $P(B|A)$ es la probabilidad de que el producto de 10, *dado que* la suma fue 7. Es decir, sabemos que la suma fue 7 queremos ver que tan probable es que el producto sea 10. Si sabemos que la suma fue 7 tenemos que *restringirnos* a las 6 posibles tiradas que hacen que la suma sea 7, es decir, el conjunto de casos posibles tiene 6 elementos (estos son $(4,3),(5,2),(6,1),(2,5),(3,4),(1,6)$). De esos 6 solamente 2 resultados hacen que el producto sea 10, es decir tenemos 2 casos favorable, por lo tanto $P(B|A) = 2/6 = 1/3$. Observemos que $P(A \cap B) = 2/36 = 1/18$ y que se cumple que

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \tag{2.1}$$

La identidad (2.1) *define* la probabilidad condicional $P(B|A)$. Para que dicha expresión tenga sentido es necesario que $P(A) \neq 0$. Observemos que $P(A \cap B) \neq P(A)P(B)$.

2.2 Independencia

Vamos a introducir ahora uno de los conceptos más importantes del curso, el concepto de independencia de sucesos. Supongamos que tiramos una moneda y sale cara. Luego volvemos a tirar la moneda y sale número. Desde un punto de vista intuitivo, el hecho de que haya salido cara en la primera tirada, *no influye* sobre el

resultado de la segunda tirada, dicho de otra manera, para saber cual será el resultado de la segunda tirada, no nos aporta información saber que la primera tirada fue cara. Lo mismo nos dice la intuición respecto al resultado de tirar un dado, y luego otro, etc. Intuitivamente, cuando el resultado de la realización de un experimento no nos aporta *información* que nos permita deducir a priori, el resultado de otro experimento, decimos que estos son independientes. Observemos que no estamos diciendo que el hecho de que se haya dado un determinado resultado es incompatible con que suceda el otro. Veamos como se formaliza en términos matemáticos esta idea intuitiva.

Ejemplo 2.2. Consideremos nuevamente los 36 resultados de tirar 2 dados, el suceso $A = \{$ la suma de los dados es 7 $\}$, y el suceso $B = \{$ el dado rojo es 1 $\}$. Sabemos que $P(A) = 1/6$ y que $P(B) = 1/6$, además, la probabilidad de que la suma sea 7 *y* que el dado rojo sea 1 (es decir $A \cap B$) es $1/36$, observemos que

$$P(A \cap B) = P(A) \times P(B). \tag{2.2}$$

Definición 2.3. Si tenemos A y B dos sucesos tal que la identidad (2.2) se cumple, decimos que los sucesos A y B son independientes.

Observar que si A y B son independientes (y $P(A) > 0$), de (2.1) se sigue que $P(B|A) = P(B)$ y si siempre que sucede A (nuevamente con $P(A) > 0$) sucede B (es decir $A \subset B$ y por lo tanto $A \cap B = A$), se sigue que $P(B|A) = P(A)/P(A) = 1$. Se deja como ejercicio verificar, usando 2.1 que $P(B|A) + P(B^c|A) = 1$

Observación 2.4. *De la monotonía de la probabilidad (ver Observación 1.12) se deduce que si un suceso tiene probabilidad nula, por ejemplo $P(A) = 0$ entonces A es independiente de cualquier otro suceso B ya que por monotonía $P(A \cap B) \leq P(A) = 0$ con lo cual se verifica (2.2). Esto intuitivamente significa que los sucesos de probabilidad nula son independientes de cualquier otro suceso.*

Ejemplo 2.5. Veamos que se confirma lo que nos dice la intuición respecto a la independencia de la primera y segunda tirada. Si en lugar de tirar los dos dados a la vez, tiramos primero el dado negro y luego el rojo la tabla de posibilidades es (1.2). Es razonable pensar que aquí también cualquier pareja de resultados tiene la misma probabilidad, es decir $1/36$. Si calculamos la probabilidad de que, por ejemplo, el dado rojo sea i (donde i es cualquier número entre 1 y 6), tenemos 6 casos (correspondientes a la columna i de la tabla) por lo tanto el suceso $A = \{$ el dado rojo es i $\}$, tiene probabilidad $1/6$. Por otro lado, si consideramos el suceso $B = \{$ el dado negro es j $\}$ (donde j es cualquier número entre 1 y 6) tenemos que $P(B) = 1/6$ (basta considerar los 6 casos de la fila j de la tabla). El suceso $A \cap B$ que corresponde a que el dado rojo sea i *y* que el dado negro sea j nos da únicamente la pareja (j, i) que tiene probabilidad $1/36$. Es decir se da la igualdad (2.2) y por lo tanto son independientes.

Definición 2.6. Consideremos tres sucesos A, B y C, si

$$\begin{aligned} P(A \cap B \cap C) &= P(A) \times P(B) \times P(C) \\ P(A \cap B) &= P(A) \times P(B) \\ P(A \cap C) &= P(A) \times P(C) \\ P(B \cap C) &= P(B) \times P(C) \end{aligned}$$

decimos que son independientes.

Ejercicio 2.7. Demostrar, a partir de los 8 posibles resultados de tirar 1 moneda 3 veces, que las tiradas son independientes.

Observemos que en el Ejemplo (1.14) tenemos dos sucesos que *no* son independientes, (como vimos, no se cumple (2.2)).

2.3 Fórmula de la probabilidad total

Muchas veces para calcular la probabilidad de un suceso disponemos de las probabilidad de dicho suceso pero condicionado a otros conjuntos que forman una *partición* del espacio Ω (una partición es un conjunto de sucesos $A_1, \ldots, A_k$, cuya unión da todo Ω y que siempre que sucede uno no sucede otro, es decir $A_i \cap A_j = \varnothing$ para todo $i \neq j$). A modo de ejemplo supongamos que se dispone de resultados de un determinado análisis clínico que se realiza para determinar si una persona tiene una determinada enfermedad. En este caso Ω son todos los individuos a los que se le realizó el análisis, y tenemos dos conjuntos que llamaremos S (para indicar a los individuos sanos, es decir no tienen dicha enfermedad) y E para indicar aquellos que si la tienen (ver Figura 2.1). Llamemos T^+ al suceso formado por los individuos (de Ω) cuyo test dio positivo. Supongamos que queremos calcular $P(T^+)$ y disponemos de $P(T^+|E)$, es decir la probabilidad de que el test de positivo *dado que* el individuo está enfermo, de $P(T^+|S)$, esto es, la probabilidad de que el test de positivo dado que está sano y de $P(S)$, la probabilidad de que un individuo elegido al azar en Ω esté sano (observar que como $\{S, E\} = \Omega$ también sabemos $P(E)$). Antes de proceder vamos a introducir algunos conceptos, llamamos

- **Falsos positivos (FP):** El conjunto de los individuos cuyo test da positivo **dado que** no tienen la enfermedad (pertenecen a S). La probabilidad de falso positivo es $P(T^+|S)$
- **Falsos negativos (FN):** El conjunto de los individuos cuyo test da negativo **dado que** son personas que tienen la enfermedad. La probabilidad de falso negativo es $P(T^-|E)$
- **Verdaderos positivos (VP):** El conjunto de los individuos cuyo test da positivo **dado que** son personas que tienen la enfermedad.
- **Verdaderos negativos (VN):** El conjunto de los individuos cuyo test da negativo **dado que** no tienen la enfermedad.

Es claro que $T^+ = [T^+ \cap S] \cup [T^+ \cap E] = VP \cup FP$ y que dicha unión es disjunta ya que E y S son disjuntos. Por lo tanto $P(T^+) = P(T^+ \cap S) + P(T^+ \cap E)$. Ninguna de estas dos probabilidades las tenemos como dato pero las podemos despejar usando 2.1 ya que por ejemplo $P(T^+ \cap S) = P(T^+|S)P(S)$ y $P(T^+ \cap E) = P(T^+|E)P(E)$, por lo tanto hemos llegado a que

$$P(T^+) = P(T^+|S)P(S) + P(T^+|E)P(E).$$

La fórmula anterior es un caso simple de lo que se conoce como la *fórmula de la probabilidad total* que enunciamos en su forma mas general a continuación.

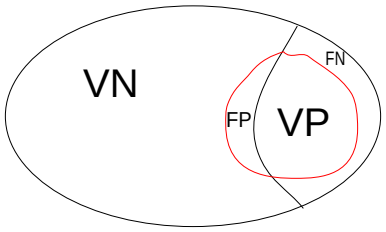


Figure 2.1: La figura representa la población dividida según el resultado del test y según si son sanos o no. Aquellos dentro del conjunto cuyo borde es rojo son los individuos a los que el test les da positivo, que hemos denotado T^+, este conjunto es la unión de los verdaderos positivos (VP) y los falsos positivos (FP). El complemento del conjunto rojo son aquellos cuyo test da negativo que llamamos T^-. Estos pueden estar sanos (y por lo tanto son verdaderos negativos VN), o enfermos (falsos negativos).

Sea $A_1, \ldots, A_k$ una partición de Ω en sucesos disjuntos 2 a 2 ($A_i \cap A_j = \varnothing$ para todo $i \neq j$) y tal que $P(A_j) > 0$ para todo $j = 1, \ldots, k$, sea B otro suceso cualquiera, entonces

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k). \tag{2.3}$$

2.4 Fórmula de Bayes

Supongamos ahora que tenemos dos sucesos A y B cualquiera, con probabilidad no nula, y conocemos $P(B|A)$, $P(A)$ y $P(B)$ y queremos calcular $P(A|B)$. Por definición $P(A|B) = \frac{P(A\cap B)}{P(B)}$. Por otra parte, $P(A\cap B) = P(B|A)P(A)$, por lo tanto llegamos a que

$$P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{P(B|A)}{P(B)}P(A) \tag{2.4}$$

La fórmula (2.4) se denomina **fórmula de Bayes**. En la terminología de la estadística Bayesiana, $P(A)$ se llama *probabilidad a priori* (que representa la probabilidad sin tener en cuenta, o antes de, haber observado B), $P(B|A)$ es la *verosimilitud* (que tan probable es haber observado B en el supuesto de que se cumpla A), y $P(A|B)$ se conoce como probabilidad a posteriori, que representa una *actualización* de la probabilidad a priori, dado que observamos B. El valor $\frac{P(B|A)}{P(B)}$ es el *impacto* de B en la probabilidad de A.

En el caso en que no se conoce $P(B)$ pero si se sabe $P(B|A^c)$, podemos usar la fórmula de la probabilidad total vista en la sección anterior, y obtenemos que $P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$, si combinamos esto con 2.4 obtenemos que

$$P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \tag{2.5}$$

De forma totalmente análoga se demuestra que en general si se tienen $A_1, \ldots, A_k$ sucesos, todos ellos con probabilidad no nula, disjuntos dos a dos (es decir $A_i \cap A_j = \varnothing$ para todo $i \neq j$) y tal que siempre ocurre alguno de ellos (es decir $A_1 \cup A_2 \cup \ldots \cup A_k = \Omega$), entonces:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}. \tag{2.6}$$

La fórmula 2.6 se deduce de 2.4, sustituyendo $P(B)$ por 2.3.

Ejemplo 2.8. Un laboratorio ha desarrollado una prueba para el diagnóstico de la hepatitis tipo C. Esta prueba tiene un 98% de exactitud entre los que tienen la enfermedad y un 80% entre los que no la tienen. Supongamos que el 0.5% de la población tiene la enfermedad. Denotemos S al conjunto de los individuos sanos y E al de los individuos que tienen la enfermedad, por otra parte denotemos T^+ a aquellos cuyo test dio positivo (que se representa como el conjunto en rojo en la figura 2.1) y como T^- a aquellos cuyo test da negativo (el complemento del conjunto en rojo). Según los datos que tenemos, $P(E) = P(VP) + P(FN) = 0.005$. Por lo tanto la probabilidad de que un individuo no tenga la enfermedad es $P(S) = 1 - 0.005 = 0.995$. Que el test tenga un 98% de exactitud entre los que tienen la enfermedad se traduce en que $P(T^+|E) = 0.98$ (esta probabilidad **no** es $P(VP)$). Por otra parte que la exactitud sea 80% entre aquellos que no la tienen es $P(T^-|S) = 0.8$, por lo tanto $P(T^+|S) = 1 - 0.8 = 0.2$. La probabilidad *a priori* de que una persona tenga la enfermedad es 0.005, no obstante si le hacemos el test y da positivo, la probabilidad de que efectivamente la tenga cambia (esto es $P(E|T^+)$). Esta probabilidad es la probabilidad a posteriori. Veamos como se modifica en nuestro ejemplo. Si usamos la fórmula de Bayes 2.4

$$P(E|T^+) = \frac{P(T^+|E)P(E)}{P(T^+|E)P(E) + P(T^+|S)P(S)} = \frac{0.98 \times 0.005}{0.98 \times 0.005 + 0.2 \times 0.995} = 0.024,$$

es decir se multiplicó casi por 5 la probabilidad de que este enfermo.

Ejemplo 2.9. La probabilidad de que haya un accidente en una fábrica que dispone de alarma es 0.1. La probabilidad de que suene la alarma si hay un accidente es 0.97 y de que suene si no hay ninguno es 0.02. Si sabemos que sonó la alarma, ¿cuál es la probabilidad de que no haya habido un accidente?.
Vamos a traducir las probabilidades que queremos a sucesos, consideremos el suceso I = hubo accidente (denotamos I^c al complemento de I es decir, no hubo accidente), y el suceso A = sonó la alarma. Lo que sabemos es que $P(I) = 0.1$, $P(A|I) = 0.97$ y $P(A|I^c) = 0.02$, lo que queremos calcular es $P(I^c|A)$. Los sucesos I y I^c son disjuntos y su unión tiene probabilidad 1. Por lo tanto si usamos la fórmula de Bayes (2.5) escribimos

$$P(I^c|A) = \frac{P(A|I^c)P(I^c)}{P(A|I)P(I) + P(A|I^c)P(I^c)} = \frac{0.9 \times 0.02}{0.1 \times 0.97 + 0.9 \times 0.02} = 0.157.$$

En R

Si queremos por ejemplo elegir k números al azar (distintos) de entre los primeros n podemos usar la función sample(n,k),

```
sample(10,3)

## [1] 4 7 5
```

Si queremos permitir que se puedan repetir alguno de los n en la muestra de tamaño k tenemos que usar sample(n,k,replace=TRUE):

```
sample(10,7,replace=TRUE)

## [1] 5 5 3 8 6 9 9
```

Chapter 3

Variables aleatorias discretas

3.1 Distribución Binomial

En esta sección vamos a centrarnos en el estudio de un tipo de experimento muy particular en el cual existen exactamente 2 posibles resultados, por ejemplo los resultados que surgen de tirar una moneda. Otro ejemplo simple consiste en considerar si al tirar un dado sale o no un cierto número. Usualmente se llama *éxito* a uno de estos resultados y *fracaso* al otro (éxito podría ser que salga 5 al tirar un dado). Es claro que como tenemos 2 posibilidades (éxito o fracaso) podemos decir que la frecuencia con que ocurre éxito si repetimos el experimento muchas veces (esto es P(éxito)) es un número p donde $0 \leq p \leq 1$, mientras que la frecuencia con que *no* ocurre el éxito (es decir ocurre el *fracaso*) es $1 - p$ (deducirlo a partir de (1.5)). Lo que nos interesará es, fijada de antemano la cantidad de veces que repetiremos el experimento (y bajo la hipótesis de que *los experimentos se realizan de forma independiente*), calcular qué tan probable es obtener determinada cantidad de éxitos. Por ejemplo, si tiramos 50 veces el dado, qué tan probable es que el 5 salga exactamente 10 veces, es decir, la probabilidad de que la cantidad de éxitos sea 10, o, lo que es lo mismo, que la cantidad de fracasos sea exactamente 40.

Veamos cómo es la construcción del marco teórico de este modelo. Es natural que, para modelar el experimento que consiste en tirar 3 veces un dado y mirar la cantidad de veces que salió el número 5 pensemos en tomar Ω como el conjunto de ternas de ceros y unos, donde 1 es éxito, es decir salió el 5. Así por ejemplo un posible resultado es $(1,0,0)$, que representa el experimento en el cual en la primera tirada salió el 5 mientras que en las dos siguientes no salió el 5. Es claro que dicho resultado es distinto de $(0,0,1)$ o de $(1,1,1)$. Observemos que Ω tiene 8 elementos (escribirlos). Observemos además que, si las tiradas son independientes, el suceso $(1,0,0)$ tiene probabilidad $1/6 \times 5/6 \times 5/6$ ya que $P((1,0,0)) = P(\{\text{En la primera tirada sale el } 5\} \cap \{\text{En la segunda tirada no sale el } 5\} \cap \{\text{En la tercera tirada no sale } 5\})$ y como son independientes, el resultado se sigue de (2.6).

Por otro lado el el suceso $(1,1,1)$ tiene menor probabilidad ($P((1,1,1)) = 1/6 \times 1/6 \times 1/6$, esto se sigue razonando de forma análoga a como hicimos con $(1,0,0)$). De esto último deducimos que *no* estamos ante un modelo de casos favorables sobre casos posibles.

Ejercicio 3.1. Asignar a cada una de las 8 ternas de ceros y unos sus probabilidades.

Volviendo al problema que mencionamos antes, queremos calcular la probabilidad de tener una determinada cantidad de éxitos. En el caso de repetir 3 veces el experimento de tirar el dado, si éxito es que salga el número 5 es claro que podemos tener:

1) 0 éxito, como en $(0,0,0)$,

2) exactamente 1 éxito, como en $(1,0,0)$ o $(0,1,0)$ o $(0,0,1)$,

3) exactamente 2 éxitos, como en $(1,1,0)$ o $(0,1,1)$ o $(1,0,1)$,

4) 3 éxitos, como en $(1,1,1)$.

Si lo que nos interesa es contar la cantidad de éxitos, los 3 resultados posibles del caso 2 se pueden agrupar en el suceso $\{$exactamente 1 éxito$\}$ y podemos asignarle a esos 3 elementos de Ω el número 1. Análogamente, los 3 elementos del caso 3 se pueden agrupar en el suceso $\{$exactamente 2 éxitos $\}$ y podemos asignarle a esos 3 elementos de Ω el número 2. Finalmente, razonando de manera análoga, al elemento del caso 1 le asignamos el 0 mientras que al del caso 4 le asignamos el 3, ver figura (3.1).

Esta función (la cual denotaremos con la letra X) que hemos definido, de Ω en los números $\{0,1,2,3\}$ se llama **variable aleatoria con distribución binomial**. En general, una variable aleatoria será una función X de Ω en los números reales $\mathbb{R}$. Y nos interesará calcular cosas como $P(X = t)$ o $P(X \leq t)$, donde t es un número real, entendiendo por eso la probabilidad de el conjunto de los elementos de Ω que hacen que $X = t$.

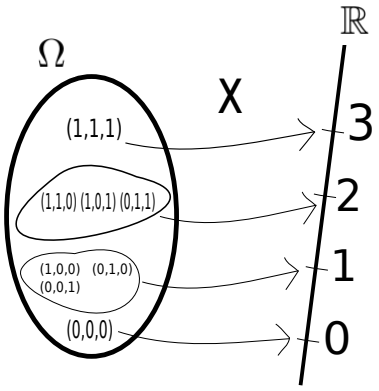


Figure 3.1: Variable aleatoria Binomial en el caso de 3 experientos

En nuestro ejemplo anterior si X tiene distribución binomial,

$$P(X = 2) = P((1,1,0) \cup (0,1,1) \cup (1,0,1)) = P((1,1,0)) + P((0,1,1)) + P((1,0,1)),$$

donde en esta igualdad hemos usado la propiedad (1.4). Dado que $P((1,1,0)) = P((0,1,1)) = P((1,0,1)) = (1/6)^2 \times 5/6$ tenemos que

$$P(X=2) = 3 \times (1/6)^2 \times 5/6. \tag{3.1}$$

De la misma forma se llega a que

$$P(X=1) = 3 \times (1/6) \times (5/6)^2 \tag{3.2}$$

Vamos a calcular $P(X=0)$, tenemos un sólo caso (caso 1, el $(0,0,0)$), que tiene probabilidad $(5/6)^3$, esto es lo mismo que hacer

$$P(X=0) = 1 \times (1/6)^0 \times (5/6)^3. \tag{3.3}$$

Verificar que en (3.1), (3.2) y (3.3) obtuvimos, para $k = 0,1,2,3$ éxitos:

$P(X=k)$ =(formas de poner k unos en 3 lugares)$\times (1/6)^k \times (5/6)^{3-k}$

Veamos como contar en términos de combinaciones la cantidad de formas de poner k unos en 3 lugares (con $k = 0,1,2,3$). Recordemos que $\binom{n}{k}$ cuenta la cantidad de subconjuntos distintos, de k elementos, que podemos hacer con n objetos diferentes. Podemos pensar que los $n = 3$ objetos distintos son los lugares donde vamos a poner los k unos, y como los unos son intercambiables (una vez elegidos en que lugares se pondrán) estamos ante un problema de combinaciones. Es decir, podemos poner los unos en la primera tirada, en la segunda, o en la tercera. Por ejemplo, para $k = 2$ elegir *primera tirada* y *tercera tirada* produce la misma cantidad de éxitos que elegirlos en el orden inverso: *tercera tirada* y *primera tirada* y da lugar a la terna $(1,0,1)$, pero es distinto que elegir *primera tirada* y *segunda tirada* ya que en este caso estamos eligiendo de los 3 posibles lugares, los dos primeros (y da lugar a la terna $(1,1,0)$). Es decir tenemos $\binom{3}{2}$ formas de elegir esos lugares. Es claro que si tuvimos 1 solo éxito, tenemos tres posibles lugares donde poner ese uno, y por lo tanto son $\binom{3}{1}$, finalmente si no se dio éxito en ninguna de las tiradas tenemos una sola posibilidad, la terna $(0,0,0)$ y nuevamente aquí tambien $\binom{3}{0} = 1$.

Definición 3.2. Bin(n,p): Siguiendo el razonamiento que hicimos antes, si tenemos n experimentos, y $p = P(\text{éxito})$ decimos que $X : \Omega \to \{0,1,2,\ldots,n\}$ tiene distribución binomial con parámetros n, p que se denota $X \sim \text{Bin}(n,p)$ si

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0,1,2,\ldots,n \tag{3.4}$$

Definición 3.3. Ber(p): En el caso particular en que $n = 1$ (es decir hacemos un sólo experimento) se dice que la variable tiene distribución de Bernoulli de parámetro p. Más adelante veremos que si X tiene distribución Binomial de parámetros n y p, se puede escribir como suma de n variables con distribución de Bernoulli de parámetro p.

Ejercicio 3.4. Usando la expresión anterior calcular la probabilidad de que al tirar 50 veces un dado salgan exactamente 10 cincos.

Ejercicio 3.5. Supongamos que $X \sim \text{Bin}(n, 1/2)$

- Intuitivamente, sin calcular, ¿Cuánto da $P(X=0) + P(X=1) + P(X=2) + \cdots + P(X=n)$?

- Usando la parte anterior y (3.4) deducir que $\binom{n}{0}+\binom{n}{1}+\binom{n}{2}+\cdots+\binom{n}{n}=2^n$ y observar que 2^n es exactamente la cantidad total de secuencias de largo n que podemos formar con ceros y unos.

En R

Para sortear números al azar con distribución Binomial tenemos el comando rbinom(N,n,p). El parámetro N indica cuantas realizaciones queremos de la variable, el parámetro n indica que cada una de estas realizaciones da como resultado un número entre $0,\ldots,n$ (cantidad de éxitos), finalmente p es la probabilidad de éxito en cada realización. Por ejemplo para tener una Bernoulli de parámetro p, tenemos que hacer rbinom(1,1,p).

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

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

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

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

3.2 Distribución Geométrica

Supongamos que estamos interesados en repetir un experimento aleatorio que tiene dos posibles resultados, uno de ellos que llamaremos éxito, y el otro fracaso, hasta que se da el éxito por primera vez, y queremos contar la cantidad total de fracazos. Supondremos además que el experimento lo repetimos de forma independiente de lo que sucedió antes. Por ejemplo podemos pensar que tiramos una moneda hasta que salga cara, o un dado hasta que salga 5 etc. En el caso del dado, el 5 puede aparecer por primera vez en la primera tirada, o puede que en la primera tirada no salga 5 pero en la segunda si, etc. La probabilidad de que salga 5 en la primera tirada es $1/6$, en este caso tuvimos 0 fracasos. La probabilidad de que salga por primera vez en la segunda tirada (es decir tuvimos un fracaso) es la probabilidad de la intersección de los sucesos $A=\{$ no sale 5 en la primera tirada $\}$ y $B=\{$ sale 5 en la segunda tirada$\}$. Dado que estamos suponiendo que los experimentos se realizan de forma independiente, es decir saber que salió 5 en la primera tirada *no influye* sobre el resultado de la segunda, tenemos que $P(A\cap B)=P(A)\times P(B)$. Como $P(A)=1/6$ mientras que $P(B)=5/6$ tenemos que la probabilidad de que salga el 5 por primera vez en la segunda tirada es $5/6\times 1/6$.

Ejercicio 3.6. Probar que la probabilidad de que salga 5 por primera vez en la tercera tirada es $(5/6)^2\times 1/6$.

En general, razonando de esta manera, es fácil ver que la probabilidad de que salga 5 por primera vez, en la tirada k, es decir tuvimos $k-1$ fracasos, es la intersección de los sucesos independientes $A=\{$ no sale 5 en las tiradas $1,2,\ldots,k-1\}$ y $B=\{$sale 5 en la tirada $k\}$. Dicha intersección tiene probabilidad $(5/6)^{k-1}\times 1/6$.

Esto nos conduce a la siguiente definición:

Definición 3.7. Decimos que una variable aleatoria X tiene **distribución geométrica** de parámetro p con $0 \leq p \leq 1$ (que denotaremos $X \sim \text{Geo}(p)$) si

$$P(X = k) = (1-p)^k \times p \quad k = 0, 1, 2, \ldots \tag{3.5}$$

Observemos que, a diferencia de lo que sucede con las variables con distribución binomial de parámetros n, p, una variable con distribución geométrica es una función que toma infinitos valores $(0, 1, 2, 3, \ldots)$ mientras que una variable con distribución binomial de parámetros n, p toma valores $0, 1, 2, \ldots, n$, donde n es un número fijado de antemano.

Ejercicio 3.8. Supongamos que tiramos un dado hasta que sale 3

Calcular la probabilidad de que salga por primera vez en la primera tirada o en la segunda.

Calcular la probabilidad de que haya que realizar más de 2 tiradas.

Vamos a demostrar una propiedad importante de la distribución geométrica, que se conoce como *pérdida de memoria*. Para eso vamos a necesitar probar un resultado intermedio, que dice que si $X \sim \text{Geo}(p)$ entonces, para todo $k \geq 0$,

$$P(X \geq k) = (1-p)^k, \tag{3.6}$$

para probar esto último observemos que,

$$P(X \geq k) = P(X = k) + P(X = k+1) + P(X = k+2) + P(X = k+3) + \ldots + \ldots$$

esto último es una suma de infinitas probabilidades, y se escribe de manera compacta como

$$P(X \geq k) = \sum_{i=0}^{\infty} P(X = k+i),$$

si usamos (3.5) tenemos que $P(X = k+i) = (1-p)^{k+i} p$ y por lo tanto

$$P(X \geq k) = \sum_{i=0}^{\infty} P(X = k+i) = \sum_{i=1}^{\infty} (1-p)^{k+i} p.$$

Veamos que es lo que estamos sumando, para $i = 0$ el primer término es p, para $i = 1$ es $(1-p)^k p$, para $i = 2$ es $(1-p)^{k+1} p$, etc, por lo tanto la suma anterior se puede escribir como

$$\sum_{i=0}^{\infty} (1-p)^k (1-p)^i p,$$

como el factor $(1-p)^k p$ no depende de i, sale de factor común para afuera de la sumatoria, es decir

$$\sum_{i=0}^{\infty} (1-p)^k (1-p)^i p = (1-p)^k p \sum_{i=0}^{\infty} (1-p)^i.$$

Ahora usamos un resultado de series, que no demostraremos que dice que

$$\sum_{i=0}^{\infty}(1-p)^i = \frac{1}{p}, \tag{3.7}$$

finalmente obtuvimos, como queríamos

$$P(X) = (1-p)^k, \quad \text{para todo } k = 0,1,2,\ldots$$

Otra forma de ver (3.6) es observar que si $X \geq k$ en los primeros k experimentos no obtuvimos éxito, y esto pasa con probabilidad $(1-p)^k$ ya son k eventos independientes

La propiedad de **pérdida de memoria** dice que, la probabilidad de obtener por lo menos $n+m$ fracasos, con $n, m \geq 0$, condicionada a que ya obtuviste por lo menos m fracasos, es decir $P(X \geq n+m|X \geq m)$, es igual a la probabilidad de obtener por lo menos n fracasos. Esto en términos de probabilidad condicional se escribe como

$$P(X \geq n+m|X \geq m) = P(X \geq n). \tag{3.8}$$

Esto significa que la variable se olvida que ya realizamos m experimentos en los que no hubo éxito (de ahí el nombre *pérdida de memoria*). Para demostrar que se se cumple (3.8)vamos a usar la definición de probabilidad condicional,

$$P(X \geq n+m|X \geq m) = \frac{P(\{X \geq n+m\} \cap \{X \geq m\})}{P(X \geq m)} \tag{3.9}$$

Veamos ahora cual es el evento $\{X \geq n+m\} \cap \{X \geq m\}$; si $X \geq n+m$ quiere decir que hasta el momento $m+n$ no tuvimos éxito, en particular hasta el momento m no tuvimos éxito (ya que $n+m \geq m$), por lo tanto el evento $\{X \geq m+n\}$ incluye al evento $\{X \geq m\}$, es decir $\{X \geq n+m\} \cap \{X \geq m\} = \{X \geq n+m\}$, y por lo tanto

$$P(\{X \geq n+m\} \cap \{X \geq m\}) = P(X \geq n+m)$$

Si ahora usamos la fórmula (3.6) tenemos que $P(X \geq n+m) = (1-p)^{m+n}$ y $P(X \geq m) = (1-p)^m$, por lo tanto (3.9) nos queda

$$P(X \geq n+m|X \geq m) = \frac{P(\{X \geq n+m\} \cap \{X \geq m\})}{P(X \geq m)} = \frac{(1-p)^{n+m}}{(1-p)^m} = (1-p)^n = P(X \geq n).$$

en la última igualdad hemos usado que (3.6)

En R

Es importante notar que en R la variable con distribución geométrica cuenta la cantidad de fracasos hasta el primer éxito. Si queremos una realización de una variable con distribución geométrica de parámetro p, tenemos el comando `rgeom(1,p)`. Esto nos da el número de fracasos antes del primer éxito, si la probabilidad

de éxito es p. Si queremos el primer momento en que se da el éxito hay que sumar 1. Para obtener N realizaciones de esta variable se usa rgeom(N,p).

```
rgeom(1,1/4)
```

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

```
rgeom(10,1/10)
```

```
##  [1]  1  0 12  1  1 12 14  6  4  1
```

3.3 Distribucción Hipergeométrica: Extracciones sin resposición

Veremos ahora otro tipo de variable aleatoria. Supongamos que tenemos una urna con 6 bolas de color rojo y 3 de color negro como en (1.8). Vamos a suponer que la diferencia es únicamente el color. Si extraemos una bola al azar, como la diferencia está solamente en el color, podemos agarrar cualquiera de las 9 bolas pero, intuitivamente, la probabilidad de extraer una bola roja es mayor (ya que son más) que la de extraer una bola negra. Dado que cualquiera de las 9 bolas tiene la misma probabilidad (es decir $1/9$), el espacio muestral Ω esta formado por los 9 posibles resultados. Por lo tanto el suceso $A = \{$ la bola extraída es roja$\}$ tiene probabilidad $6/9 = 2/3$, mientras que el suceso $B = \{$ la bola extraída es negra $\}$ tiene probabilidad $3/9$. Consideremos ahora el experimento que consiste en sacar 5 bolas (al mismo tiempo) de las 9, supongamos que queremos calcular la probabilidad de extraer 3 bolas de color rojo, y 2 de color negro. Como las bolas son todas idénticas, cualquier conjunto de 5 bolas tiene la misma probabilidad, por lo tanto en total tenemos $\binom{9}{5}$ (*casos posibles*) formas de extraer las 5 bolas. Como vimos en (1.8), los *casos favorables* son $\binom{6}{3} \times \binom{3}{2}$ es decir, la probabilidad que queremos calcular es

$$\frac{\binom{6}{3} \times \binom{3}{2}}{\binom{9}{5}}.$$

Ejemplo 3.9. Razonando de la misma forma que antes, si en lugar de sacar 5 bolas sacamos 3 y queremos tener 2 rojas y 1 negra, obtenemos que la probabilidad es

$$\frac{\binom{6}{2} \times \binom{3}{1}}{\binom{9}{3}} = \frac{15}{28}. \tag{3.10}$$

Una pregunta que surge naturalmente es qué sucede si en lugar de sacar las 5 bolas al mismo tiempo, las sacamos de a una (y las que vamos sacando no las devolvemos a la urna, es decir son *extracciones sin reposición*). Veamos, para el caso en que extraemos 3 bolas (sólo a efectos de hacer menos engorrosas las cuentas), que el resultado es *el mismo* que el que obtuvimos en el ejemplo (3.10). Observemos que

tenemos que sumar la probabilidad de los sucesos (disjuntos) $A = \{$Sale roja en la primer extracción, roja en la segunda, y negra en la tercera$\}$, $B = \{$Sale roja en la primer extracción, negra en la segunda, roja en la tercera$\}$ y $C = \{$Sale negra en la primer extracción, roja en la segunda, y roja en la tercera$\}$. Vamos a calcular la probabilidad de C y quedará como ejercicio verificar que A y B tienen la misma probabilidad.
Observemos que la probabilidad de que salga negra en la primer extracción $3/9$. En la segunda extracción, *dado que* sacamos una bola negra, nos quedan 8 bolas en la urna, 6 son de color rojo y 2 de color negro, la probabilidad de sacar una bola roja ahora es $6/8$. Finalmente, en la tercer extracción, *dado que* sacamos una bola negra en la primera y una roja en la segunda, tenemos 7 bolas en la urna, de las cuales 5 son rojas y 2 son negras, por lo tanto la probabilidad de sacar una bola roja es $5/7$. De esta manera, llegamos a que $P(C) = 3/9 \times 6/8 \times 5/7$. Como dijimos antes, A, B y C tienen la misma probabilidad (verificarlo!) por lo tanto la probabilidad que queremos es $3 \times P(C) = 3 \times \left(3/9 \times 6/8 \times 5/7\right) = 15/28$ que es lo mismo que obtuvimos en el ejemplo (3.10). La pregunta que nos queda por responder ahora es: ¿por qué multiplicamos las probabilidades que fuimos calculando?. Veamos que esto sale de la aplicación *en cadena* de la fórmula de probabilidad condicional dada en (2.1). Antes de eso observemos que si despejamos en (2.1) obtenemos que.

$$P(B|A) \times P(A) = P(B \cap A) \tag{3.11}$$

Volviendo a lo que queríamos calcular:

$$P(C) = P\Big(\{\text{negra en la primera}\} \cap \{\text{roja en la segunda}\} \cap \{\text{roja en la tercera}\}\Big)$$

si tomamos en (3.11) $A = \{$ negra en la primera $\} \cap \{$ roja en la segunda$\}$ y $B = \{$roja en la tercera $\}$ obtenemos que

$$P(C) = P\Big(\{\text{roja en la tercera}\}\Big|\{\text{ roja en la segunda}\} \cap \{\text{ negra en la primera}\}\Big) \times P\Big(\{\text{ roja en la segunda}\} \cap \{\text{ negra en la primera}\}\Big) \tag{3.12}$$

Si volvemos a aplicar (3.11) con $B = \{$ roja en la segunda$\}$ y $A = \{$ negra en la primera$\}$ obtenemos que

$$P\Big(\{\text{ roja en la segunda}\} \cap \{\text{ negra en la primera}\}\Big) = P\Big(\{\text{roja en la segunda}\}\Big|\{\text{negra en la primera}\}\Big) \times P\Big(\{\text{negra en la primera}\}\Big) = \frac{6}{8} \times \frac{3}{9} \tag{3.13}$$

Por otra parte

$$P\Big(\{\text{roja en la tercera}\}\Big|\{\text{ roja en la segunda}\} \cap \{\text{ negra en la primera}\}\Big) = \frac{5}{7}. \tag{3.14}$$

De (3.14) (3.13) y (3.12) obtenemos $6/8 \times 3/9 \times 5/7$ que es a lo que habíamos llegado antes con lo cual queda demostrado que es lo mismo extraer las 3 bolas al mismo tiempo o de a una, siempre y cuando no se vuelvan a colocar en la urna. El problema de hacer el cálculo mediante la segunda forma está en que hay que contar

todas las posibles formas de sacar una bola roja, una negra, y otra roja (en algún orden), que dio lugar a los sucesos A, B y C. En este ejemplo contarlos es simple. Luego de que sabemos *cuantos* son, basta calcular la probabilidad de uno de ellos y multiplicarlo por el número de casos ya que los diferentes casos tienen siempre la misma probabilidad.

Definición 3.10. En general, cuando tenemos un conjunto de N objetos, de los cuales $0 \leq d \leq N$ son de un tipo (llamemosle A) y $N-d$ son de otro, la probabilidad de obtener x elementos de A (con $x = 0, 1, 2, \ldots, d$) en una extracción de n elementos es

$$\frac{\binom{d}{x} \times \binom{N-d}{n-x}}{\binom{N}{n}}. \tag{3.15}$$

Decimos que una variable aleatoria $X : \Omega \to \mathbb{R}$ tiene **distribución hipergeométrica** (que denotaremos $X \sim \text{HipGeo}(d, n, N)$) si

$$P(X = x) = \frac{\binom{d}{x} \times \binom{N-d}{n-x}}{\binom{N}{n}} \quad x = 0, 1, 2, \ldots, \min\{n, d\}.$$

En R

Para sortear nn datos al azar con distribución hipergeométrica (ver 3.10) se puede usar la función rhyper(nn,d,N-d,n). Por ejemplo si hacemos rhyper(10,3,5,2) nos va a devolver una secuencia de 10 números donde cada uno puede ser 0, 1 o 2. En cada una de las 10 tiradas, la probabilidad de que salga 0, 1 o 2 viene dada por la fórmula (3.15).

```
rhyper(10,3,5,2)

##  [1] 2 1 0 2 0 1 2 0 1 1
```

3.4 Distribución Multinomial

La distribución multinomial es una generalización de la binomial, en cada experimento los resultados posibles pueden ser mas de 2. A modo de ejemplo supongamos que tiramos un dado n veces, cada uno de los posibles resultados en lugar de ser un 0 o un 1, es un valor entre 1 y 6, y nos interesa contar la cantidad de veces que salió el 3. Otro ejemplo es pensar que hacemos 5 extracciones con resposición de una urna que tiene bolas de 3 colores, rojo (R) azul (A) y negro (N). Supongamos que la probabilidad de sacar una bola roja es p_1 la de sacar una bola azul es p_2 y la de sacar una bola negra es p_3. Como son los únicos colores posibles $p_1 + p_2 + p_3 = 1$. Si queremos calcular la probabilidad de sacar 2 bolar rojas, 1 azul y 2 negras sin importar el orden, tenemos que contar de cuantas maneras posibles se puede realizar dicha extracción (una posibles puede ser RARANN y otra distinta es RRAANN por ejemplo) y multiplicarlo por $p_1^2 p_2 p_3^2$ que es la probabilidad de cualquiera de esas extracciones. Para contar la cantidad de maneras posibles de realizar dicha extracción observemos que tenemos $\binom{5}{2}$ formas distintas de elegir las dos extracciones donde sacamos la bola roja. Nos

quedaron $5-2$ lugares para las otras 3 extracciones. Por cada una de las posibilidades anteriores tenemos $\binom{5-2}{1} = \binom{3}{1}$ formas de elegir donde poner la bola azul, es decir en que extracción la sacamos. Ahora nos quedan solamente 2 lugares y por lo tanto tenemos $\binom{2}{2}$ formas de elegir donde poner las bolas azules. En total nos dio

$$\binom{5}{2}\binom{3}{1}\binom{2}{2} = \frac{5!}{2!1!2!}.$$

En el caso general, razonando de esta forma, si hacemos $n > 0$ extracciones *con resposición* de una urna que tiene bolas de k colores distintos (y estos son los únicos colores que hay en la urna) y tenemos que sacar $0 \leq x_1 \leq n$ de un color, $0 \leq x_2 \leq n$ de otro color distinto, etc, hasta $0 \leq x_k \leq n$ del color k, con $x_1 + x_2 + \cdots + x_k = n$ tenemos

$$\binom{n}{x_1}\binom{n-x_1}{x_2}\binom{n-x_1-x_2}{x_3}\cdots\binom{n-x_1-x_2-x_3\cdots-x_{k-1}}{x_k} \tag{3.16}$$

formas distintas de realizar dicha extracción. Si se desarrollan las combinaciones de (3.16) se llega a que lo anterior es igual a

$$\frac{n!}{x_1!x_2!\ldots x_k!}. \tag{3.17}$$

Denotemos p_1 a la probabilidad de sacar una bola del color 1, p_2 la probabilidad de sacar una bola del color 2, hasta p_k, la de sacar una bola del color k. Como son los únicos colores $p_1 + p_2 + \cdots + p_k = 1$. La probabilidad de extraer exactamente x_1 del color 1, x_2 del color 2 hasta x_k del color k (con las restricciones sobre los x_i que usamos para deducir (3.17) es

$$\frac{n!}{x_1!x_2!\ldots x_k!}p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k}. \tag{3.18}$$

Se deja como ejercicio verificar que en el caso en que $k = 2$ se obtiene la distribución binomial con parámetro $p = p_1$. Observar que aquí los resultados posibles de repetir n veces el experimento no son números reales, sino n uplas que pueden ser por ejemplo los n colores que salieron. Si denotamos X_i la variable aleatoria que indica la cantidad de veces que salió el color i en las n repeticiones tenemos una variable con distribución binomial de parámetro n, p_i siendo p_i la probabilidad de que en una extracción salga una bola del color i.

3.5 Distribución de Poisson

Hasta ahora hemos definido variables aleatorias a partir de cierto tipo de experimentos, vimos que ellas encierran la información que nos permite calcular probabilidades relacionadas a ellos. Por ejemplo, la variable aleatoria con distribución binomial de parámetros n y p la asociamos al modelo que nos permite contar la cantidad de éxitos en la repetición de n experimentos independientes en los cuales la probabilidad de éxito es p. La distribución geométrica a contar intentos hasta obtener éxito, y la hipergeométrica a contar extracciones de una urna que contiene dos tipos de elementos. En el caso de la distribución de Poisson que definiremos ahora, existen algunos fenómenos fisicos que se modelan con dicha distribución. Vamos a dar la definición y hacer algunos comentarios.

Definición 3.11. Una variable aleatoria con **distribución de Poisson** es un función $X : \Omega \to \mathbb{R}$ que toma solamente los valores $0, 1, 2, 3\ldots$ con ciertas probabilidades que dependen de un parámetro $\lambda > 0$, esto se

denota $X \sim \text{Poisson}(\lambda)$. Las probabilidades son:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, 3, \ldots. \tag{3.19}$$

Observemos que, al igual que lo que sucedía con la distribucción geométrica, el recorrido de la variable es infinito (todos los números naturales). A diferencia de la geométrica (que tomaba valores a partir de 1) la Poisson toma el valor 0 con probabilidad $e^{-\lambda}$ como se sigue inmediatamente a partir de tomar $k = 0$ en (3.19). Si $X \sim \text{Poisson}(\lambda)$, dado que los *únicos* valores que toma son $0, 1, 2, 3, \ldots$ tenemos que $P(X = 0) + P(X = 1) + P(X = 2) + \cdots = 1$ (esta suma tiene infinitos sumandos). De esto último se deduce que si k tiende a infinito, $P(X = k)$ tiende a cero muy rápidamente. Una pregunta natural es: ¿qué papel juega λ en la definición de la variable aleatoria?. Supongamos que tomamos, para fijar ideas $\lambda = 3$. El concepto de probabilidad como frecuencia nos dice que si sorteamos por ejemplo $n = 1$ millón de números con distribución de Poisson(3), aproximadamente $e^{-3} \times n$ de veces va a salir el 0 (es decir la proporción de veces que sale 0 es $P(X = 0)$), aproximadamente $\lambda e^{-\lambda} \times n$ veces va a salir 1 (es decir la proporción de veces que sale 1 es $P(X = 1)$), aproximadamente $(\lambda^2/2) \times e^{-\lambda} \times n$ veces va a salir 2, etc. Si promediamos esos números, es decir sumamos los n números que salieron (de modo tal que si el 0 salió 1560 veces, sumamos 1560 veces 0, si el 1 salió 3450 sumamos 3450 veces 1 etc) y dividimos entre 1 millón, el número que vamos a obtener va a estar próximo a λ que en nuestro caso es 3. Es decir, λ es un parámetro (que se denomina *valor esperado*) que tiene que ver con lo que obtenemos si promediamos *muchos* datos que siguen la distribución Poisson(λ). El concepto de valor esperado es importante en la teoría de la probabilidad, y aparecerá a lo largo del texto en muchas oportunidades.

Una propiedad interesante de la distribución de Poisson, que hace que se use para modelar ocurrencias de un evento por unidad de tiempo (por ejemplo llamadas a una central telefónica recibidas en un período de 1 hora) es que la suma de variables independientes con distribución de Poisson de parámetros λ_1 y λ_2 es otra variable con distribución de Poisson de parámetro $\lambda_1 + \lambda_2$. Esto, intuitivamente, significa que el número de ocurrencias de un determinado evento, cuando lo miramos en dos intervalos de tiempo iguales pero disjuntos, es la suma de el número de ocurrencias en cada intervalo. En el ejemplo de las llamdas telefónicas, la cantidad de llamadas que se reciben entre la 1 y las 2 y entre las 3 y las 4, es la suma de las que se recibe entre la 1 y las 2, más las que se reciben entre las 3 y las 4 (asumiendo la hipótesis, razonable en este caso, de que las llamadas entre la 1 y las 2 son independientes de las que se reciben entre las 3 y las 4).

La distribución de Poisson se usa para modelar algunos fenómenos de la naturaleza, por ejemplo, el número de mutaciones de determinada cadena de ADN después de cierta cantidad de radiación, el número de núcleos atómicos inestables que se han desintegrado en un determinado período de tiempo. El número de autos que pasan a través de un cierto punto en una ruta, etc.

En R

Si queremos sortear n variables de acuerdo a la distribución (3.19) el comando es `rpois(n,λ)`.

```r
rpois(17,3)
```

```
##  [1] 3 2 4 2 5 6 1 6 3 1 3 1 3 4 3 6 3
```

Chapter 4

Variables aleatorias continuas

4.1 Distribución Uniforme

Hasta ahora hemos visto variables que toman una cantidad finita, o numerable de valores. Como vimos con la distribución geométrica y de Poisson, si toma infinitos valores, no puede tomarlos con la misma probabilidad ya que la suma total de las probabilidades de los valores que toma es 1. Esto quiere decir en particular que nunca vamos a poder sortear al azar, de forma equiprobable, un número natural. En este capítulo vamos a introducir un nuevo tipo de variable aleatoria, que, no solo puede tomar infinitos valores, sino que nos permitirá definir qué quiere decir sortear al azar un número *entre* 0 y 1. Dos preguntas que surgen naturalmente son

1) ¿Puede la variable tomar todos los números entre 0 y 1 con *la misma probabilidad* y además que esta sea no nula?

2) ¿En caso de que 1) no sea posible, puede tomar todos los números entre 0 y 1 con probabilidad no nula (aunque no sea la misma)?

Vamos a intentar responder estas preguntas, empecemos por la primera que es la más fácil. Es claro que entre 0 y 1 tenemos infinitos números, por ejemplo $1/2, 1/3, 1/4, 1/5, 1/6, \ldots$. Al igual que lo que sucede con los números naturales, no podemos asignarle probabilidad positiva y *la misma* a dichos números. ¿Por qué? Supongamos que le asignamos a los infinitos números de la forma $1/n$ con $n = 2, 3, 4, 5, \ldots$ probabilidad $p > 0$, es decir tenemos una variable aleatoria X tal que $P(X = 1/n) = p$ para todo $n > 1$. Entonces

$$P\left(X = \frac{1}{2}\right) + P\left(X = \frac{1}{3}\right) + P\left(X = \frac{1}{4}\right) + \cdots + P\left(X = \frac{1}{n}\right) = n \times p,$$

como p es un número no nulo, fijo, $n \times p$ se puede hacer tan grande como se quiera al ir agregando sumandos (y por lo tanto mayor que 1). Con lo cual tiene que ser $p = 0$. Esto responde la pregunta 1), veamos la 2). La pregunta dos es mucho más difícil de responder y simplemente daremos una idea intuitiva de dónde está el

problema. Observemos que ahora, el problema no está en que sean infinitos números ya que una variable aleatoria con distribución geométrica toma infinitos valores $(1,2,3,\dots)$ y sin embargo pudimos asignarle probabilidad positiva a cada uno de ellos: $(1-p)^k \times p$, sino en que entre 0 y 1 hay *muchos más* valores que puede tomar. Esto tiene que ver con el hecho de que los números entre 0 y 1 no se pueden poner en una lista y contarlos, son *no numerables*, lo cual hace que la respuesta a la pregunta formulada en 2) sea negativa.

Volviendo al problema de sortear un número entre 0 y 1 de manera equiprobable, podemos decir que intuitivamente, si nuestro número a sortear puede ser cualquier número entre 0 y 1 y queremos que el sorteo sea *equiprobable* es razonable pensar que cualquier número entre 0 y $1/2$ tiene la misma probabilidad de salir que un número entre $1/2$ y 1. A su vez, la suma de las probabilidades de dichos intervalos es 1. Si bien no son disjuntos ($[0,1/2] \cap [1/2,1] = 1/2$) la probabilidad de que X tome el valor $1/2$ es 0, por lo que vimos antes (no puede tomar ningún valor con probabilidad positiva. Por lo tanto

$$P\Big(X \in \Big[0,\frac{1}{2}\Big]\Big) = \frac{1}{2} = P\Big(X \in \Big[\frac{1}{2},1\Big]\Big).$$

Razonando de la misma manera

$$P\Big(X \in \Big[0,\frac{1}{4}\Big]\Big) = \frac{1}{4} = P\Big(X \in \Big[\frac{1}{4},\frac{1}{2}\Big]\Big) = P\Big(X \in \Big[\frac{1}{2},\frac{3}{4}\Big]\Big) = P\Big(X \in \Big[\frac{3}{4},1\Big]\Big),$$

y lo mismo podemos hacer para los n intervalos $[0,1/n],[1/n,2/n],\dots,[(n-1)/n,1]$ y concluir que tienen que tener probabilidad $1/n$. De esta manera vemos que la probabilidad de que la variable pertenezca a un intervalo $[a,b]$ depende *únicamente* de la longitud del intervalo: $b-a$. Consideremos la función f constante e igual a 1, definida del intervalo $[0,1]$ en los reales. Por lo que acabamos de ver, $P(X \in [a,b])$ no es otra cosa que *el área* de la función f entre los puntos a y b. Ver figura (4.1).

Definición 4.1. Una variable aleatoria tiene **distribución uniforme** en $[0,1]$ (que denotaremos $X \sim \mathrm{U}([0,1])$ si para todo $0 \le x \le y \le 1$ tenemos que

$$P\big(X \in [x,y]\big) = y - x.$$

Observemos que en este caso $P(X \in [x,y])$ coincide con área encerrada por el gráfico de la función constante $f = 1$ y el eje x y además, la probabilidad de que la variable pertenezca a un determinado intervalo depende *únicamente* de la longitud del intervalo.

Vamos a definir formalmente que quiere decir que una variable X tenga distribución uniforme en $[a,b]$, con $a < b$:

Definición 4.2. Una variable aleatoria tiene **distribución uniforme** en $[a,b]$ (que denotaremos $X \sim \mathrm{U}([a,b])$ si para todo $a \le x \le y \le b$ tenemos que

$$P\big(X \in [x,y]\big) = \frac{y-x}{b-a}. \tag{4.1}$$

Observemos que en este caso $P(X \in [x,y])$ coincide con área encerrada por el gráfico de la función constante $f = 1/(b-a)$ y el eje x. Nuevamente la probabilidad de que la variable pertenezca a un determinado intervalo depende *únicamente* de la longitud del intervalo y como caso particular, cuando $a = 0$ y $b = 1$, tenemos la distribución uniforme en $[0,1]$.

```r
plot.new()
sq=seq(from=0,to=1,by=.1)
plot(sq,dunif(sq),type="l",ylim=c(0,1.1),ylab="",xlab="")
rect(0.1,0,.2,1,col="blue")
rect(0.4,0,.5,1,col="blue")
```

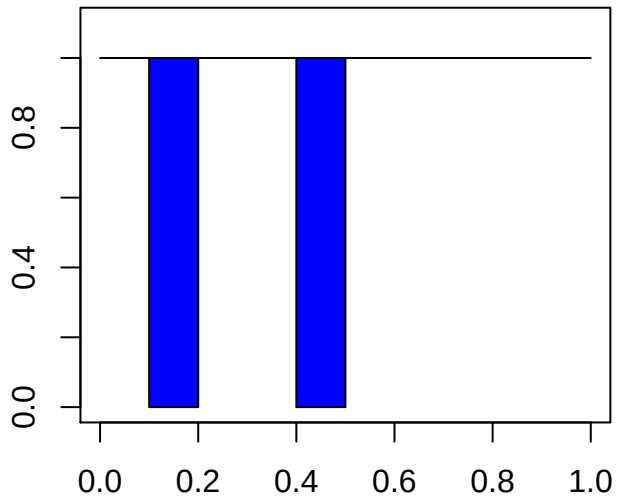


Figure 4.1: Gráfico de la densidad de una variable con distribución uniforme en $[0,1]$. Las áreas en azul son iguales y son la probabilidad de que una variable con distribución uniforme en $[0,1]$ pertenezca al intervalo $[0,1;0.2]$ para el primer rectángulo y al $[0.4;0.5]$ para el segundo.

Ejercicio 4.3. Hemos definido la variable aleatoria uniforme en un intervalo fijo $[a,b]$ ¿Se puede definir una variable aleatoria con distribución uniforme en $\mathbb{R}$?

En R

Para sortear n datos con distribución uniforme en $[a,b]$ (por defecto toma $a=0, b=1$) el comando es runif(n,a,b).

```
runif(5)

## [1] 0.5736179 0.8471996 0.2973667 0.5150698 0.8964859

runif(7,-1,1)

## [1] -0.28553989  0.84670100 -0.01165732  0.74440703 -0.91884933 -0.37446990
## [7] -0.94069110
```

4.2 Distribución de una variable aleatoria

La función que para cada $x \in \mathbb{R}$ nos da la probabilidad de que la variable sea menor o igual que x (es decir $P(X \leq x)$) se llama función de distribución, y usualmente se denota como F_X.

A modo de ejemplo supongamos que tenemos una variable aleatoria X que toma únicamente el valor 0, es decir $P(X=0)=1$. Veamos como es su función de distribución. Si razonamos como antes, vemos que $F_X(x)=0$ para todo $x<0$ mientras que $F_X(x)=1$ si $x>1$. No obstante $F_X(0)=1$ ya que (nuevamente usando (1.3)), $F_X(0)=P(X<0)+P(X=0)=0+1=1$. Esta función no es continua en $x=0$. Se deja como ejercicio verificar que $P(X \in (0,1))=1$, lo cual *no* coincide con $F_X(1)-F_X(0)=0$, y graficar F_X en este caso.

Supongamso que tenemos ahora $X \sim \text{Ber}(p)$. En este caso la variable toma los valores 0 con probabilidad $1-p$ y 1 con probabilidad p. Por lo tanto si $x<0$, $F_X(x)=P(X\leq x)=0$ ya que X no toma valores negativos. Por otra parte, para todo $x \geq 1$, $F_X(x)=P(X\leq x)=P(X=0)+P(X=1)=(1-p)+p=1$. Finalmente, si $x \in [0,1)$ $F_X(x)=P(X=0)=1-p$. El gráfico de esta función para $p=0.2$ se muestra en 4.2

Supongamos que tenemos X con distribución uniforme en $[0,1]$, vamos a calcular $F_X(x)$ para todo x. Observemos que, al definir la distribución uniforme en $[a,b]$ con $a<b$ (ver Definición 4.2) dimos la probabilidad de que la variable pertenezca a un intervalo $[x,y]$ para todo $a \leq x \leq y \leq b$. En nuestro caso la variable aleatoria sólo toma valores en el intervalo $[0,1]$ (estamos tomando $a=0$ y $b=1$), es decir $P(X \in [0,1])=1$ y por lo tanto $P(X \notin [0,1])=0$. Si tenemos un $x<0$ es claro que $F_X(x)=P(X\leq x)=0$, esto se deduce usando (1.3). De igual forma se deduce que $F_X(x)=1$ para todo $x \geq 1$. Finalmente si $x \in [0,1]$, si usamos ahora la fórmula (4.1) obtenemos que $F_X(x)=P(X \in [0,x])=x$. Se deja como ejercicio deducir la fórmula general de

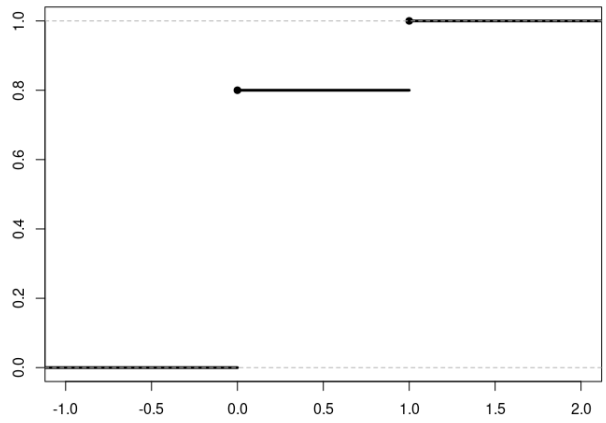


Figure 4.2: Gráfico de la función de distribución de una variable $X \sim \text{Ber}(0.2)$.

$F_X(x)$ cuando $X \sim \text{U}([a,b])$.

La función de distribución tiene algunas propiedades: es claro que para todo x, $0 \leq F_X(x) \leq 1$ (ya que $0 \leq P(X \leq x) \leq 1$ para todo x). Por otra parte, los límites a $-\infty$ y $+\infty$ son 0 y 1 respectivamente, es decir $F_X(x) \to 0$ si $x \to -\infty$, y $F_X(x) \to 1$ si $x \to +\infty$. Una propiedad fácil de ver es que F_X es no decreciente, lo cual significa que si $x < x'$ entonces $F_X(x) \leq F_X(x')$. Esto se sigue simplemente de que $F_X(x') = F_X(x) + P(X \in (x,x']) \geq F_X(x)$, donde en la primera igualdad hemos usado (1.3) y en la segunda desigualdad que $P(X \in (x,x']) \geq 0$. Por otra parte, se puede ver que $F_X(x)$ es *continua por derecha*, lo cual significa que $F_X(t_n) \to F_X(x)$ si $t_n \to x$ pero por derecha, es decir $t_n \geq x$ para todo $n > n_0$ para algún valor n_0. Esto no implica que la función F_X sea continua en todo punto ya que podría pasar que $F_X(t_n)$ tiende, cuando t_n tiende por izquierda (es decir por valores menores que x), a un valor estrictamente menor que $F_X(x)$. Observemos que de (1.3) se deduce que por ejemplo, $P(X \in (a,b])$ no es otra cosa que $P(X \in (-\infty,b]) - P(X \in (-\infty,a])$ y esto es, por definición, $F_X(b) - F_X(a)$. Si la función F_X es continua en a, se puede ver que $P(X \in [a,b])$ coincide también con $F_X(b) - F_X(a)$ (si no lo es, hay que sutituir $F_X(a)$ por el límite por izquierda de $F_X(x)$ cuando x tiende a a). Esto nos muestra que si conocemos la función de distribución podemos saber la probabilidad de que la variable pertenezca a cualquier intervalo $[a,b]$, para todo $a < b$.

4.3 Densidad asociada a una variable aleatoria

Una función $f \geq 0$ que tiene la propiedad de que para todo $a,b \in \mathbb{R}$, $P(X \in [a,b])$ es igual al área que encierra la gráfica de f y el eje de las x, en el intervalo $[a,b]$ se llama *función de densidad*. Para aquellos que hayan

visto cálculo integral, esto se puede expresar como

$$P\big(X \in [a,b]\big) = \int_a^b f(x)dx. \tag{4.2}$$

En particular si $a = b$,

$$P(X = a) = \int_a^a f(x)dx = 0. \tag{4.3}$$

Es decir, si una variable aleatoria tiene densidad, la probabilidad de que tome un valor dado cualquiera es 0. De esto se deduce que si una variable aleatoria toma una cantidad finita de valores $\{x_1,\ldots,x_n\}$ como por ejemplo la binomial, no puede tener densidad ya que si tuviera, por (4.3), $P(X = x_i) = 0$ para todo $i = 1,\ldots,n$ pero esto no puede ser. Se puede probar que una variable que toma una cantidad numerable de valores (como la geométrica o la Poisson) tampoco puede tener densidad. Es importante aclarar que la función de distribución se puede definir siempre (como $F_X(x) = P(X \leq x)$), aunque la variable no tenga densidad.

Se puede demostrar que si una variable aleatoria tiene densidad, su función de distribución no sólo es continua por derecha, sino que también lo es por izquierda, con lo cual es continua en todo punto.

A veces, en lugar de definir cual es la variable, diremos cual es la densidad. Una propiedad importante de las densidad es que el área total que queda bajo el gráfico de la misma es 1 (¿por qué?).

En R

En general los comandos de R para obtener el valor de la densidad de una variable comienzan con la letra d seguido del nombre que usan en R para identificar la variable, por ejemplo, dunif(x,a,b) nos da el valor en x, de la densidad de la uniforme en $[a,b]$. ¿ Cuanto debería dar por ejemplo dunif(2,0,3)?

```
dunif(2,0,3)
```

```
## [1] 0.3333333
```

```
dnorm(0,1,2)
```

```
## [1] 0.1760327
```

```
dbinom(3,6,1/3)
```

```
## [1] 0.2194787
```

Verificarlo en R. Si tenemos un conjunto de datos $X_1,\ldots,X_n$ y queremos hacer un gráfico de como es su densidad (aproximadamente) podemos usar plot(density(datos)) por ejemplo plot(density(runif(10000))) grafica algo que se parece a la densidad de la uniforme.

```
plot(density(runif(10000)),main="",xlab="Estimación de la uniforme")
lines(density(runif(1000)),col="red")
```

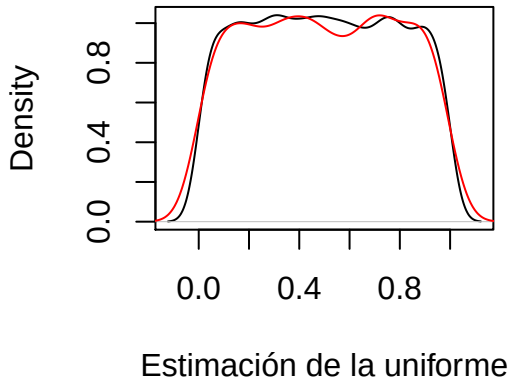


Figure 4.3: Estimación de la densidad uniforme con 10000 datos (en negro) y con 1000 datos (en rojo)

4.4 Distribución Normal

Veremos ahora una de las distribuciones más importantes del curso (y probablemente de la probabilidad toda), la **distribución normal** (o distribución gaussiana). La importancia de esta distribución radica, entre otras cosas, en un teorema que se denomina Teorema Central del Límite, el cual da la distribución a la que tiende el promedio de observaciones, más adelante enunciaremos este resultado con precisión ya que será de utilidad en el curso. Su origen data de 1733 y hasta la actualidad se siguen descubriendo propiedades y definiendo generalizaciones de la misma. Vamos a empezar con el caso más simple, la distribución normal de parámetros 0 y 1.

Definición 4.4. Decimos que una variable aleatoria X tiene distribución normal de parámetros 0 y 1 (que se denota $X \sim N(0,1)$ si su función de densidad es

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{4.4}$$

El gráfico de la función f se muestra en negro en la Figura (4.5).

Veamos algunas propiedades que se deducen de la *forma* de la f. Como se ve en (4.4) la función es simétrica respecto de 0 es decir $f(x) = f(-x)$ (esto se conoce como función par). Como f es par, si $0 < a < b$ entonces $P(X \in [a,b]) = P(X \in [-b,-a])$, además, también por la propiedad de simetría respecto de 0 se

```
plot(density(rnorm(10000)),main="",xlab="Estimación de la normal")
lines(density(rnorm(1000)),col="red")
```

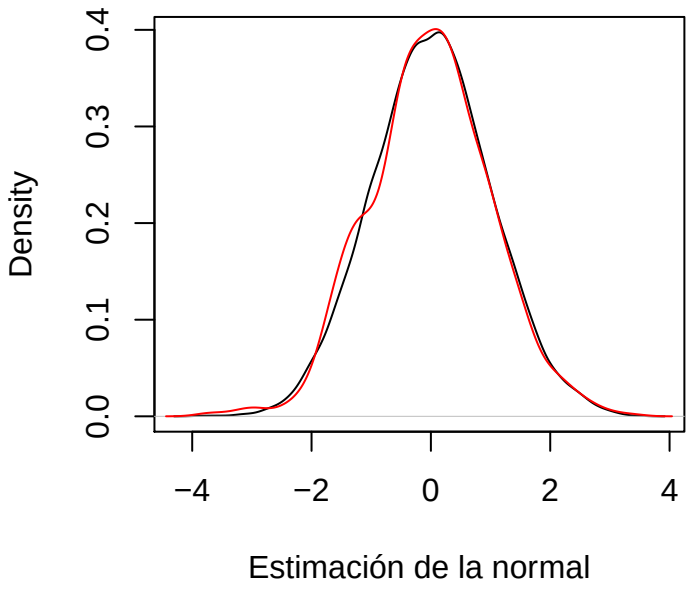


Figure 4.4: Estimación de la densidad normal con media 0 y varianza 1 con 10000 datos (en negro) y con 1000 datos (en rojo)

obtiene que

$$P(X > a) = P(X < -a), \tag{4.5}$$

en particular $P(X < 0) = P(X > 0) = 1/2$. En general la propiedad (4.5) va a ser cierta para cualquier variable cuya densidad sea par. Por otro lado es claro que el máximo de la función se da en 0 (y es $1/\sqrt{2\pi}$ como se sigue de (4.5)). De donde se deduce que para todo $-a < 0 < a$, $P(X \in [-a,a]) > P(X \in [l,s])$ para *cualquier* intervalo $[l,s]$ *que tenga longitud* $2a$ (ya que el área va a ser más pequeña). Si bien el área que encierra la gráfica de f con el eje x no se puede calcular de manera simple a partir de (4.4), (como hacíamos con la función de densidad de una variable con distribución uniforme) existen tablas que dan para ciertos valores de $t > 0$ el valor de $P(X < t)$. Este valor se puede calcular en R con `pnorm(t)`. En general se denota $\Phi(t) = P(X \leq t)$ con $X \sim N(0,1)$.

Observación 4.5. *Observemos que si* $a < b$, $\Phi(b) - \Phi(a) = P(X \in [a,b])$.

Ejercicio 4.6. Deducir, a partir de (4.5) que

$$\Phi(t) = 1 - \Phi(-t).$$

Esta propiedad es *muy* importante a la hora de usar una tabla de la distribución normal ya que como dijimos antes, dichas tablas solo dan los valores de $\Phi(t)$ para algunos $t > 0$. Dado que la función definida en (4.4) tiende a 0 muy rápido cuando x tiende a $+\infty$ o a $-\infty$ es que en general las tablas no dan el valor $\Phi(t)$ para valores mayores a 4 (en R verificar que $\Phi(4) = 0.999$ con el comando `pnorm(4)`).

El lector ya sospechará que el 0 en la definición de $N(0,1)$ tiene que ver con el hecho de que sea simétrica respecto de 0, efectivamente esto es así. Está relacionado al comportamiento que tienen los números que se *sortean* con distribución $N(0,1)$. En una muestra de muchos datos que siguen dicha distribución, y bajo el supuesto de que los datos son independientes unos de otros, vamos a obtener aproximadamente la misma cantidad de números positivos que negativos (no sólo la misma cantidad sino también de similar magnitud, pero de signo opuesto). Esto hace que si los promediamos de un valor próximo a 0. Por ahora no vamos a justificar mucho más ese parámetro, más adelante veremos que es lo que se denomina *valor esperado*. La relación del 1 en $N(0,1)$ con el comportamiento de los datos es un poco más difícil de explicar a partir de (4.4), tiene que ver con qué tanto se alejan de 0 esos números sorteados con distribución normal. Si en lugar de 1 ponemos 1 millón, vamos a ver que los números se hacen mucho más grandes en magnitud. En virtud de esto es que vamos a definir:

Definición 4.7. Una variable aleatoria X tiene distribución normal con media μ y varianza $\sigma^2 > 0$, que denotamos $N(\mu, \sigma^2)$, si su densidad es

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{4.6}$$

El gráfico de esta función es idéntico al que se muestra en la Figura (4.5) salvo por el hecho de que el punto de simetría es μ. Esto hace que promediar *muchos* datos independientes, con distribución $N(\mu, \sigma^2)$ da un valor próximo a μ. Se puede demostrar, y será lo que se usará para calcular probabilidades de una variable con distribucción $N(\mu, \sigma^2)$ el siguiente resultado:

Teorema 4.8. *Si X tiene distribución $N(0,1)$ entonces, si $\sigma > 0$,*

$$\sigma X + \mu \sim N(\mu, \sigma^2).$$

Proof. Llamemos $Z = \sigma X + \mu$ observar que $X = \frac{1}{\sigma}(Z - \mu)$ (aquí hemos usado que $\sigma > 0$), tenemos que probar que Z tiende densidad dada por (4.6). Para eso calculamos

$$P(Z \le t) = P(Z - \mu \le t - \mu) = P\Big(\frac{1}{\sigma}(Z-\mu) \le \frac{1}{\sigma}(t-\mu)\Big) = P\Big(X \le \frac{1}{\sigma}(t-\mu)\Big).$$

Como X tiene distribución normal con esperanza 0 y varianza 1, si usamos (4.4), la probabilidad anterior se

```
plot.new()
sq=seq(from=-4,to=4,by=.1)
plot(sq,dnorm(sq,mean=0,sd=1),type="l",ylim=c(0,.51),ylab="",xlab="")
points(sq,dnorm(sq,mean=0,sd=1.5),type="l",col="red")
points(sq,dnorm(sq,mean=0,sd=0.8),type="l",col="yellow")
points(sq,dnorm(sq,mean=0,sd=1.8),type="l",col="blue")
```

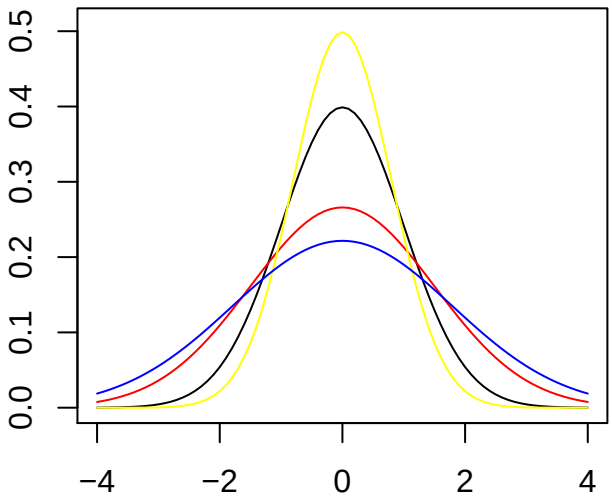


Figure 4.5: Gráfico de la densidad de una variable con distribución Normal con media 0 y diferentes valores del desvío σ.

puede calcular como

$$\int_{-\infty}^{\frac{t-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}x^2\Big)dx,$$

ahora hacemos el cambio de variable $x = (y-\mu)/\sigma$ [1] tenemos que $dx = (1/\sigma)dy$, y además $y = \sigma x + \mu$ por lo tanto los nuevos límites de integración son $-\infty$ (ya que $\sigma > 0$) y t, y nos queda,

$$\int_{-\infty}^{\frac{t-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}x^2\Big)dx = \int_{-\infty}^{t} \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\Big[\frac{y-\mu}{\sigma}\Big]^2\Big)dy$$

con lo cual llegamos a que la densidad de Z esta dada por (4.6) y esto concluye la demostración. □

De ésto último se sigue que, si $Z \sim N(\mu, \sigma^2)$ entonces

$$\Phi(t) = P\Big(\frac{Z-\mu}{\sigma} \le t\Big) = P\big(Z \le \sigma t + \mu\big).$$

Si tomamos ahora $t = \frac{1}{\sigma}(u-\mu)$ tenemos que

$$\Phi\Big(\frac{u-\mu}{\sigma}\Big) = P\big(Z \le u\big).$$

Ejercicio 4.9. Sea $X \sim N(1,4)$. Calcular $P(X < 2)$. Usando la igualdad anterior con $u = 1$ tenemos que $P(X < 2) = \Phi\Big(\frac{2-1}{2}\Big) = \Phi(1/2)$ y, usando la tabla $\Phi(1/2) \approx 0.691$. En R esto se calcula como $\texttt{pnorm}(1/2)$ (recordar que por defecto R asume que estamos con una normal con media 0 y varianza 1). Si queremos calcular directamente el valor $P(X < 2)$ tenemos que hacer $\texttt{pnorm}(2,1,2)$

Observación 4.10. *La mediana de una variable X con distribución normal con media μ es μ. Esto se sigue de que, como la densidad de X es simétrica respecto de μ, $P(X \le \mu) = 1/2 = P(X \ge \mu)$. Por lo tanto $Q(1/2) = \mu$.*

En R

Para sortear n variables aleatorias con distribución normal con media μ y varianza σ^2 se escribe $\texttt{rnorm}(\texttt{n},\mu,\sigma)$. Observar que el tercer parámetro no es σ^2 sino σ. El comando para obtener el valor de la densidad en un punto x, como dijimos, es $\texttt{dnorm}(\texttt{x},\mu,\sigma)$ mientras que la función cuantil para un cierto $u \in (0,1)$ se calcula como $\texttt{qnorm}(\texttt{u},\mu,\sigma)$. La distribución se calcula como $\texttt{pnorm}(\texttt{x},\mu,\sigma)$ y por defecto $\texttt{pnorm}(\texttt{x})$=$\Phi(x)$.

```
rnorm(4,-2,4)

## [1]  3.0546368 -4.3970563 -1.8082659  0.3032558
```

[1] la fórmula del cambio de variable dice que $\int_a^b f(g(x))g'(x)dx = \int_{g(a)}^{g(b)} f(u)du$, en el ejemplo $g(x) = \sigma x + \mu$ y f esta dada por (4.6)

```
pnorm(0)

## [1] 0.5

pnorm(qnorm(0.975))

## [1] 0.975
```

4.5 Distribución Exponencial

Vamos a definir ahora una nueva variable aleatoria a partir de su densidad: la variable aleatoria con distribución exponencial de parámetro λ. Dicha variable se emplea típicamente para modelar el tiempo entre la llegada consecutiva de dos personas a una fila (y por lo tanto es una variable que toma únicamente valores positivos). Tiene algunas propiedades que la hacen importante en la teoría de probabilidad, una de ellas, quizás la mas conocida es la de *pérdida de memoria* que, a grandes rasgos dice que, haber esperado un tiempo h para que suceda un cierto fenómeno no nos aporta información de si este sucederá o no en tiempo $t+h$. Dicho de otra manera, el fenómeno *se olvida* de lo que pasó hasta h. En términos de probabilidad condicional esto se escribe como

$$P\big(X > t+h | X > h\big) = P(X > t). \tag{4.7}$$

Veamos cual es la definición formal

Definición 4.11. Decimos que una variable aleatoria X tiene **distribución exponencial** de parámetro $\lambda > 0$ (que denotamos $X \sim \text{Exp}(\lambda)$) si su densidad es

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{si } \ t > 0 \\ 0 & \text{en otro caso} \end{cases} \tag{4.8}$$

El gráfico de dicha densidad, para 3 valores de λ, se muestra en la Figura (4.6).

Observemos que $f(0) = \lambda$. El parámetro λ, al igual que el μ de la distribución normal, tiene que ver con el promedio de *muchos* datos independientes, que tienen distribución exponencial. Se puede demostrar, aunque no lo haremos ahora que el promedio de esos datos estará próximo a $1/\lambda$. Es decir cuanto más grande es λ más chico es el promedio. Esto coincide con el hecho, observable en la Figura (4.6) de que al aumentar λ la gráfica toma valores más grandes cerca de 0 y valores más chicos hacia infinito.

Observación 4.12. *Veamos que si* $X \sim \text{Exp}(\lambda)$ *entonces*

$$P(X < t) = \begin{cases} 1 - e^{-\lambda t} & si \ \ t \geq 0 \\ 0 & si \ t < 0 \end{cases} \tag{4.9}$$

```
plot.new()
sq=seq(from=0,to=4,by=.1)
plot(sq,dexp(sq,rate=.5),type="l",ylim=c(0,2),ylab="",xlab="")
points(sq,dexp(sq,rate=1),type="l",col="red")
points(sq,dexp(sq,rate=1.5),type="l",col="yellow")
points(sq,dexp(sq,rate=2),type="l",col="blue")
```

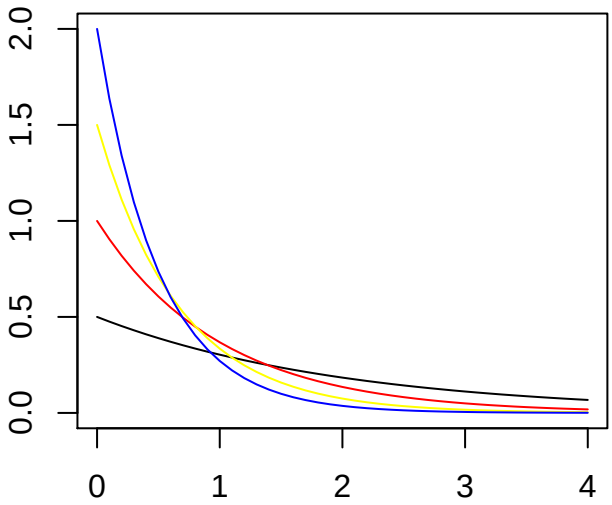


Figure 4.6: Gráficos de las densidades de la variable exponencial, para diferentes valores de λ.

Para ver eso observemos que, como la densidad de una variable con distribución exponencial es nula hasta 0, si $t < 0$ se cumple (4.9), y si $t \geq 0$,

$$P(X < t) = \int_0^t \lambda e^{-\lambda s} ds,$$

por la regla de Barrow lo que tenemos que hacer es hallar una primitiva de $\lambda e^{-\lambda s}$, evaluarla en t y en 0 y restar dichos valores. Es fácil ver que la función $-e^{-\lambda s}$ es primitiva (verificarlo derivando). Por lo tanto

$$P(X < t) = \int_0^t \lambda e^{-\lambda s} ds = -e^{-\lambda s}\Big|_0^t = -e^{-\lambda t} + 1 = 1 - e^{-\lambda t},$$

es decir se verifica (4.9).

Vamos a demostrar que se cumple (4.7). Observemos que de (4.9) tenemos que, para todo $h \geq 0$

$$P(X > h) = 1 - P(X < h) = e^{-\lambda h}.$$

Por definición

$$P(X > t + h | X > h) = \frac{P(\{X > t + h\} \cap \{X > h\})}{P(X > h)},$$

Observemos que $P(\{X > t + h\} \cap \{X > h\}) = P(X > t + h)$ ya que si pasó más de $t + h$, seguro que paso más de h. En términos de sucesos lo que estamos diciendo es que $\{X > t + h\} \subset \{X > h\}$ por lo tanto $\{X > t + h\} \cap \{X > h\} = \{X > t + h\}$. Si usamos eso tenemos que

$$P(X > t + h | X > h) = \frac{P(X > t + h)}{P(X > h)} = \frac{e^{-\lambda(t+h)}}{e^{-\lambda h}} = \frac{e^{-\lambda t} e^{-\lambda h}}{e^{-\lambda h}} = e^{-\lambda t} = P(X > t).$$

Observación 4.13. *Veamos que si $X \sim \text{Exp}(\lambda)$ entonces su función cuantil es $Q(u) = -\frac{1}{\lambda}\log(1-u)$, con $u \in (0,1)$. Por (4.9) sabemos que $F(x) = 1 - \exp(-\lambda x)$, por lo tanto si hacemos $u = F(x)$, $1 - u = \exp(-\lambda x)$ es decir $\log(1-u) = -\lambda x$, de donde $x = \frac{-1}{\lambda}\log(1-u)$, es decir $Q(u) = \frac{-1}{\lambda}\log(1-u)$. En particular la mediana de la exponencial es $Q(1/2) = \frac{1}{\lambda}\log(2)$.*

En R

Para obtener la densidad, la distribucion, la función cuantil o sortear datos con distribución exponencial tenemos que usar `dexp(x,λ)`, `pexp(q,λ)`, `qexp(u,λ)` y `rexp(n,λ)` respectivamente. Si no especificamos el valor λ por defecto toma $\lambda = 1$.

4.6 Distribución T de Student

Definiremos ahora una variable aleatoria que aparecerá en los próximos capítulos y es muy importante para hacer test de hipótesis.

Definición 4.14. Diremos que una variable aleatoria tiene **distribución** T **de Student con** $k > 0$ ***grados de libertad*** (que denotamos T_k) si su densidad es

$$f(x) = \frac{C(k)}{\left(1 + \frac{x^2}{k}\right)^{\frac{k+1}{2}}}$$

donde $C(k)$ es una constante (que depende de k) que hace que $f(x)$ sea una densidad, es decir que el área entre el eje $y = 0$ y la gráfica de f sea 1. Se puede dar una fórmula explícita para $C(k)$ que aquí no daremos ya que no haremos uso de ella.

Se deja como ejercicio verificar que f es simétrica respecto de 0. Si bien, a diferencia de lo que sucedía con (4.4), existen fórmulas expícitas para $P(T_k < x)$ es decir el área de f hasta x, lo que se hace en general es usar una tabla que da, para algunos valores de t y de k, los valores de $P(T_k < t)$ (esto se puede calcular en R con el comando `pt(t,k)`). Se puede demostrar que para valores de k grandes, $P(T_k < x) \approx P(N(0,1) < x)$ *para todo* x más aun

$$P(T_k < x) \to P(N(0,1) < x) \quad \text{si } k \text{ tiende a infinito.}$$

La Figura (4.7) muestra, para diferentes valores de k el gráfico de f.

Una propiedad que nos será de utilidad más adelante, que se deduce de que T_k es simétrica respecto de 0 es que, si llamamos $F_{T_k}(x)$ a la función que (al igual que Φ para la normal) nos da el área hasta x, es decir si $P(T_k \le x) = F_{T_k}(x)$ entonces, para todo x

$$F_{T_k}(x) = 1 - F_{T_k}(-x) \tag{4.10}$$

Se deja como ejercicio expresar la igualdad anterior en términos de la funcion cuantil Q.

En R

Supongamos que tenemos `k` grados de libertad, para obtener: la densidad en un punto `x` usamos `dt(x,k)`, la función de distribución en un punto `x` `pt(x,k)`, la función cuantil en `u` `qt(u,k)`, `n` realizaciones `rt(n,k)`.

4.7 Distribución *Chi-cuadrado:* χ^2_k

Finalmente, otra distribución que aparecerá, pero que no estudiaremos en profundidad, es la **distribución Chi-cuadrado con** k **grados de libertad**, que se denota usualmente χ^2_k.

Definición 4.15. Diremos que una variable X tiene distribución χ^2_k si su densidad es

$$f(x) = \begin{cases} C(k)x^{k/2-1}e^{-x/2} & \text{si } x > 0 \\ 0 & \text{si } x \le 0 \end{cases} \tag{4.11}$$

donde $C(k)$ es una constante (que depende de k) que hace que el área bajo la gráfica de f sea uno. (No es la misma constante que para la T de student pero también existen fórmulas explicitas para $C(k)$).

Observemos que la densidad f es cero si $x \leq 0$ (al igual que lo que sucedía con la distribución exponencial), es decir si una variable tiene distribución χ^2_k esta sólo toma valores positivos. La Figura (4.8) muestra el gráfico de f para diferentes valores de k.

En R

Nuevamente, si tenemos k grados de libertad, para obtener: la densidad en un punto x usamos dchisq(x,k), la función de distribución en un punto x pchisq(x,k), la función cuantil en u qchisq(u,k), n realizaciones rchisq(n,k).

```
sq=seq(from=-5,to=5,by=.1)
plot(sq,dt(sq,df=2),type="l",ylim=c(0,.45),ylab="",xlab="")
points(sq,dt(sq,df=1),type="l",col="red")
points(sq,dt(sq,df=5),type="l",col="yellow")
points(sq,dt(sq,df=10),type="l",col="blue")
points(sq,dt(sq,df=40),type="l",col="magenta")
```

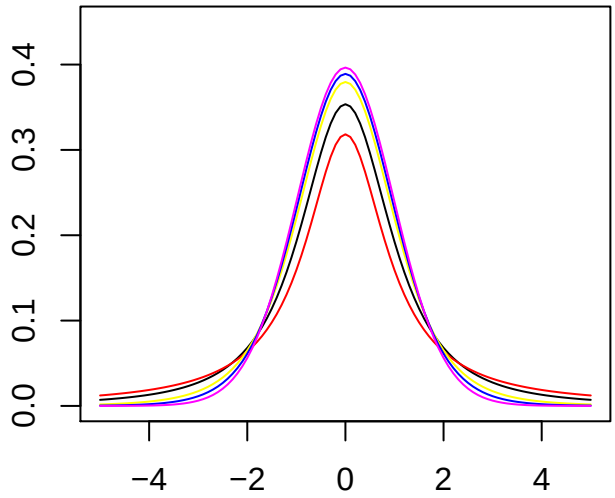


Figure 4.7: Gráficos de las densidades de una variable con distribución T_k para diferentes valores de k

```
plot.new()
sq=seq(from=0,to=8,by=.1)
plot(sq,dchisq(sq,df=2),type="l",ylim=c(0,1),ylab="",xlab="")
points(sq,dchisq(sq,df=1),type="l",col="red")
points(sq,dchisq(sq,df=3),type="l",col="yellow")
points(sq,dchisq(sq,df=4),type="l",col="blue")
points(sq,dchisq(sq,df=5),type="l",col="magenta")
```

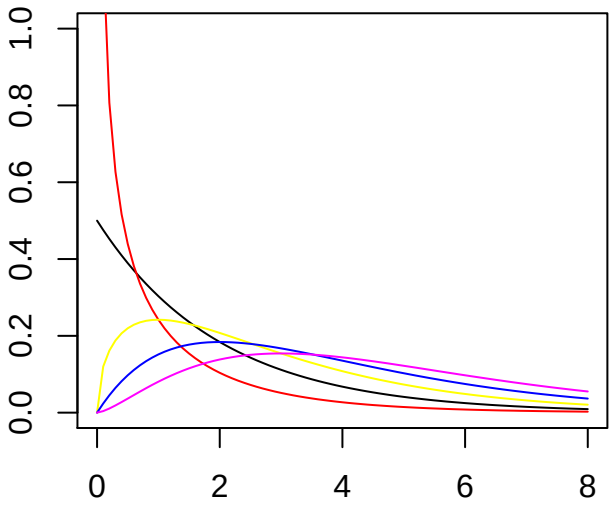


Figure 4.8: Gráficos de las densidades de una variable con distribución χ^2_k para diferentes valores de k.

Chapter 5

Esperanza y Varianza de una variable aleatoria

5.1 Esperanza

En esta sección vamos a definir dos valores asociados a una variable aleatoria que serán de importancia en el resto del curso. No daremos las definiciones formales sino simplemente la idea intuitiva de qué son, para qué se usan, y cómo se calculan para las variables aleatorias que hemos definido. Vamos a comenzar con el concepto de esperanza de una variable aleatoria. Al comienzo del apartado sobre probabilidad dijimos que, intuitivamente, la probabilidad de que se de un determinado resultado en un experimento, está asociado a lo que sucede cuando repetimos ese experimento muchas veces. Si tiramos una moneda mil veces es intuitivo que *aproximadamente* la mitad de las veces va a salir cara, y la otra mitad va a salir número, si representamos cada una de estas ocurrencias con un 1 cuando sale cara y 0 cuando sale número y promediamos la secuencia de ceros y unos que obtenemos vamos a tener que sumar cerca de 500 veces 0 y cerca de 500 veces uno, y dividir sobre el total de tiradas, es decir el promedio es *aproximadamente* $500/1000 = 1/2$. Consideremos la variable X que toma los valores 0 y 1 con probabilidad $1/2$, observemos que el promedio que calculamos es $0 \times P(X = 0) + 1 \times P(X = 1) = 1/2$. Esto se expresa diciendo que el *valor esperado* de X es $1/2$. El nombre valor esperado puede dar lugar a confusión, *no* quiere decir que si arrojamos una moneda (cuyos resultados son únicamente 0 (cara) y 1 (número) vamos a obtener $1/2$ (que en términos del experimento no representa nada) sino que el promedio se está próximo a $1/2$.

Veamos otro ejemplo: la probabilidad de que al tirar un dado salga 5 es $1/6$ porque intuitivamente, si tiramos un dado, digamos mil veces para fijar ideas, $1/6$ de las mismas va a salir 5 ($1000/6 \approx 166$ de veces). Veamos que pasa si cada vez que obtenemos un resultado, lo anotamos, y luego promediamos los mil resultados que obtuvimos, pudimos haber obtenido: $1,3,3,3,3,4,5,3,5,6,6,6,4,5,1,1,5,1,2,1,1$.... Como dijimos, el 5 sale aproximadamente $1/6$ de las veces, es decir tenemos que sumar 166 veces 5. Es razonable suponer que lo mismo va a pasar con el 1, con el 2, con el 3 con el 4 y con el 6. Es decir, vamos a tener que sumar 166 veces 1, 166 veces 2, 166 veces 3, 166 veces 4 y 166 veces 6. Por lo tanto el promedio que

queremos hacer va a dar aproximadamente:

$$\frac{1\times 166+2\times 166+3\times 166+4\times 166+5\times 166+6\times 166}{1000}\approx 3.5,$$

que es igual a:

$$1\times\frac{166}{1000}+2\times\frac{166}{1000}+3\times\frac{166}{1000}+4\times\frac{166}{1000}+5\times\frac{166}{1000}+6\times\frac{166}{1000} \tag{5.1}$$

a su vez, como $\frac{166}{1000}\approx 1/6$, (5.1) es una *aproximación* de

$$1\times\frac{1}{6}+2\times\frac{1}{6}+3\times\frac{1}{6}+4\times\frac{1}{6}+5\times\frac{1}{6}+6\times\frac{1}{6}.$$

Vamos a definir la variable aleatoria X que sea i si salió i al tirar el dado, es claro que X toma valores 1,2,3,4,5,6 con probabilidad $1/6$ cada uno. Lo que escribimos es

$$1\times P(X=1)+2\times P(X=2)+\cdots+6\times P(X=6). \tag{5.2}$$

Observemos que en (5.2) desapareció completamente el número mil, que era el número de veces que tirábamos el dado. Es decir el número que se calcula en (5.2) es una propiedad de la variable X, que, como vimos, esta relacionada con lo que pasa al promediar los resultados de experimentos *independientes* que siguen la distribución de X. Dicho número se llama valor esperado de X. Es claro que no todas las variables toman los valores 1 a 6 (una variable con distribución geométrica toma infinitos valores) ni los toman con la misma probabilidad. No obstante, el valor esperado se calcula de forma parecida a (5.2), de tal manera que represente el número al que se aproxima el promedio de los resultados de repetir el experimento de forma independiente. Esto se debe a que en el promedio de los resultados, cada número aparece tantas veces como la probabilidad de que aparezca, multiplicado por la cantidad de experimentos que se hacen. La tabla (5.1) tiene los valores esperados de las variables aleatorias que hemos presentado hasta ahora.

Distribución	Valor esperado
$\text{Bin}(n,p)$	np
$\text{Geo}(p)$	(1-p)/p
$\text{HipGeo}(d,n,N)$	$\frac{nd}{N}$
$\text{Poisson}(\lambda)$	λ
$\text{U}[a,b]$	$(a+b)/2$
$N(\mu,\sigma^2)$	μ
$\text{Exp}(\lambda)$	$1/\lambda$
T_k	0
χ^2_k	k

Table 5.1: Esperanza de algunas variables

En R

El comando que nos permite calcular promedios es mean, por ejemplo si hacemos mean(rnorm(100,1,2)) lo que hace es primero general 100 realizaciones de una variable aleatoria con distribución normal con media 1 y varianza 4 (recordar que el tercer parámetro es el desvío y no la varianza de la normal), y luego calcula el promedio. Esto debería dar cerca de 1. Observar que al correr distintas veces este comando, obtenemos valores distintos (pero próximos a 1). Verificar computacionalmente que mean(rnorm(1000,1,2)) esta aún mas próximo a 1 y si hacemos mean(rnorm(100000,1,2)) nos acercamos todavía mas!.

```
mean(rnorm(1000,1,2))

## [1] 1.124912

mean(rnorm(100000,1,2))

## [1] 0.9998967
```

5.1.1 Esperanza de $X \sim \mathbf{Bin}(n,p)$

Veamos ahora cómo calcular la esperanza de una variable con distribucción binomial. Para fijar ideas tomemos $n = 4$ y $p = 1/6$ y denotemos $X \sim \text{Bin}(4, 1/6)$. Recordemos que la distribución binomial cuenta la cantidad de éxitos que se tienen en la repetición n veces de un experimento cuya probabilidad de éxito es p. Si repetimos muchas veces la secuencia de 4 experimentos, el valor esperado será próximo a la cantidad de éxitos promedio que obtenemos. Al igual que como hicimos con los dados, el valor esperado de X que se denota $E(X)$ es

$$E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) + 2 \times P(X = 2) + 3 \times P(X = 3) + 4 \times P(X = 4),$$

Usando (3.2) con $n = 4$ y $p = 1/6$ tenemos que

$$E(X) = \binom{4}{1} \times \frac{1}{6} \times \left(\frac{5}{6}\right)^3 + \binom{4}{2} \times \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^2 + \binom{4}{3} \times \left(\frac{1}{6}\right)^3 \times \frac{5}{6} + \binom{4}{4} \times \left(\frac{1}{6}\right)^4 = 4/6$$

Observemos que nos dio exactamente np. En general si $X \sim \text{Bin}(n,p)$ se tiene que $E(X) = np$, por ejemplo si $p = 1$ el número *promedio* esperado de éxitos es n, lo cual es muy razonable dado que $p = P(\text{éxito})$. Si $p = 0$ da 0 que también es muy intuitivo. Veamos la demostración de esto. Para eso vamos a usar la fórmula del binomio de Newton:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \tag{5.3}$$

Sea $X \sim \text{Bin}(n,p)$, dijimos que la esperanza de una variable discreta esta dada por la suma de los valores que toma, multiplicado por la probabilidad de que tome esos valores, es decir,

$$E(X) = \sum_{k=0}^{n} k\binom{n}{k} p^k (1-p)^{n-k}$$

el primer sumando ($k = 0$) da cero, por lo tanto

$$\sum_{k=0}^{n} k\binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k\binom{n}{k} p^k (1-p)^{n-k},$$

si ahora usamos la fórmula para $\binom{n}{k}$ tenemos que

$$\sum_{k=1}^{n} k\binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

ahora usamos que $n! = n(n-1)!$ y que $k! = k(k-1)!$ y escribimos $p^k = pp^{k-1}$ por lo tanto

$$\sum_{k=1}^{n} k\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \sum_{k=1}^{n} np\frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k}$$

observemos que el factor np sale multiplicando para afuera de la sumatoria, y que $\frac{(n-1)!}{(k-1)!(n-k)!} = \binom{n-1}{k-1}$, por lo tanto obtuvimos

$$\sum_{k=1}^{n} np\frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} = np\sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k}.$$

Veamos que $\sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} = 1$, para eso vamos a usar (5.3).

$$\sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} = \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} = (p + (1-p))^{n-1} = 1.$$

5.1.2 Esperanza de $X \sim \mathbf{Geo}(p)$

No vamos a hacer la cuenta para calcular $E(X)$ para la distribución geométrica, que implica sumar los infinitos valores:

$$E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) + 2 \times P(X = 2) + 3 \times P(X = 3) + \cdots = \sum_{i=0}^{\infty} i(1-p)^i p$$

Se puede demostrar que dicho valor es $(1-p)/p$. Aquí también, valores grandes de p hacen que $E(X)$ sea chico, esto quiere decir que el tiempo que, *en promedio*, tenemos que esperar hasta que ocurra el primer éxito se hace menor a medida que hacemos más probable que este ocurra. Muy razonable, ¿no?. Por otro lado $1/p$ tiende a infinito si p tiende a 0. Lo cual significa que si es poco probable que se de el éxito, vamos a tener que esperar *en promedio* mucho hasta que se de por primera vez.

5.1.3 Esperanza de $X \sim$ Poisson(λ)

Sea X con distribución de Poisson de parámetro λ, veamos que su esperanza es λ, tenemos que calcular,

$$E(X) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda}$$

nuevamente el primer sumando es 0. Y por lo tanto la suma anterior puede arrancar en $k = 1$. Vamos a usar primero que $\lambda^k = \lambda \lambda^{k-1}$ y que $k! = k(k-1)!$, por lo tanto

$$E(X) = \sum_{k=1}^{\infty} \lambda \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}$$

observemos los factores λ y $e^{-\lambda}$ salen fuera de la sumatoria ya que no dependen de k, por lo tanto

$$\sum_{k=1}^{\infty} \lambda \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

se puede demostrar, aunque no lo haremos, que la sumatoria anterior es e^{λ} y por lo tanto $E(X) = \lambda$.

5.1.4 Esperanza de una variable aleatoria discreta: caso general

Definición 5.1. Vamos ahora a definir cómo se calcula la esperanza de una variable aleatoria que toma valores $x_1, x_2, \ldots, x_k, \ldots$ (una cantidad finita o infinita numerable). Supongamos que $P(X = x_i) = p_i > 0$ y que $p_1 + p_2 + p_3 + \cdots = 1$ es decir esos son todos los valores que toma. Se define la esperanza de X, que denotamos $E(X)$ como:

$$E(X) = x_1 p_1 + x_2 p_2 + x_3 p_3 + \cdots + x_k p_k + \ldots,$$

siempre que la suma (o serie en el caso de infinitos sumandos) anterior sea finita. Como pasaba con el dado, cuyo valor esperado era 3.5, el valor esperado de una variable no necesariamente es uno de los valores que toma. Tampoco es cierto que en general tenga que ser positivo ya que los x_i pueden ser negativos. No obstante, se puede demostrar que está entre el mínimo y el máximo de los valores que toma X (si dichos valores son finitos, si por ejemplo el mínimo es $-\infty$ y el máximo es un valor C, la esperanza estará entre $-\infty$ y C). Esto último es muy intuitivo si lo pensamos en términos de promediar las veces que salieron los x_i ya que el promedio nunca va a dar mayor que si siempre sumamos el mayor de los datos ni menos que lo que daría si sumamos el más chico.

5.1.5 Esperanza de una variable continua

En el caso de variables que tienen densidad, las fórmulas anteriores no se aplican ya que la variable, como vimos, toma cualquier número con probabilidad 0 y además toma una cantidad *no numerable* de valores. En ese caso se usa la igualdad

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx. \tag{5.4}$$

- $X \sim \mathrm{U}([a,b])$. Tenemos que calcular

$$E(X) = \frac{1}{b-a}\int_a^b x dx = \frac{1}{b-a}\frac{x^2}{2}\Big|_a^b = \frac{1}{2(b-a)}(b^2-a^2) = \frac{1}{2(b-a)}(b-a)(b+a) = \frac{a+b}{2}$$

- $X \sim \mathrm{Exp}(\lambda)$, entonces $E(X) = 1/\lambda$. Recordemos que si X tiene distribución exponencial su densidad es (4.8), por lo tanto

$$E(X) = \int_0^\infty \lambda x e^{-\lambda x} dx$$

si usamos la fórmula de integración por partes [1] (una primitiva de $e^{-\lambda x}$ es $\frac{-1}{\lambda}e^{-\lambda x}$),

$$\int_0^\infty \lambda x e^{-\lambda x} dx = \lambda x \frac{-1}{\lambda} e^{-\lambda x}\Big|_0^\infty - \int_0^\infty \lambda \frac{-1}{\lambda} e^{-\lambda x} dx.$$

Veamos el primer sumando

$$\lambda x \frac{-1}{\lambda} e^{-\lambda x}\Big|_0^\infty = -x e^{-\lambda x}\Big|_0^\infty = \lim_{x\to+\infty} -x e^{-\lambda x} = 0$$

el segundo término es

$$-\int_0^\infty \lambda \frac{-1}{\lambda} e^{-\lambda x} dx = \int_0^\infty e^{-\lambda x} dx = \frac{-1}{\lambda} e^{-\lambda x}\Big|_0^\infty = \frac{1}{\lambda}$$

Por lo tanto $E(X) = 1/\lambda$.

- $X \sim N(\mu, \sigma^2)$ entonces $E(X) = \mu$, para eso vamos a usar que si $g(x)$ es una función impar (es decir cumple $g(x) = -g(-x)$) su integral en toda la recta es 0). Tenemos que calcular

$$E(X) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} x \exp\Big[-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big] dx$$

Hacemos el cambio de variable $u = (x-\mu)/\sigma$, y por lo tanto $du = \frac{1}{\sigma}dx$, y $x = \sigma u + \mu$. Observar que los límites de integración no cambian ya que $\sigma > 0$,

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} x \exp\Big[-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big] dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} (\sigma u + \mu) \exp\Big(-\frac{1}{2}u^2\Big) du$$

si separamos la integral anterior en suma de dos integrales, obtenemos

$$E(X) = \sigma \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u \exp\Big(-\frac{1}{2}u^2\Big) du + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}u^2\Big) du$$

[1] $\int_a^b f(x)g'(x)dx = fg|_a^b - \int_a^b g(x)f'(x)dx$

La primera integral da 0 ya que (se deja como ejercicio verificarlo) la función $u\exp(-\frac{1}{2}u^2)$ es impar. Por otra parte, $\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$ es la densidad de la normal con media 0 y varianza 1 por lo tanto si la integramos en toda la recta da 1. Finalmente obtenemos que $E(X) = \mu$ como queríamos.

5.2 Varianza

Supongamos que tenemos mil datos independientes, $X_1, \ldots, X_{1000}$ que corresponden a observaciones de una variable aleatoria X que sabemos que tiene distribución normal con esperanza 10 y varianza σ^2 desconocida. Por lo que dijimos antes, si promediamos dichos datos nos da próximo a 10 ya que la esperanza de $X \sim N(10, \sigma^2)$ es 10. Pensemos que esos datos son resultados de mediciones de algo cuyo valor real sabemos que tiene que ser 10, con un aparato que tiene un cierto error al medir. Lo que queremos ahora es tener una idea de qué tanto se equivoca al medir el aparato. Es decir, que tan alejados de 10 son los valores que medimos. La varianza σ^2 nos da una idea de la *disperción* de los datos entorno al valor esperado. Cuanto más grande es la varianza, más dispersos son los datos, y, en nuestro ejemplo, más grande es el error que comete el aparato. En el caso extremo en que $\sigma^2 = 0$ es decir la varianza es nula, los datos van a ser todos iguales a 10, en cuyo caso el error es 0. La definición de la varianza se hace por medio de la esperanza de una nueva variable. Lo que hacemos es, dada la variable X restarle la esperanza, es decir tomar $X - 10$ (con lo cual los datos que antes tenían esperanza $E(X)$ ahora, *centrados*, tienen esperanza 0) y luego promediar las nuevas observaciones $X_1 - 10, X_2 - 10, \ldots, X_{1000} - 10$ pero de modo tal que los datos que son muy grandes *influyan más* en el promedio que los datos pequeños. Es decir, en lugar de promediar $X_1 - 10, \ldots, X_{1000} - 10$ que sabemos que nos va a dar próximo a 0, promediamos $(X_1 - 10)^2, (X_2 - 10)^2, \ldots, (X_{1000} - 10)^2$. Observemos que la función x^2 hace justamente lo que decíamos antes. Si una observación, supongamos la X_4 para fijar ideas, es tal que $X_4 - 10$ es muy grande (que quiere decir que ese dato se alejó mucho de la esperanza), al elevar al cuadrado da un número aún mayor. Mientras que si $X_4 - 10 < 1$, entonces $(X_4 - 10)^2 < (X_4 - 10)$. La definición formal es la siguiente:

Definición 5.2. Sea X una variable aleatoria tal que $E(X^2) < \infty$ entonces

$$\mathrm{Var}(X) = E\left[(X - E(X))^2\right]$$

Veamos cómo se calcula la varianza a partir de la fórmula que dimos en (5.1), para variables que toman una cantidad finita o numerable de valores:

Observación 5.3. *Vamos ahora a calcular la varianza de una variable aleatoria que toma valores* $x_1, x_2, \ldots, x_k, \ldots$ *(una cantidad finita o infinita numerable). Supongamos que* $P(X = x_i) = p_i > 0$ *y que* $p_1 + p_2 + p_3 + \cdots = 1$ *es decir esos son todos los valores que toma. Sabemos, por* (5.1) *que*

$$E(X) = x_1 p_1 + x_2 p_2 + x_3 p_3 + \cdots + x_k p_k + \ldots,$$

De la formula anterior tenemos que para calcular $E\left((X - E(X))^2\right)$ *tenemos que ver que valores toma* $(X - E(X))^2)$ *y con qué probabilidades, y luego usar dicha fórmula. Observemos que* $(X - E(X))^2$ *toma los valores* $(x_1 - E(X))^2$ *con probabilidad* p_1, $(x_2 - E(X))^2$ *con probabilidad* p_2, *etc,* $(x_k - E(X))^2$ *con*

probabilidad p_k, ... por lo tanto, usando la formula anterior

$$E\big((X-E(X))^2\big) = (x_1-E(X))^2 p_1 + (x_2-E(X))^2 p_2 + (x_3-E(X))^2 p_3 + \cdots + (x_k-E(X))^2 p_k + \ldots \quad (5.5)$$

La observación anterior nos da la forma de calcular, por ejemplo, al varianza para la binomial, la geométrica, la hipergeométrica y la Poisson. Para las variables que tienen densidad no es posible usar dicha fórmula (no obstante la definición (5.2) sigue siendo cierta), pero se calculan usando la fórmula

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} (x-E(X))^2 f(x)dx. \quad (5.6)$$

La tabla (5.2) contiene los resultados de las varianzas de las variables que hemos visto hasta ahora.

A modo de ejemplo supongamos que X tiene distribución de Bernoulli de parámetro p, en este caso X toma únicamente los valores 0 con probabilidad $1-p$ y 1 con probabilidad p. La esperanza de X, $E(X)$ es p, por lo tanto si usamos (5.5)

$$\mathrm{Var}(X) = (0-p)^2(1-p) + (1-p)^2 p = p(1-p)(p+1-p) = p(1-p).$$

Veamos ahora como calcular la varianza de la variable X que daba el resultado de tirar un dado, es decir la variable aleatoria X que sea i si salió i al tirar el dado, como dijimos toma valores 1,2,3,4,5,6 con probabilidad $1/6$ cada uno. y su esperanza es 3.5, por lo tanto su varianza se calcula, usando (5.5) como

$$\mathrm{Var}(X) = (1/6)[(1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2] \approx 2.9166.$$

- Si $X \sim \mathrm{U}([a,b])$,

$$\mathrm{Var}(X) = \frac{1}{b-a}\int_a^b \Big[x-\frac{a+b}{2}\Big]^2 dx = \frac{1}{3(b-a)}\Big[x-\frac{a+b}{2}\Big]\Big|_a^b = \frac{1}{3(b-a)}\Big(\Big[b-\frac{a+b}{2}\Big]^3 - \Big[a-\frac{a+b}{2}\Big]^3\Big)$$
$$= \frac{1}{3(b-a)}\Big(\Big[\frac{b-a}{2}\Big]^3 - \Big[\frac{a-b}{2}\Big]^3\Big) = \frac{1}{3(b-a)}\Big(\Big[\frac{b-a}{2}\Big]^3 + \Big[\frac{b-a}{2}\Big]^3\Big) = \frac{1}{3(b-a)}2\Big[\frac{b-a}{2}\Big]^3 = \frac{(b-a)^2}{12}$$

- Si $X \sim \mathrm{Exp}(\lambda)$,

$$\mathrm{Var}(X) = \int_0^\infty \Big[x-\frac{1}{\lambda}\Big]^2 \lambda\exp(-\lambda x)dx,$$

para eso vamos a usar la fórmula de integración por partes, integrando $\lambda\exp(-\lambda x)$ (cuya primitiva es $-\exp(-\lambda x)$ y derivando $(x-1/\lambda)^2$.

$$\int_0^\infty \Big[x-\frac{1}{\lambda}\Big]^2 \lambda\exp(-\lambda x)dx = -\Big[x-\frac{1}{\lambda}\Big]^2\exp(-\lambda x)\Big|_0^\infty - \int_0^\infty -2\Big[x-\frac{1}{\lambda}\Big]\lambda\exp(-\lambda x)dx$$

El primer sumando es

$$-\Big[x-\frac{1}{\lambda}\Big]^2\exp(-\lambda x)\Big|_0^\infty = \Big[\lim_{x\to\infty} -\Big[x-\frac{1}{\lambda}\Big]^2\exp(-\lambda x)\Big] + (1/\lambda)^2 = 1/\lambda^2,$$

el segundo término es 0, ya que estamos calculando $2E(X - E(X))$.

- Si $X \sim N(0,1)$, como $E(X) = 0$ nos queda

$$\text{Var}(X) = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}x^2\Big) dx,$$

vamos a usar ahora la fórmula de integración por partes [2] con $a = -\infty$ y $b = \infty$, donde tomamos $g'(x) = x\frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}x^2\Big)$ y $f(x) = x$. Observar que $f(x)g'(x)$ es exactamente lo que queremos integrar. Por lo tanto nos queda que

$$g(x) = -\frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}x^2\Big) \quad y \quad f'(x) = 1.$$

El término del medio en la fórmula de partes es

$$f(x)g(x)\Big|_{-\infty}^{+\infty} = -x\frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}x^2\Big)\Big|_{-\infty}^{+\infty},$$

por órdenes, se puede ver que ambos límites dan 0. Nos queda por integrar

$$-\int_{-\infty}^{\infty} f'(x)g(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}x^2\Big) dx.$$

Si observamos lo que nos quedó, no es otra cosa que la integral de la densidad de X en toda la recta, es decir esta última integral da 1 ya que es $P(X \in (-\infty, +\infty))$. Por lo tanto obtuvimos que $Var(X) = 1$. Más adelante veremos que si $Z \sim N(\mu, \sigma^2)$, $Var(Z) = \sigma^2$.

Distribución	Varianza
$\text{Bin}(n,p)$	$np(1-p)$
$\text{Geo}(p)$	$(1-p)/p^2$
$\text{HipGeo}(d,n,N)$	$\frac{nd(N-d)(N-n)}{N^2(N-1)}$
$\text{Poisson}(\lambda)$	λ
$\text{U}[a,b]$	$(b-a)^2/12$
$N(\mu,\sigma^2)$	σ^2
$\text{Exp}(\lambda)$	$1/\lambda^2$
T_k con $k > 2$	$k/(k-2)$
χ^2_k	$2k$

Table 5.2: Varianza de algunas variables

[2] $\int_a^b f(x)g'(x)dx = fg|_a^b - \int_a^b g(x)f'(x)dx$

Chapter 6

Ley fuerte de los Grandes Números y Teorema Central del Límite

6.1 Variables aleatorias independientes

En la sección sobre independencia de sucesos vimos qué quería decir que dos sucesos fuesen independientes. En esta sección definiremos qué quiere decir que dos variables X e Y sean independientes. La idea intuitiva de fondo es la misma que para sucesos, lo que decimos es que saber el resultado de una variable no nos aporta información para deducir el resultado que tendrá la otra. Por ejemplo, si X es una variable que dada una persona nos da la altura (es decir definida en $\Omega = \{\text{conjunto de todas las personas}\}$, e Y es una variable (definida en el mismo Ω) que dada la persona nos da el color de ojos (supongamos que los hemos numerado 1= verdes, 2 azul, etc) es razonable pensar que dichas variables son independientes en el sentido de que saber el color de ojos de una persona nada nos dice de la altura, y que saber su altura nada nos dice del color de ojos que tendrá. En términos matemáticos estamos diciendo, por ejemplo

$$P(1.60 < X < 1.75 | Y = 1) = P(1.60 < X < 1.75),$$

o, como vimos, (dado que $P(Y = 1) > 0$), considerando los sucesos $\{1.60 < X < 1.75\}$ y $\{Y = 1\}$ lo anterior es equivalente a decir que

$$P(\{1.60 < X < 1.75\} \cap \{Y = 1\}) = P(\{1.60 < X < 1.75\}) \times P(\{Y = 1\}).$$

y esto, por definición, no es otra cosa que decir que los sucesos $\{1.60 < X < 1.75\}$ y $\{Y = 1\}$ son independientes. En virtud del ejemplo anterior, vamos a introducir la siguiente definición

Definición 6.1. Dos variables X e Y son independientes si, para todo par de intervalos $[a,b]$ y $[c,d]$ se cumple que

$$P\big(\{X \in [a,b]\} \cap \{Y \in [c,d]\}\big) = P\big(\{X \in [a,b]\}\big) \times P(\{Y \in [c,d]\}),$$

es decir los sucesos $\{X \in [a,b]\}$ y $\{Y \in [c,d]\}$ son independientes, *para todo* a,b,c,d.

La definición anterior se generaliza de manera intuitiva a n variables $X_1, \ldots, X_n$. Muchas veces, cuando tenemos una muestra de n datos, vamos a hacer el supuesto de que son independientes. Es un supuesto *muy* fuerte pero sin el cual muchos razonamientos que haremos no son ciertos. En la siguiente sección vamos a ver algunas propiedades de las variables independientes.

Ejercicio 6.2. Verificar que si X es una variable aleatoria constante (es decir existe c tal que $P(X = c) = 1$), entonces X es independiente de cualquier otra variable Y.

6.1.1 Suma de variables aleatorias

Hasta ahora hemos hablado de sumar datos, promediarlos, etc. Vamos a introducir aquí un poco más de formalización respecto de lo que significan esos procedimientos. Supongamos que tiramos 3 veces un dado, y que éxito es que salga 5 en la cara superior. Contar la cantidad de éxitos nos lleva (o debería) a pensar en una variable binomial de parametros $n = 3$, $p = 1/6$. Sabemos que cada una de las 3 tiradas se hacen de forma independiente (porque es parte de los supuestos del experimento). Por otro lado, si representamos éxito con un 1, y fracaso con un 0, contar la cantidad de éxitos no es otra cosa que contar la cantidad de unos que que hay en las ternas de ceros y unos. Es decir, si ahora consideramos las tres variables: X_1 que vale 1 si salió éxito en el primer experimento y 0 en caso contrario, X_2 la variable que vale 1 si salió éxito en el segundo experimento y 0 en caso contrario X_3 la variable que vale 1 si salió éxito en el tercer experimento y 0 en caso contrario, es claro que dichas variables son independientes. Si llamamos X a la variable que cuenta la cantidad de éxitos, de la cual sabemos que $X \sim \text{Bin}(3, 1/6)$, tenemos que la cantidad de éxitos es $X_1 + X_2 + X_3$. Es decir

$$X = X_1 + X_2 + X_3.$$

Veamos algunos cálculos para convencernos de esto: $P(X = 3) = P(X_1 + X_2 + X_3 = 3)$ y esto se da únicamente si $X_1 = X_2 = X_3 = 1$. Usando la formula de $\text{Bin}(3, 1/6)$ tenemos que $P(X = 3) = (1/6)^3$.
Observemos que las variables X_1, X_2 y X_3 son independientes, en el sentido de la definición anterior ya que hicimos el supuesto de que los experimentos se realizaban de forma independiente. Por lo tanto

$$P(\{X_1 = 1\} \cap \{X_2 = 1\} \cap \{X_3 = 1\}) = P(\{X_1 = 1\}) \times P(\{X_2 = 1\}) \times P(\{X_3 = 1\}) \\ = 1/6 \times 1/6 \times 1/6,$$

es decir llegamos al mismo resultado. Para calcular $P(X = 2)$ tenemos que considerar *todas* las formas de sumar 2 a partir de las variables X_1, X_2 y X_3 que nos conduce a 3 casos, $((1,1,0), (1,0,1)$ y $(0,1,1))$. Queda como ejercicio hacer la cuenta de que también se cumple que $P(X = 2) = P(X_1 + X_2 + X_3 = 2)$. Lo que hicimos fue *sumar* variables, para obtener una nueva variable aleatoria (definida en el mismo Ω). Este procedimiento tiene interés en general. Asimismo a veces puede ser de interés considerar la variable aX que dados los resultados de X los multiplica por una constante a. Estas nuevas variables tienen algunas propiedades interesantes:

Teorema 6.3. • *Si X e Y son variables* independientes *entonces*

$$E(XY) = E(X)E(Y) \tag{6.1}$$

Es importante aclarar que el resultado anterior no *es cierto sin la hipótesis de independencia.*

- *Para cualquier variable X y cualquier número a,*

$$E(aX) = aE(X) \quad y \quad \text{Var}(aX) = a^2\text{Var}(X).$$

- *Para* cualquier *par de variables X, Y se tiene que*

$$E(X+Y) = E(X) + E(Y).$$

- *Si X e Y son independientes se verifica*

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$

Ejercicio 6.4. Se deja como ejercicio deducir que si X es una variable aleatoria constante (toma un valor c con $P(X = c) = 1$) entonces $E(X) = c$ y $Var(X) = 0$.

Ejemplo 6.5. Usando estas propiedades deducir de (5.2) que también vale $\text{Var}(X) = E(X^2) - (E(X))^2$. Otra consecuencia inmediata de estas propiedades, y de la propiedad (4.8) es que si $X \sim N(\mu, \sigma^2)$, su varianza es σ^2 (aquí se usa que $X = \sigma Z + \mu$ con $Z \sim N(0,1)$ y que ya probamos que $Var(Z) = 1$).

Observación 6.6. *Es importante aclarar que, incluso en el caso en que X e Y son independientes e igualmente distribuidas, su suma no necesariamente tiene la distribución común. Por ejemplo la suma de dos variables independientes X e Y, con distribución exponencial de parámetro λ tiene una distribución que se denomina distribución de Erlag (de parámetros 2 y λ) cuya densidad esta dada por $f(x) = (1/2)\lambda^2 x \exp(-\lambda x)$ si $x \geq 0$ y $f(x) = 0$ si $x < 0$. Si X e Y son independientes y uniformes en $[0,1]$ su suma tiene una distribución conocida como Irwin-Hall, cuya densidad $f(x)$ vale 0 salvo cuando $x \in [0,2]$ en cuyo caso vale $f(x) = x$ si $x \in [0,1]$ y $f(x) = 2 - x$ si $x \in [1,2]$ (se deja como ejercicio graficarla). Un caso muy particular es el de la distribución normal, en este caso si X tiene distribución con esperanza μ_X y varianza σ_X^2, y Y es otra variable, independiente de X, con distribución normal con esperanza μ_Y y varianza σ_Y^2, su suma si tiene distribución normal, y en este caso la esperanza de la suma es la suma de las esperanzas, es decir $\mu_X + \mu_Y$, e igualmente, la varianza de la suma es la suma de las varianzas (es decir $\sigma_X^2 + \sigma_Y^2$).*

6.2 Covarianza y coeficiente de correlación

Como vimos la hipótesis de independencia entre variables es muy importante. Dadas dos variables X e Y veremos que se puede definir una cantidad $\rho(X,Y)$ que se denomina correlación (y es un valor en el intervalo $[-1,1]$), de modo que vale 0 si las variables son independientes (podría ser 0 pero que no sean independientes), y es 1 o -1 si y sólo si si estamos en el caso de máxima dependencia, es decir una es una combinación lineal de la otra, o lo que es lo mismo, existe $a \neq 0$ y b tal que con probabilidad 1, $Y = aX + b$. Para eso vamos a definir primero la covarianza entre dos variables.

Definición 6.7. Dadas dos variables aleatorias X,Y, supongamos que existen sus esperanzas, es decir $E|X| < \infty$ y $E|Y| < \infty$, definimos su covarianza como

$$\operatorname{cov}(X,Y) = E\Big[\big(X - E(X)\big)\big(Y - E(Y)\big)\Big] \tag{6.2}$$

Es inmediato verificar que $\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y)$, por lo tanto de esta última fórmula junto con (6.1) se sigue que *si* X e Y son independientes *entonces* $\operatorname{cov}(X,Y) = 0$, es muy importante tener en cuenta que **el recíproco no es cierto en general**, es decir *que la covarianza sea 0* ***no implica*** *que las variables sean independientes.* Para ver esto último consideremos el siguiente ejemplo, sea X la variable que toma únicamente los valores $-2,-1,1,2$ con probabilidad $1/4$ (y por lo tanto $E(X) = 0$), y Y la variable aleatoria X^2, es inmediato verificar que X e Y no son independientes. No obstante $XY = X^3$ toma los valores $-8,-1,1,8$ con probabilidad $1/4$ y por lo tanto $E(XY) = 0$, de donde se sigue que $\operatorname{cov}(X,Y) = 0$. Algunas propiedades importantes que se deducen de la propia definición son

- $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{cov}(X,Y)$.
- $\operatorname{cov}(X,a) = 0$ para todo número real real a.
- $\operatorname{cov}(X,X) = \operatorname{Var}(X)$
- $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$
- $\operatorname{cov}(aX,bY) = ab\operatorname{cov}(X,Y)$ para todo a,b números reales.
- $\operatorname{cov}(X+a,Y+b) = \operatorname{cov}(X,Y)$

Definición 6.8. Dadas dos variables aleatorias X,Y con varianzas $0 < \sigma_X^2 < \infty$ y $0 < \sigma_Y^2 < \infty$ respectivamente, el coeficiente de correlación se define como

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{6.3}$$

Algunas propiedades importantes de la correlación son

- $-1 \le \rho(X,Y) \le 1$
- $|\rho(X,Y)| = 1$ si y sólo si existe $a \neq 0$ y b tal que con probabilidad 1, $Y = aX + b$.
- $\rho(aX,bY) = \rho(X,Y)$ para todo $a > 0, b > 0$. Esta propiedad dice que el coeficiente de correlación es invariante por cambios de escala en las variables.
- $\rho(X+a,Y+b) = \rho(X,Y)$.

6.3 Ley de los Grandes Números y Teorema Central del Límite

Vamos a enunciar un par de teoremas que serán de utilidad más adelante y cuya demostración excede el contenido del curso, el primero, que se conoce como Ley de los grandes números, justifica formalmente el hecho de que la frecuencia en que sucede un evento (por ejemplo que sale cierto número al tirar un dado), tiende bajo ciertas hipótesis a la probabilidad (teórica) de que dicho evento suceda. El segundo es el Teorema Central del Liímite. Si bien se pueden encontrar diversas formas con distintas hipótesis, el siguiente teorema se debe a Kolmogorov,

Teorema 6.9. ***(Kolmogorov)****. Sea $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas tal que $E|X_i| < \infty$, entonces*

$$\frac{X_1 + \ldots + X_n}{n} \to E(X_1) \tag{6.4}$$

Como son idénticamente distribuidas (es decir todas tienen la misma distribución) se puede demostrar que todas tienen la misma esperanza, es decir $E(X_1) = E(X_2) = \ldots = E(X_n) = \ldots$. En el resultado anterior hay un pasaje al límite, que viene dado por el símbolo $\to$, definir con precisión que quiere decir que la variable promedio $\frac{X_1+\ldots+X_n}{n}$ converge al número $E(X_1)$ requiere de definiciones que no daremos pero intuitivamente lo que estamos diciendo, si pensamos $X_1, X_2, \ldots, X_n$ como los resultados de un experimento que se repite n veces de forma independiente, es que promediar estos valores da para valores de n *grandes* un valor póximo a $E(X_1)$.
El grafico 6.3 muestra el promedio de n variables normales con media 0 y varianza 1, para diferentes valores de n, como se ve, dicho promedio tiende a 0, que es el valor esperado de la distribución normal

La equación (6.4) nos dice en particular que $\frac{X_1+\ldots+X_n}{n} - E(X_1) \to 0$, la pregunta que surge naturalmente es, a que velocidad, o dicho de otra manera, ¿ es posible encontrar $\alpha_n \to \infty$ una sucesión de números, de modo que compense la diferencia anterior, es decir,

$$\alpha_n\Big(\frac{X_1 + \ldots + X_n}{n} - E(X_1)\Big) \to Z$$

siendo Z alguna variable aleatoria?. La respuesta es que si las variables $X_1, \ldots, X_n$ son independientes e idénticamente distribuidas con varianza σ^2 finita se puede tomar $\alpha_n = \frac{\sqrt{n}}{\sigma}$ y Z como la Normal con esperanza 0 y varianza 1, como demuestra el siguiente Teorema.

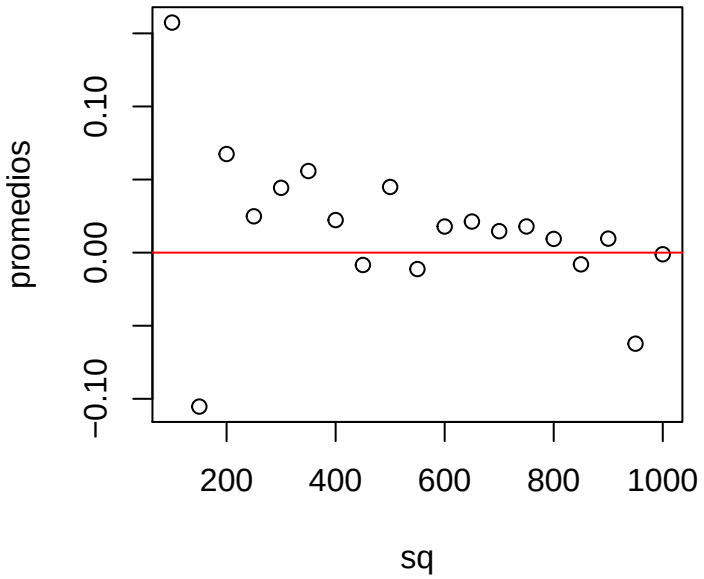
Teorema 6.10. ***(Teorema Central del Límite, P. Levy)*** *Sean $X_1, \ldots, X_n$ independientes e idénticamente distribuidas, $Var(X_1) = \sigma^2 < \infty$. Entonces para todo $a < b$ números reales,*

$$P\Big(a \le \frac{\sqrt{n}}{\sigma}\Big[\frac{X_1 + \ldots X_n}{n} - E(X_1)\Big] \le b\Big) \to \Phi(b) - \Phi(a),$$

donde Φ es la función de distribución de la Normal con esperanza 0 y varianza 1.

Lo que estamos diciendo es que la probabilidad de que el promedio *normalizado*, es decir una vez que le

```
plot.new()
sq=seq(from=100,to=1000,by=50)
promedios=c()
for(i in 1:length(sq)){promedios[i]=mean(rnorm(sq[i]))}
plot(sq,promedios)
abline(h=0,col="red")
```



restamos la esperanza, y multiplicamos por $\sqrt{n}/\sigma$, pertenezca a un cierto intervalo $[a,b]$ tiende, para valores de n grandes, al número $\Phi(b)-\Phi(a)$.

Part II

Estadística

Chapter 7

Estimación

7.1 Estimación de la esperanza y varianza de una variable aleatoria

En este capítulo veremos algunas aplicaciones de los conceptos introducidos en el capítulo anterior. En general se cuenta con datos $X_1, \ldots, X_n$ que resultan de realizar n veces un cierto experimento, y se quiere deducir propiedades de la distribución de los mismos: su valor esperado, su varianza, mediana, cuantiles, etc. Por ejemplo, si suponemos que los n datos son alturas de cierto grupo de personas, y si suponemos además que los mismos se distribuyen uniformemente en algún intervalo $[a,b]$ que desconocemos, nos puede interesar aproximar a o b o el valor esperado $(a+b)/2$. En otros casos los datos pueden provenir de mediciones de un fenómeno físico que suponemos sigue una distribución normal, cuya esperanza y varianza desconocemos, y nos interesa estimar dichos parámetros. Si se tienen dos muestras, una forma de tener una idea de si hay dependencia lineal entre ellas es estimar su correlación, esto lo veremos en la sección 7.3. Los antes mencionados son métodos de estimación paramétricos (ya que lo que se busca es estimar un parámetro). Para saber si los datos provienen o no de una determinada distribución F_0, tenemos que estimar la función de distribución de los datos, y por lo tanto aquí estamos ante un problema de estimación no-paramétrico (lo que queremos estimar es una función y no un parámetro). Veremos como estimar la función de distribución de los datos, por medio de la distribución empírica. La consistencia de estos estimadores (es decir, que cuando la cantidad de datos tiende a infinito, los estimadores convergen al valor verdadero del parámetro) se siguen en general de la ley de los grandes números.

7.2 Estimación de $E(X)$ y $\mathbf{Var}(X)$

Como dijimos antes, dados n datos $X_1, \ldots, X_n$ independientes e identicamente distribuidos segun una cierta variable X, si queremos estimar $E(X)$ es razonable considerar como *estimador* el promedio $\overline{X_n}$, dado por

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n}, \tag{7.1}$$

ya vimos que (ver Teorema 6.9)

$$\overline{X_n} \to E(X) \quad \text{cuando } n \text{ tiende a infinito.} \tag{7.2}$$

Dado que $\mathrm{Var}(X)$ lo definimos a partir de la esperanza de una nueva variables, es decir $Var(X) = E\big((X - E(X))^2\big)$ tenemos que, si conociéramos $E(X)$ un estimador para $\mathrm{Var}(X)$ es

$$\widehat{\mathrm{Var}}(X) = \frac{(X_1 - E(X))^2 + (X_2 - E(X))^2 + \cdots + (X_n - E(X))^2}{n},$$

en caso de no conocerlo podemos sustituir $E(X)$ por el estimador $\overline{X_n}$ obteniendo así el estimador de $\mathrm{Var}(X)$:

$$\widetilde{\mathrm{Var}}(X) = \frac{(X_1 - \overline{X_n})^2 + (X_2 - \overline{X_n})^2 + \cdots + (X_n - \overline{X_n})^2}{n}, \tag{7.3}$$

en este caso también es posible demostrar, bajo ciertas hipótesis que

$$\widetilde{\mathrm{Var}}(X) \to \mathrm{Var}(X) \quad \text{cuando } n \text{ tiede a infinito.}$$

En general, en lugar de estimar $\mathrm{Var}(X)$ a partir de $\widetilde{\mathrm{Var}}(X)$ usaremos el estimador de $\sqrt{\mathrm{Var}(X)}$

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2}$$

es decir $S_n^2 \to \mathrm{Var}(X)$. Observemos que la relación entre $\widetilde{\mathrm{Var}}(X)$ y S_n^2 es:

$$\widetilde{\mathrm{Var}}(X) = \frac{n-1}{n} S_n^2.$$

Observación 7.1. *Se deja como ejercicio, y es útil a veces para hacer cuentas, verificar que*

$$S_n^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\big(\overline{X_n}\big)^2\right].$$

Vamos a enunciar sin demostrar algunas propiedades de S_n que serán de utilidad para calcular intervalos de confianza.

Teorema 7.2. *Sean $X_1, \ldots, X_n$ independientes e idénticamente distribuidas con distribución $N(\mu, \sigma^2)$, entonces*

1) $\overline{X_n} \backsim N\left(\mu, \dfrac{\sigma^2}{n}\right)$.

2) $\overline{X_n}$ *y* S_n^2 *son independientes.*

3) $\dfrac{n-1}{\sigma^2} S_n^2 \backsim \chi^2_{n-1}$.

4) $\sqrt{n}\dfrac{(\overline{X_n}-\mu)}{S_n} \backsim T_{n-1}$.

En R

Como vimos antes, el promedio de datos se calcula con el comando mean, por otra parte el valor de S_n se calcula con sd, por ejemplo

```
sd(rnorm(20,0,2))

## [1] 1.816023

sd(rnorm(100000,0,2))

## [1] 1.99303
```

7.3 Estimación del coeficiente de correlación

Supongamos que tenemos datos $X_1,\ldots,X_n$ iid de X y $Y_1,\ldots,Y_n$ iid de Y, denotemos $\overline{X_n}$ al promedio de las X_i y $\overline{Y_n}$ al promedio de las Y_i, el coeficiente de correlación $\rho(X,Y)$ lo estimamos mediante

$$\hat{\rho}(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\overline{X_n})(Y_i-\overline{Y_n})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X_n})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y_n})^2}} \tag{7.4}$$

se puede demostrar, usando la ley de los grandes números, que $\hat{\rho}(X,Y) \to \rho(X,Y)$ cuando $n \to \infty$. En R se puede usar el comando cor para estimar la correlación mientras que cov estima la covarianza.

```
muestra1=rnorm(100,0,1)
muestra2=runif(100,0,1)
error=rnorm(100,0,1/4)
cor(muestra1,muestra2) #son independientes

## [1] -0.1999587

cor(muestra1,muestra1^2) #no hay dependencia lineal

## [1] -0.254071
```

```
cor(muestra1,3*muestra1+2+error) #hay dependencia lineal

## [1] 0.9963122
```

7.4 Estimación de la distribución $F_X(x)$ de una variable aleatoria X

Supongamos ahora que queremos estimar $F_X(x)$, a partir de una muestra de n datos $X_1,\dots,X_n$, independientes e identicamente distribuidos según una cierta variable X. Por definición $F_X(x)$ nos da la probabilidad de que la variable sea menor o igual que x. Si seguimos un modelo del tipo casos favorables sobre casos posibles, es intuitivo que un estimador de $F_X(x)$ es contar cuantos datos X_i son menores o iguales que x (casos favorables) y dividir este número entre n, es decir

$$\frac{\#\{X_i : X_i \le x\}}{n},$$

veamos por qué esto es razonable (es decir, se aproxima a $F_X(x)$ para valores de n *grandes*). Consideremos, para $j = 1,\dots,n$, las variables aleatorias auxiliares

$$Y_j = \begin{cases} 1 \text{ cuando } X_j \le x \\ 0 \text{ cuando } X_j > x \end{cases}$$

Esto se denota también $Y_j = \mathbb{I}_{\{X_j \le x\}}$. Observemos que

$$\frac{Y_1 + \cdots + Y_n}{n} = \frac{\#\{X_i : X_i \le x\}}{n}.$$

Estas variables Y_j tienen todas la misma distribución de Benoulli, con parámetro de éxito $p = P(X \le x) = F_X(x)$. Por otra parte su esperanza, como sabemos, también es p. Se puede ver además que son independientes, por lo tanto si usamos ahora (7.1) y (7.2) obtenemos que

$$\frac{Y_1 + Y_2 + \ldots + Y_n}{n} \to E(Y) = P(X \le x) = F_X(x).$$

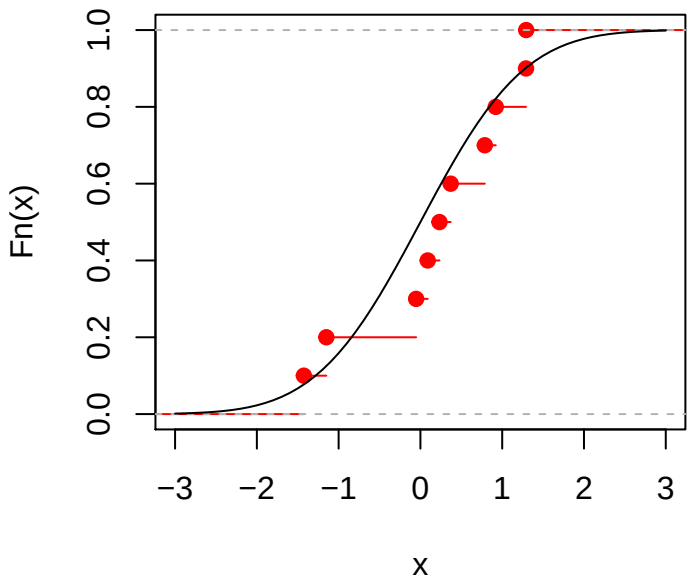
La función que para cada x toma el valor $\#\{X_i : X_i \le x\}/n$ se conoce como distribución empírica, a partir de la muestra $X_1,\dots,X_n$, y usualmente se denota como F_n. Supongamos que no hay dos valores repetídos, si ordeamos los datos de menor a mayor y denotamos $X_{(1)}$ al menor, $X_{(2)}$ el siguiente, hasta $X_{(n)}$, es decir

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)},$$

de la propia fórmula $F_n(X_{(i)}) = \#\{X_j : X_j \le X_{(i)}\}/n$ se sigue que $F_n(X_{(i)}) = i/n$ para todo $i = 1,\dots,n$, ya que $X_{(i)}$ tiene i datos menores o iguales que el. Por otra parte si, por ejemplo $x \in [X_{(j)}, X_{(j+1)})$ para algún j, tambien de la fórmula deducimos que $F_n(x) = j/n$. Finalmente si $x < X_{(1)}$, $F_n(x) = 0$ y si $x \ge X_{(n)}$, $F_n(x) = 1$.

En la figura 7.4 se grafica, en rojo, la función F_n para el caso en que X tiene distribución normal con media 0 y varianza 1 y tenemos 10 datos. En negro se grafica la distribución de la normal con media 0 y varianza 1.

```
plot.new()
sq=seq(from=-3,to=3,by=.1)
x=rnorm(10,0,1)
teorica=pnorm(sq)
plot(ecdf(x),main="",col="red",xlim=c(-3,3))
points(sq,teorica,type="l",xlab="",ylab="")
```



7.5 Método de los Momentos y Máxima verosimilitud

En general la distribución de las variables aleatorias que vimos dependen de ciertos parámetros, por ejemplo la distribución la exponencial depende del parámetro λ, la distribución normal depende de la media μ y la varianza σ^2. Otro ejemplo es el caso de la uniforme en un intervalo $[a,b]$, aqui a y b son parámetros que podemos querer estimar a partir de una muestra. Muchas veces sabemos que los datos que tenemos son normales pero no conocemos estos parámetros o son uniformes pero no conocemos a y/o b. Para estimar estos parámetros veremos dos métodos clásicos, presentaremos el método y veremos algunos ejemplos. No

demostraremos las condiciones bajo las cuales los estimadores convergen a los valores verdaderos.

7.5.1 Método de los momentos

El método de los momentos permite estimar los parámetros siempre y cuando sean finitos los *momentos* de la variable aleatoria. Dada una variable aleatoria X y un número natural $k > 0$, el k-ésimo momento de X es $E(X^k)$ y decimos que tiene momento k si $E(X^k) < \infty$. Se puede demostrar que si una variable aleatoria tiene momento k, también son finitos los momentos $1, \ldots, k-1$. El método de los momentos plantea el siguiente sistema de ecuaciones:

$$\begin{cases} E(X) &= \overline{X_n} \\ E(X^2) &= \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2 \\ \vdots & \vdots \\ E(X^k) &= \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^k \end{cases}$$

Los $E(X^k)$ se llaman momentos poblacionales y las expresiones al otro lado de la igualdad, momentos muestrales. Los parámetros a estimar aparecen en el cálculo de los momentos poblacionales. El sistema de ecuaciones anterior no necesariamente es un sistema lineal pero en algunos casos se puede resolver y se despejan los parámetros. Veamos algunos ejemplos

Ejemplo 7.3. Comencemos con un ejemplo simple supongamos que X tiene distribución normal con media μ y varianza $\sigma^2 > 0$. Recordemos que $\mathrm{Var}(X) = E(X^2) - (E(X))^2$. Por lo tanto despejamos el segundo momento poblacional como $E(X^2) = \sigma^2 + \mu^2$. Denotemos $\frac{1}{n}\sum_{i=1}^{n} X_i^k = \overline{X_n^2}$. Por lo tanto el sistema queda

$$\begin{cases} \mu &= \overline{X_n} \\ \sigma^2 + \mu^2 &= \overline{X_n^2} \end{cases}$$

Si despejamos en la segunda ecuación, $\sigma^2 = \overline{X_n^2} - (\overline{X_n})^2$

Ejemplo 7.4. Veamos como queda el sistema para el caso en que X tiene distribución uniforme en $[a,b]$ y queremos despejar a y b. Sabemos que $E(X) = (a+b)/2$ y $\mathrm{Var}(X) = (b-a)^2/12$. Recordemos que $\mathrm{Var}(X) = E(X^2) - (E(X))^2$. Por lo tanto despejamos el segundo momento poblacional como $E(X^2) = (b-a)^2/12 + (a+b)^2/4$. Denotemos $\frac{1}{n}\sum_{i=1}^{n} X_i^k = \overline{X_n^2}$. Por lo tanto el sistema queda

$$\begin{cases} (a+b)/2 &= \overline{X_n} \\ (b-a)^2/12 + (a+b)^2/4 &= \overline{X_n^2} \end{cases}$$

Si usamos la primer ecuación en la segunda obtenemos

$$\frac{(b-a)^2}{12} = \overline{X_n^2} - (\overline{X_n})^2$$

es decir

$$(b-a) = \pm 2\sqrt{3\left(\overline{X_n^2} - (\overline{X_n})^2\right)}$$

si escribimos la primer ecuación como $a+b = 2\overline{X_n}$ y le sumamos la ecuación anterior obtenemos que

$$b = \overline{X_n} \pm \sqrt{3\left(\overline{X_n^2} - (\overline{X_n})^2\right)}$$

y por lo tanto usando que $a = 2\overline{X_n} - b$ tenemos que

$$a = 2\overline{X_n} - \overline{X_n} \pm \sqrt{3\left(\overline{X_n^2} - (\overline{X_n})^2\right)} = \overline{X_n} \pm \sqrt{3(\overline{X_n^2} - (\overline{X_n})^2)}$$

como $a < b$ tenemos únicamente 2 soluciones posibles

$$a = \overline{X_n} - \sqrt{3\left(\overline{X_n^2} - (\overline{X_n})^2\right)}$$

y

$$b = \overline{X_n} + \sqrt{3\left(\overline{X_n^2} - (\overline{X_n})^2\right)}.$$

Ejemplo 7.5. Para el primer ejemplo vamos a introducir la distribución Γ de parametros k y λ. Una variable X tiene distribución $\Gamma(k,\lambda)$ si es suma de k variables exponenciales de parámetro λ, independientes. Por lo tanto por la linealidad de la esperanza $E(X) = k\lambda$ y $\mathrm{Var}(X) = k/\lambda^2$. Recordemos que $\mathrm{Var}(X) = E(X^2) - (E(X))^2$. Por lo tanto despejamos el segundo momento poblacional $E(X^2) = k/\lambda^2 + k^2/\lambda^2$. Denotemos $\frac{1}{n}\sum_{i=1}^n X_i^k = \overline{X_n^2}$. Por lo tanto el sistema queda

$$\begin{cases} \frac{k}{\lambda} &= \overline{X_n} \\ \frac{k}{\lambda^2} + \frac{k^2}{\lambda^2} &= \overline{X_n^2} \end{cases}$$

De la ecuación 2, usando que $k/\lambda = \overline{X_n}$ obtenemos

$$\frac{k}{\lambda^2} = \overline{X_n^2} - (\overline{X_n})^2,$$

y como $k = \lambda\overline{X_n}$ obtenemos que

$$\frac{\lambda\overline{X_n}}{\lambda^2} = \overline{X_n^2} - (\overline{X_n})^2,$$

es decir

$$\lambda = \frac{\overline{X_n}}{\overline{X_n^2} - (\overline{X_n})^2},$$

y

$$k = \frac{\overline{X_n}^2}{\overline{X_n^2} - (\overline{X_n})^2}.$$

7.5.2 Máxima verosimilitud

La idea de este método es buscar los parámetros que hagan que la muestra que observamos sea la mas probable. Primero introduciremos el método por medio de un ejemplo muy simple, supongamos que tenemos una variable aleatoria X Bernoulli de parámetro p y queremos hallar p. Supongamos que tenemos una muestra de n realizaciones iid de esta variable en la cual n_1 veces salió el valor 1 y n_0 el valor 0 con $n_1 + n_0 = n$. Por lo tanto la probabilidad de que hayamos observado dicha muestra es simplemente $p^{n_1}(1-p)^{n_0}$. Esto nos da una función $L(p)$ que para diferentes valores de p nos da la probabilidad de que hayamos obtenido esa muestra. Por lo tanto si maximizamos en p estaríamos hallando el parámetro que hace que haber observado n_0 veces el 0 y n_1 el 1, es lo mas probable. Para eso derivamos $L(p)$ respecto de p y maximizamos (observemos que $p \in [0,1]$). Si derivamos

$$L'(p) = n_1 p^{n_1-1}(1-p)^{n_0} - n_0 p^{n_1}(1-p)^{n_0-1} = (1-p)^{n_0-1} p^{n_1-1}(n_1(1-p) - n_0 p)$$

Si igualamos la derivada a 0 nos queda que $(n_1(1-p) - n_0 p) = 0$, es decir $p = n_1/(n_1+n_0) = n_1/n$, lo cual nos dice que el valor de p que hace que la muestra que observamos sea la mas probable es simplemente la frecuencia de las veces con que observamos el 1, que era de esperarse.

Veamos otro ejemplo simple, supongamos que tenemos una variable que toma únicamente los valores 0, 1, y 2 con probabilidades $(1-\theta)/3$, $1/3$ y $1+\theta/3$ respectivamente con $\theta \in [0,1]$. Supongamos que en nuestra muestra observamos n_0 veces el 0, n_1 veces el 1 y n_2 veces el 2. Por lo tanto la función a maximizar es simplemente

$$L(\theta) = \left(\frac{1-\theta}{3}\right)^{n_0}\left(\frac{1}{3}\right)^{n_1}\left(\frac{1+\theta}{3}\right)^{n_2}.$$

En la función anterior podemos quitar el término del medio ya que no depende de θ y por lo tanto un cierto valor θ_0 maximiza $L(\theta)$ si y sólo si maximiza

$$\left(\frac{1-\theta}{3}\right)^{n_0}\left(\frac{1+\theta}{3}\right)^{n_2}. \tag{7.5}$$

No estamos diciendo que el valor que toma la función anterior es igual al que toma $L(\theta)$ sino que el máximo se realizará en el mismo θ_0. Por otra parte como la función logaritmo es creciente, θ_0 maximiza $L(\theta)$ si y sólo si maximiza $log(L(\theta))$ con lo cual muchas veces se toma logaritmo antes de derivar. Aquí simplemente derivamos (7.5) y obtenemos

$$\frac{1}{3^{n_0+n_2}}(1-\theta)^{n_0-1}(1+\theta)^{n_2-1}(-n_0 + n_2(1-\theta))$$

si igualamos a 0 y despejamos θ nos queda $1 - n_0/n_2 = \theta$.

Para formalizar esta idea tenemos que definir primero al función de verosimilitud. Dada una muestra $X_1, \ldots, X_n$ iid de una variable X cuya distribución depende de ciertos parámetros $\theta = (\theta_1, \ldots, \theta_k)$. Supongamos primero que X es una variable aleatoria discreta que toma únicamente los valores $z_1, \ldots, z_r$ con probabilidades

$p_1(\theta),\ldots,p_r(\theta)$ tal que $p_1(\theta)+p_2(\theta)+\ldots+p_r(\theta)=1$. Si en nuestra muestra el valor z_1 fue tomado n_1 veces, el valor z_2 fue tomado n_2, hasta el valor z_r el cual fue tomado una cantidad n_r de veces. La función de verosimilitud que tenemos que maximizar es

$$L(\theta)=p_1(\theta)^{n_1}p_2(\theta)^{n_2}p_3(\theta)^{n_3}\cdots p_r(\theta)^{n_r}. \tag{7.6}$$

Observemos que esta es una función a valores reales, que toma valores en $\mathbb{R}^k$ que es donde varían los k posibles parámetros $(\theta_1,\ldots,\theta_k)$. Como vimos en el ejemplo anterior no siempre todas las p_i dependen de los parámetros, en cuyo caso ese término se puede quitar a la hora de maximizar la función $L(\theta)$. No necesariamente existe un único máximo de dicha función. No obstante el método de máxima verosimilitud busca el o los parámetros que maximizan L.

Caso contínuo

Si en lugar de una variable discreta tenemos una variable continua X cuya densidad que denotamos $f(x|\theta)$ depende de ciertos parámetros $\theta=(\theta_1,\ldots,\theta_k)$ y tenemos una muestra iid de X se procede de forma análoga. En este caso, si nuestras observaciones fueron $x_1,\ldots,x_n$ (usamos letras minúsculas para indicar que nos referimos a números reales y no variables aleatorias) la función a maximizar respecto de θ es el producto de las densidades, evaluadas en estos puntos, es decir.

$$L(\theta)=\prod_{i=1}^{n}f_X(x_i|\theta).$$

Veamos un ejemplo simple,

Ejemplo 7.6. Supongamos que tenemos $x_1,\ldots,x_n$ realizaciones iid de $X\sim N(\mu,1)$, la función de verosimilitud es

$$L(\mu)=\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}\big[(x_1-\mu)^2+(x_2-\mu)^2+\cdots(x_n-\mu)^2\big]\Big).$$

Como dijimos antes, podemos primero eliminar la constante $\frac{1}{\sqrt{2\pi}}$ y luego tomar logaritmo, por lo tanto la función a maximizar es

$$-\frac{1}{2}\big[(x_1-\mu)^2+(x_2-\mu)^2+\cdots(x_n-\mu)^2\big],$$

por lo tanto si derivamos respecto de μ e igualamos a 0 el máximo se obtiene en $\overline{X_n}=\mu$ que es muy razonable.

Ejemplo 7.7. Supongamos que tenemos $x_1,\ldots,x_n$ realizaciones iid de $X\sim\text{Exp}(\lambda)$, la función de verosimilitud para λ es

$$L(\lambda)=\prod_{i=1}^{n}\lambda\exp\{-\lambda x_i\}=\lambda^n\exp\Big\{-\lambda\sum_{i=1}^{n}x_i\Big\},$$

con $x_i\geq 0\ \forall i$, derivando obtenemos

$$L'(\lambda)=\lambda^{n-1}\exp\left\{-\lambda\sum_{i=1}^{n}x_i\right\}\left(n-\lambda\sum_{i=1}^{n}x_i\right)$$

y por lo tanto, como $\lambda \neq 0$, si hacemos $L'(\lambda) = 0$ obtenemos $\lambda = \frac{n}{\sum_{i=1}^{n} x_i}$, es fácil ver, mirando el signo de $L'(\lambda)$ que es un máximo. Por lo tanto $\hat{\lambda} = \frac{1}{\bar{X}_n}$ es el estimador máximo verosimil (E.M.V.) de λ.

Ejemplo 7.8. Supongamos que tenemos $x_1, \ldots, x_n$ realizaciones iid de $X \backsim \mathrm{U}_{[0,b]}$ y queremos estimar b, la función de verosimilitud es entonces

$$L(b) = \prod_{i=1}^{n} \frac{1}{b} \mathbb{I}_{[0,b]}(x_i) = \begin{cases} \frac{1}{b^n} & \text{si } 0 < x_1, \ldots, x_n < b \\ 0 & \text{si no} \end{cases} = \begin{cases} \frac{1}{b^n} & \text{si } b > \max\{x_1, \ldots, x_n\} \\ 0 & \text{si no} \end{cases}$$

Como la función $1/b^n$ es decreciente obtenemos que $\hat{b} = \max\{x_1, \ldots, x_n\}$.

Chapter 8

Estadística descriptiva

8.1 Función cuantil: cuantiles teóricos

Dada una probabilidad $u \in (0,1)$ y una variable aleatoria X, nos puede interesar conocer un valor z (que depende de u) que hace que $P(X \leq z) \geq u$, y no solo eso, conocer cual es el menor de los z que lo verifican, que denotaremos por ahora como a. Por ejemplo si X corresponde a una medición, podemos querer el valor crítico a para la cual con probabilidad muy alta, u, la X no supera ese umbral. Este valor z es el que se obtiene mediante la función cuantil que definiremos. Si X tiene distribución F_X, el valor que queremos es el mas pequeño que hace que $F_X(z) \geq u$ ya que $P(X \leq z)$ es por definición $F_X(z)$. Veamos como son los z tal que $F_X(z) \geq u$. Como F_X es no decreciente, si para un z_0, $F_X(z_0) \geq u$, cualquier valor $z_1 \geq z_0$ también va a verificar $F_X(z_1) \geq u$ ya que $u \leq F_X(z_0) \leq F_X(z_1)$. Esto implica que los z que verifican $F_X(z) \geq u$ son una semirrecta que puede ser (a,∞) o $[a,\infty]$ para algún a que queremos determinar. Como la función F_X es continua por derecha es fácil ver que $F_X(a) \geq u$. Por lo tanto $\{z : F_X(z) \geq u\}$ es una semirrecta de la forma $[a,+\infty)$. Si F_X es invertible existe un único a tal que $F_X(a) = u$, y el valor a se obtiene como $a = F_X^{-1}(u)$[1]. Es claro que si la X es discreta, como por ejemplo la binomial o la geométrica, la función F_X no es invertible, no obstante, aún en el caso en que X tenga densidad, la función F_X no tiene por qué ser invertible[2].

En la figura 8.1 se muestra a la izquierda la densidad de una variable aleatoria, la línea azul representa el valor $Q(1/2)$, es decir el área encerrada por el gráfico de la densidad, desde $-\infty$ hasta dicho valor es $1/2$. Mirando el gráfico de la distribución de la variable, que se representa a la derecha en 8.1, vemos que $Q(1/2)$ se obtiene intersectando la recta horizontal a altura $1/2$ con la función de distribución de la variable, esto es equivalente a hacer $F_X^{-1}(1/2)$. En rojo se representa $Q(0.95)$, es decir el 95% del área total (que es 1), es encerrado por el gráfico entre $-\infty$ y $Q(0.95)$. Nuevamente esto corresponde, en el gráfico de la derecha, a cortar con la recta horizontal de altura 0.95 al gráfico de la distribución de la variable. En magenta se representa en ambos casos $Q(0.9)$.

[1] ya que $a = F_X^{-1}(F_X(a)) = F_X^{-1}(u)$

[2] Un ejemplo de variable con densidad cuya distribución no es invertible es la variable aleatoria que tiene densidad $f(x) = 1/2$ si $x \in [0,1/2] \cup [1,3/2]$ y 0 en otro caso

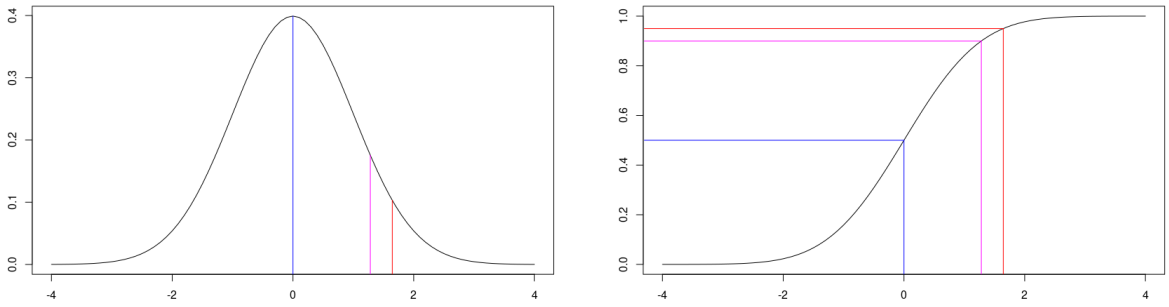


Figure 8.1: A la izquierda la densidad de una variable aleatoria, en rojo se representa $Q(1/2)$, en magenta $Q(0.9)$ y en rojo $Q(0.95)$. A la derecha se representa la distribución de dicha variable.

En general, la función cuantil de una variable aleatoria X nos da para cada $u \in (0,1)$ el infimo de los valores de $x \in \mathbb{R}$ para los cuales la probabilidad de que la variable sea menor o igual que x es mayor o igual que u, es decir

$$Q(u) = \inf\Big\{x \in \mathbb{R} : P(X \leq x) \geq u\Big\}.$$

o lo que es lo mismo $Q(u) = \inf\{x \in \mathbb{R} : F_X(x) \geq u\}$. Como la función F es continua por derecha, el ínfimo anterior se realiza en un punto, es decir, es un mínimo, por lo tanto

$$Q(u) = \min\Big\{x \in \mathbb{R} : F_X(x) \geq u\Big\}.$$

Es inmediato que en general $F_X(Q(u)) \geq u$, veamos primero un ejemplo donde $F_X(Q(u)) > u$. Supongamos que X toma únicamente el valor 0. En este caso, como vimos antes,

$$F_X(x) = \begin{cases} 0 \text{ si } x < 0 \\ 1 \text{ si } x \geq 0 \end{cases}$$

Por lo tanto para cualquier $u \in (0,1)$, $Q(u) = 0$. Observar que $F_X(0) = 1$ por lo tanto $F_X(Q(u)) > u$ para todo $u \in (0,1)$.

Veamos un caso menos trivial, supongamos que $X \sim \mathrm{U}(0,1)$, en este caso sabemos que

$$F_X(x) = \begin{cases} 0 \text{ si } x < 0 \\ x \text{ si } 0 \leq x \leq 1 \\ 1 \text{ si } x > 1 \end{cases}$$

Como F_X es invertible en $[0,1]$ (y su inversa es ella misma), tenemos que $Q(u) = u$.

Recordemos que la función de distribución F_X de una variable aleatoria siempre se puede definir, tenga esta densidad o no. Por lo tanto la función cuantil también se puede definir siempre.

En general en R la función cuantil de una distribución empieza con la letra q y el nombre de la distribución, por ejemplo para la uniforme es qunif para la binomial es qbinom, etc.
Veamos ahora algunas definiciones que usaremos luego.

- El valor $Q(1/2)$ se conoce como mediana (o segundo cuartil), se denota como Q_2.
- El valor $Q(1/4)$ es el **primer cuartil**, se denota como Q_1.
- El valor $Q(3/4)$ es el **tercer cuartil**, se denota como Q_3.
- El valor $Q_3 - Q_1$ es el **rango inter-cuartílico.**

```
qnorm(1/2) #devuelve la mediana de la normal 0,1
```

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

```
qnorm(1/2,1,4) #devuelve la mediana de la normal con media 1 y varianza 2
```

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

```
qnorm(0.95) #cuantil 0.95 de la normal 0,1
```

```
## [1] 1.644854
```

```
qnorm(3/4)-qnorm(1/4) #rango intercuartilico de la normal 0,1
```

```
## [1] 1.34898
```

8.2 Cuantiles empíricos y Boxplot

Supongamos que tenemos una muestra $X_1, \ldots, X_n$ iid de datos que tienen la misma distribución que una cierta variable X. Como vimos, hay ciertos valores relacionados a X que son importantes por ejemplo su varianza y su esperanza, que se estiman con los estimadores que vimos en la sección anterior. Pero tambien, como vimos en la sección 8.1, la mediana y los cuartiles, así como el rango inter cuartílico nos dan información de la variable. Veremos ahora como estimar estos valores a partir de la muestra.

8.2.1 Cuantiles empíricos y Boxplot

Recordemos que la función cuantil se define como $Q(u) = \min\{x : F_X(x) \geq u\}$ para $0 < u < 1$. Si lo que tenemos es una muestra $X_1, \ldots, X_n$ estimamos la función cuantil Q con la función Q_n (que depende de la muestra), definida como

$$Q_n(u) = \min\{x : F_n(x) \geq u\},$$

siendo F_n la función de distribución empírica asociada a la muestra. Por lo tanto la mediana se estima por $Q_n(1/2)$, el primer cuartil por $Q(1/4)$ y el tercer cuartíl por $Q(3/4)$. Denotemos $X_{(1)} \leq \cdots \leq X_{(n)}$ la muestra ordenada, se deja como ejercicio verificar que

$$Q_n(u) = \begin{cases} X_{(1)} & \text{si } 0 < u \leq 1/n \\ X_{(2)} & \text{si } 1/n < u \leq 2/n \\ X_{(3)} & \text{si } 2/n < u \leq 3/n \\ \vdots & \vdots \\ X_{(n)} & \text{si } (n-1)/n < u \leq 1 \end{cases}$$

En R la mediana se calcula simplemente como median(c($X_1, \ldots, X_n$)) no obstante se pueden obtener los cuartiles la mediana, el máximo y el mínimo con el comando summary. En el caso de que hayan valores repetidos, R cuenta ese valor tantas veces como aparece, por ejemplo la mediana de $1, 2, 2$ es 2 y no 1.5.

```
summary(c(1,2,4,4,5,3))

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.250   3.500   3.167   4.000   5.000

summary(c(1,2,2))

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.500   2.000   1.667   2.000   2.000

summary(runif(100))

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.001339 0.257713 0.527333 0.533861 0.814378 0.994667
```

Una forma útil de representar estos valores es mediante el **boxplot** (ver figura 8.2) en el cual se representa una caja cuya linea horizontal del medio corresponde a la mediana (es decir está a la misma altura que la mediana muestral), la de mas abajo esta a altura Q_1 y la siguiente a altura Q_3. De la parte inferior de la caja sale una linea vertical punteada que va hasta la altura $Q_1 - 1.5 \times RIC$ donde RIC es el rango inter cuartílico

que, recordemos, se define como $Q_3 - Q_1$. De la parte superior una linea punteada que va hasta $Q_3 + 1.5 \times RIC$. Finalmente los puntos que se representan fuera de estas dos últimas lineas corresponden a observaciones que cayeron fuera de dicho rango.

```
plot.new()
par(mfrow=c(1,2))
datos=rnorm(100)
boxplot(datos)
boxplot(rexp(100,1/2))
```

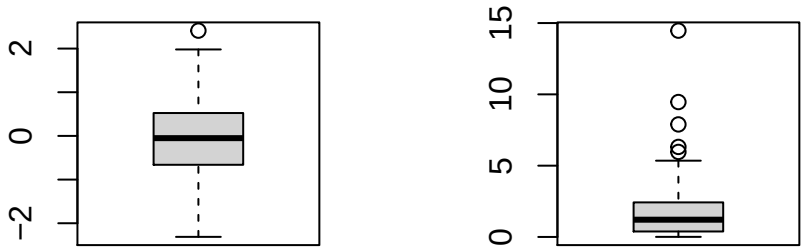


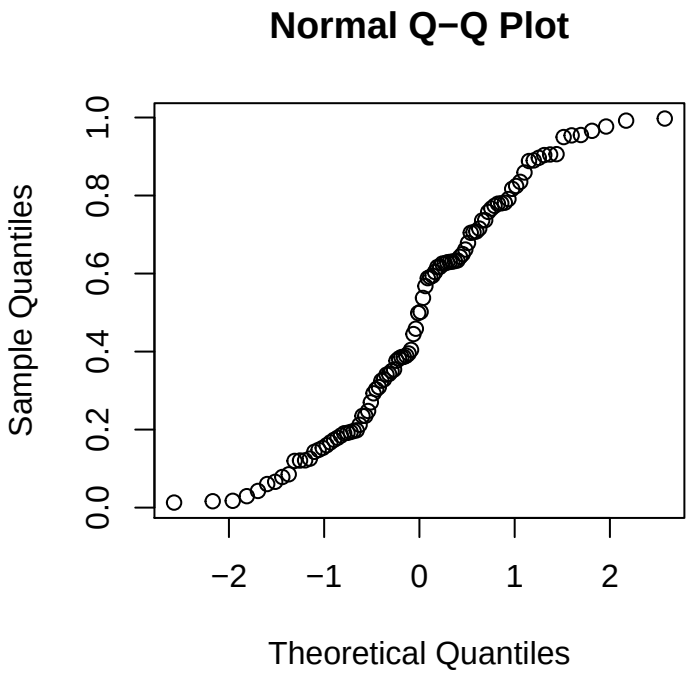
Figure 8.2: Box plots para 100 datos normales, y 100 datos exponenciales de parámetro 2

8.3 Q-Q plots

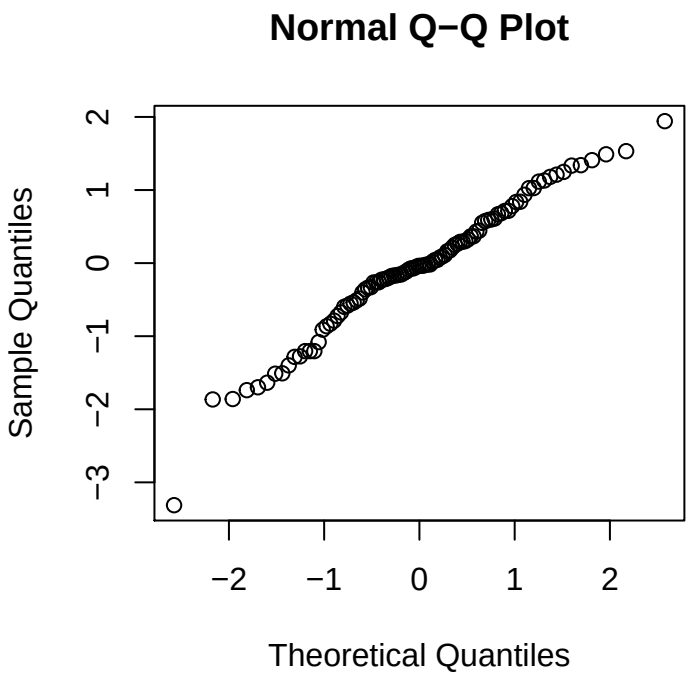
Supongamos que tenemos una muestra $X_1, \ldots, X_n$ iid que vienen de una cierta distribución que desconocemos, una forma de visuarlizar si por ejemplo vienen de una distribución F_0 (supongamos que F_0 es invertible, recordar que su inversa es la función que da los cuantiles) es mediante el gráfico de cuentiles conocido como qq-plot. En dicho gráfico se pone en el eje de las y la muestra ordenada $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$, y en el eje de las x el valor de los cuantiles teóricos dados por la F_0, es decir se grafican los pares $(X_{(i)}, F_0^{-1}(i/n))$. Si efectivamente la muestra viene de F_0 dichos puntos deberían estar próximos a la diagonal $y = x$ ya que, por ejemplo, la mediana empírica es el dato (si n es par) $X_{(n/2)}$, y la mediana teórica es $F_0^{-1}(1/2)$. Estos valores deberian estar cerca, y por lo tanto el par $(X_{(n/2)}, F_0^{-1}(1/2)))$ estar cerca de la diagonal $y = x$. Análogamente el primer cuartil empírico es el dato $X_{(n/4)}$ (supongamos que n es divisible entre 4 para facilitar la comprensión) mientras que el cuartil teórico es $F^{-1}(1/4)$. Nuevamente estos valores deberian estar próximos y el par

$(X_{(n/4)}, F_0^{-1}(1/4)))$ estar próximo a la diagonal. Razonando de esta manera vemos que todos los pares antes mencionados deberían estar próximos. En R se puede usar el comando qqnorm para el caso en que F_0 sea la normal.

```
qqnorm(runif(100)) #en este caso se alejan mucho de la diagonal
```




```
qqnorm(rnorm(100)) #aquí se parecen
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
−2
−1
0
1
2

Chapter 9

Intervalos de confianza, pruebas de hipótesis

9.1 Intervalos de confianza para la esperanza y para proporciones

En esta sección vamos a ver como se construyen intervalos de confianza para el valor esperado (que denotaremos μ) de una variable X. Veremos únicamente el caso en que la muestra es lo suficientemente grande como para poder usar el Teorema Central del Límite (ver Teorema 6.10). No obstante en el apéndice se detallan los otros casos. Cuando los datos son normales (tanto cuando conocemos su varianza σ^2 como cuando no) es posible construir intervalos *exactos*, es decir tal que $P(\mu \in I_n) = 1 - \alpha$. En general, cuando no sabemos que distribución tienen los datos hay que hacer uso del Teorema Central del Límite y se obtiene un intervalo *aproximado* I_n (que depende de la muestra $X_1, \ldots, X_n$), tal que $P(\mu \in I_n) \approx 1 - \alpha$. No obstante se puede demostrar que

$$P(\mu \in I_n) \to 1 - \alpha \qquad \text{cuando } n \text{ tiende a infinito.}$$

Por último veremos el caso particular de intervalos de confianza para proporciones, es decir en este caso las variables X_i son Bernoulli(p) y queremos un intervalo para la proporción p de éxito. Buscaremos un intervalo I_n (el subíndice n indica que lo construiremos a partir de una muestra $X_1, \ldots, X_n$ de datos independientes e idénticamente distribuidos, con la misma distribución que X) tal que, fijado un nivel $\alpha \in (0,1)$, $P(p \in I_n) \approx 1 - \alpha$.

Para deducir la forma del intervalo veamos primero el caso de datos cualquiera, y supongamos (una hipótesis poco realista en la práctica) que conocemos σ^2 la varianza de los datos.

Ahora los datos son $X_1, \ldots, X_n$ independientes e idénticamente distribuidos como una cierta variable X con varianza $0 < \sigma^2 < \infty$ y esperanza μ. No supondremos que X es normal pero si (a efectos de facilitar el razonamiento) que conocemos σ. Queremos encontrar dado $\alpha > 0$, un intervalo I_n tal que $P(\mu \in I_n) = 1 - \alpha$ pero en el caso general no sera posible encontrar un intervalo exacto sino que encontraremos un interval aproximado I_n tal que $P(I_n) \to 1 - \alpha$ cuando $n \to \infty$. Para eso vamos a usar el Teorema Central del Límite. Es

razonable suponer que $I_n = [\mu - k, \mu + k]$ para un k que tenemos que determinar,

$$P\big(\mu - k \le \overline{X}_n \le \mu + k\big) = P\left(\frac{\mu - k - \mu}{\sigma/\sqrt{n}} \le \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \le \frac{\mu + k - \mu}{\sigma/\sqrt{n}}\right).$$

Por el Teorema Central del Límite 6.10, para valores de n *grandes*, se aproxima la probabilidad anterior por

$$\Phi\left(\frac{k}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{-k}{\sigma/\sqrt{n}}\right). \tag{9.1}$$

Usamos que $\Phi(-t) = 1 - \Phi(t)$, y por lo tanto (9.1) es igual a

$$\Phi\left(\frac{k}{\sigma/\sqrt{n}}\right) - \left[1 - \Phi\left(\frac{k}{\sigma/\sqrt{n}}\right)\right] = 2\Phi\left(\frac{\sqrt{n}k}{\sigma}\right) - 1 = 1 - \alpha.$$

El valor de k se despeja igual que antes, es decir,

$$I_n = \left[\overline{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \; ; \; \overline{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right]. \tag{9.2}$$

Es importante tener presente que el intervalo anterior es aleatorio, es decir depende del azar, por lo tanto al cambiar la muestra cambia I_n. Por ejemplo si para $\alpha = 0.05$ y una muestra dada nos da que $I_n = [1,2]$, esto no significa que el 95% de las veces que sortiemos una muetra $X_1, \ldots, X_n$, μ va a estar en $[1,2]$ ya que al cambiar la muestra cambia el intervalo.

Dado que sabemos que S_n es un estimador de σ en el caso (mas realista) en que lo desconocemos, es razonable sustituir en I_n σ por S_n, es decir, el intervalo que obtenemos es:

$$\left[\overline{X}_n - \frac{S_n}{\sqrt{n}} z_{1-\alpha/2} \; ; \; \overline{X}_n + \frac{S_n}{\sqrt{n}} z_{1-\alpha/2}\right]. \tag{9.3}$$

Veamos un ejemplo para el caso en que los datos tienen distribución exponencial.

Ejemplo 9.1. Se tiene una muestra de 1000 datos de una distribución exponencial de parámetro λ y se sabe que la suma de los datos es 492 y la suma de los cuadrados 467, calcular un intervalo de confianza al 95% para λ.

Observemos primero que nos piden un intervalo para λ y no para el valor esperado de la exponencial, que como sabemos es $1/\lambda$. No obstante si tenemos $I_n = [a,b]$ con $0 < a < b$ un intervalo de confianza para el valor esperado, es claro que $J_n = [1/b, 1/a]$ es un intervalo de confianza para λ. Veamos como calcular entonces un intervalo para el valor esperado. Observemos que estamos en el caso de datos con distribución cualquiera, y σ desconocido.

Buscamos un intervalo I_n de la forma

$$\left[\overline{X_n} - \frac{S_n}{\sqrt{n}} z_{(1-0.05/2)} \; ; \; \overline{X_n} + \frac{S_n}{\sqrt{n}} z_{(1-0.05/2)}\right],$$

si usamos la Observación (7.1) tenemos que $S_n^2 = 1/(999)(467 - 1000(0.492)^2) = 0.22516$, (y $S_n = 0.4745$) por otro lado, $\alpha = 0.05$ es decir $1 - \alpha/2 = 0.975$ y de la tabla de la distribución normal $z_{0.975} = 1.96$ (si hacemos $\mathsf{qnorm}(0.975)$ en R nos da 1.959964). El intervalo para el valor esperado es entonces

$$I_n = \left[0.467 - \frac{0.4745}{31.62} 1.96 \; ; \; 0.467 + \frac{0.4745}{31.62} 1.96\right] = [0.4376; 0.4964].$$

Por lo tanto el intervalo de confianza para λ es

$$J_n = [1/0.4964 \; ; \; 1/4376] = [2.0144 \; ; \; 2.2852].$$

9.1.1 Intervalos de confianza para proporciones

Si bien este es un caso particular del caso anterior lo estudiaremos por separado. Tenemos ahora un experimento cuyos resultados son 0 o 1 (éxito o fracaso), y una muestra $X_1, \ldots, X_n$ de datos independientes e identicamente distribuidos, que corresponden a un 0 o un 1 según si fue éxito o no. Queremos un intervalo I_n para la probabilidad p de éxito. Veamos un ejemplo

Ejemplo 9.2. Para estimar la proporción de roedores de una cierta especie que padecen determinada infección se realiza un examen histológico a 182 individuos y se encuentra que 72 están infectados. Dar un intervalo de confianza 95% para la proporción total de roedores infectados.
Observemos primero que aquí los 182 datos son 0 o 1 segun el roedor está o no infectado. Si llamamos X a la variable que toma el valor 1 con probabilidad $p = P(\text{infectado})$ y 0 con probabilidad $1 - p$ (es decir X tiene distribución de Bernoulli de parámetro p) tenemos que, usando la formula para el cálculo de la esperanza de una variable discreta,

$$E(X) = 0 \times (1 - p) + 1 \times p = p.$$

Por otro lado es fácil ver que $\mathrm{Var}(X) = p(1 - p)$ Lo que nos pide el ejercicio es un intervalo de confianza para p. Ya sabemos, por la Ley de los Grandes Números que un estimador de p es $\overline{X}_n$, y es fácil ver que el estimador $\widetilde{\mathrm{Var}}(X)$ que definimos en la sección anterior se escribe como

$$\widetilde{\mathrm{Var}}(X) = \overline{X_n}(1 - \overline{X_n}).$$

En este caso, dado que sabemos $\overline{X_n}$ en lugar de usar S_n usamos $\sqrt{\widetilde{\mathrm{Var}}(X)}$ por lo tanto el intervalo nos queda

$$\left[\overline{X}_n - \frac{\sqrt{\overline{X_n}(1 - \overline{X_n})}}{\sqrt{n}} z_{1-\alpha/2} \; ; \; \overline{X}_n + \frac{\sqrt{\overline{X_n}(1 - \overline{X_n})}}{\sqrt{n}} z_{1-\alpha/2}\right]. \tag{9.4}$$

En nuestro ejemplo $\overline{X_n} = 72/182 \approx 0.3956$ y $\sqrt{182} \approx 13.49$, de donde

$$I_n = \left[0.3956 - \frac{\sqrt{0.3956(1-0.3956)}}{13.49}1.96\ ;\ 0.3956 + \frac{\sqrt{0.3956(1-0.3956)}}{13.49}1.96\right]$$

9.2 Pruebas de hipótesis

Como dijimos al comienzo del capítulo, a veces queremos no sólo un intervalo para la esperanza, μ, de una cierta variable X, sino poder afirmar con cierta certeza (de modo que la probabilidad de equivocarnos sea α, con α chico) si μ supera o no un valor crítico μ_0. La idea intuitiva es que, si tenemos n datos $X_1, \ldots, X_n$, estimamos μ mediante el promedio $\overline{X_n}$ y rechazamos que supera μ_0 si el promedio es menor que $\mu_0 - \delta$ donde $\delta > 0$ es un valor que depende de la cantidad de datos que tenemos, de α etc. Es intuitivo que si queremos tener más seguridad de que no nos estamos equivocando (es decir que α sea más chico) vamos a necesitar más datos. Lo que se plantea en estos casos son dos hipótesis, la *hipótesis nula* (que usualmente se denota como H_0) y que en el caso del valor crítico es $H_0 : \mu \geq \mu_0$ y la *hipótesis alternativa* $H_1 : \mu < \mu_0$. Tenemos dos tipos de errores posibles, o bien la media μ de la población era mayor o igual que μ_0 pero con la información que tenemos rechazamos que lo sea (es decir rechazamos H_0 cuando en realidad era cierta). O bien $\mu < \mu_0$ pero *no rechazamos* H_0 (es decir no rechazamos H_0 cuando en realidad era cierta H_1).

Es importante tener en cuenta en las aplicaciones que si μ_0 es un valor crítico que superarlo implica algún tipo de riesgo o costo grande, es *más grave* el primer error que dijimos, es decir que la media μ superaba el valor crítico, pero aún así rechazamos que lo superara. La probabilidad de dicho error es el nivel (que usualmente se denota con la letra griega α) de la prueba, es decir

$$\alpha = P_{H_0}(\text{rechazar } H_0). \tag{9.5}$$

Por otro lado, el segundo tipo de error, que también queremos que sea chico, esta relacionado con la *potencia* de la prueba, usualmente se denota

$$\beta = P_{H_1}(\text{no rechazar } H_0),$$

y la potencia (que en general queremos que sea 1 o esté proxima a 1), es $1 - \beta = P_{H_1}(\text{ rechazar } H_0)$.

Dado que no sabemos el valor exacto de μ (sino simplemente rechazaríamos H_0 si $\mu \leq \mu_0$) lo estimamos por medio de $\overline{X_n}$. Tenemos que ver cómo determinar *en que región* de valores de $\overline{X_n}$ rechazamos H_0. Es intuitivo que si queremos testear $H_0 : \mu \geq \mu_0$ y calculamos $\overline{X_n}$ y nos da mayor que μ_0 no rechazamos H_0. Pero, ¿qué pasa si nos da $\mu_0 - 1/10^9$?, en este caso si bien no es mayor que μ_0 la diferencia es muy chica como para que estemos seguros de que no estamos cometiendo un error producto de que tenemos pocos datos. En virtud de esto, es razonable pensar que la *región crítica* donde rechazaremos H_0 es de la forma $RC = \{\overline{X_n} < \mu_0 - \delta\}$. Y tenemos que determinar cual es la tolerancia δ permitida. Si tenemos en cuenta (9.5) queremos encontrar δ tal que

$$P_{H_0}\left(\overline{X_n} < \mu_0 - \delta\right) = \alpha.$$

Observemos que una vez que determinamos δ rechazamos H_0 si $\overline{X_n}$ cayó en la región crítica, es decir $\overline{X_n} < \mu_0 - \delta$. Y *no rechazamos* H_0 en caso contrario ($\overline{X_n} \geq \mu_0 - \delta$), en este caso no decimos que H_0 es verdadera sino que no la rechazamos a nivel α. *Cuando rechazamos la hipótesis nula, tenemos evidencia estadística de que la hipótesis nula es falsa. En cambio, si no podemos rechazar la hipótesis nula, no tenemos evidencia estadística de que la hipótesis nula sea verdadera. Esto se debe a que en general no tenemos control sobre β, no establecimos la probabilidad β de no rechazar la hipótesis nula para que fuera pequeña. De hecho fijado n achicar α en general aumenta β*

En lo que sigue consideraremos diferentes pruebas de hipótesis y veremos cuales son los δ que hay que tomar. Al igual que hicimos con los intervalos de confianza, sólo veremos el caso (más realista) en que la varianza de los datos es desconocida, y n es suficientemente grande como para poder aplicar el Teorema Central del Límite. Los otros casos se pueden ver en el apéndice.

9.2.1 Pruebas de hipótesis unilaterales

Supongamos que $X_1, \ldots, X_n$ es una muestra iid de una cierta variable aleatoria X, supongamos que $\text{var}(X) < \infty$, las pruebas de hipótesis

$$\begin{cases} H_0: & \mu & \leq \mu_0 \\ H_1: & \mu & > \mu_0 \end{cases} \qquad \begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & > \mu_0 \end{cases} \qquad \begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & = \mu_1 \end{cases}$$

con $\mu_1 > \mu_0$ tienen región crítica

$$RC = \left\{ \overline{X_n} > \mu_0 + \frac{z_{1-\alpha}}{\sqrt{n}} S_n \right\}.$$

Para demostrar esto tenemos que hallar $\delta > 0$ tal que $P_{H_0}(\overline{X}_n > \mu_0 + \delta) = \alpha$. Usaremos el T.C.L,

$$\alpha = P_{H_0}\Big(\frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} > \frac{\delta}{S_n/\sqrt{n}} \Big) \approx 1 - \Phi\Big(\frac{\delta}{S_n/\sqrt{n}} \Big)$$

o lo que es lo mismo

$$1 - \alpha = \Phi\Big(\frac{\delta}{S_n/\sqrt{n}} \Big).$$

Si aplicamos Φ^{-1} a ambos lados de la igualdad anterior obtenemos

$$\frac{\delta}{S_n/\sqrt{n}} = \Phi^{-1}(1-\alpha) = z_{1-\alpha},$$

de donde se sigue que

$$\delta = z_{1-\alpha} \frac{S_n}{\sqrt{n}}.$$

Análogamente, las pruebas de hipótesis

$$\begin{cases} H_0: & \mu & \geq \mu_0 \\ H_1: & \mu & < \mu_0 \end{cases} \qquad \begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & < \mu_0 \end{cases} \qquad \begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & = \mu_1 \end{cases}$$

donde $\mu_1 < \mu_0$, tienen región crítica

$$RC = \left\{ \overline{X_n} < \mu_0 - \frac{z_{1-\alpha}}{\sqrt{n}} S_n \right\}.$$

9.2.2 Pruebas de hipótesis bilaterales

En este caso queremos contrastar

$$\begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & \neq \mu_0 \end{cases}.$$

Vamos a rechazar H_0 cuando $\overline{X}_n$ esté a mas de un cierto δ de μ_0, es decir $RC = \{|\overline{X}_n - \mu_0| > \delta\}$. El valor de δ depende del nivel α de la prueba, de la varianza de los datos, y de n. Razonando igual que antes (usando el TCL), se llega a que la región crítica, si trabajamos a nivel $\alpha \in (0,1)$, es

$$RC = \left\{ |\overline{X_n} - \mu_0| \geq S_n \frac{z_{1-\alpha/2}}{\sqrt{n}} \right\}.$$

Observar que el cuantil que aparece es $z_{1-\alpha/2}$ y no $z_{1-\alpha}$ como en las pruebas unilaterales vistas anteriormente.

9.2.3 La potencia de la prueba

La potencia de la prueba como vimos se define como $1-\beta$ y es la probabilidad de detectar (bajo H_1) que no se cumple la hipótesis nula. Consideremos la prueba de hipótesis

$$\begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & = \mu_1 \end{cases}$$

con $\mu_1 > \mu_0$ para el caso de datos normales y supongamos para facilitar las cuentas que σ es conocido. Vamos a calcular el β de dicha prueba.

$$\beta = P_{H_1}\left(\overline{X_n} \leq \mu_0 + \frac{\sigma z_{1-\alpha}}{\sqrt{n}} \right) = P_{H_1}\left(\frac{\sqrt{n}(\overline{X_n} - \mu_1)}{\sigma} \leq \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma} + z_{1-\alpha} \right) = \Phi\left\{ z_{1-\alpha} - \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma} \right\}$$

Por ejemplo si $\sigma = 1$, $\alpha = 5\%$, $z_{1-\alpha} = 1{,}645$, $\mu_0 = 0$, tenemos la siguiente variación de β según n (para

dos posibles valores de μ_1),

$\mu_1 = 0.5$	
n	β
4	0.740
9	0.558
16	0.361
25	0.196
36	0.088
44	0.047

$\mu_1 = 0.25$	
n	β
4	0.874
9	0.814
16	0.740
25	0.653
36	0.557
44	0.495

Es decir que por ejemplo para $\mu_1 = 0.5$ y para $n = 16$ el test tiene potencia 1-0.361=0.63, esto quiere decir que el 36% de las veces no rechazamos H_0 cuando en realidad deberíamos rechazarla. En la tabla de la derecha se observar que cuanto mas parecida es la hipotesis alternativa a la nula (es decir en nuestro caso cuando μ_1 se acerca a $\mu_0 = 0$) beta aumenta (y por lo tanto la potencia baja).

9.2.4 p-valor

Definición 9.3. Supongamos que tenemos una muestra $X_1, \ldots, X_n$ de una cierta variable X. En general, dada una prueba de hipótesis

$$\begin{cases} H_0 : \ \theta \in A \\ H_1 : \ \theta \notin A \end{cases}$$

con $A \subset \mathbb{R}$, cuya región crítica sea $RC = \{T_n \geq k\}$ con $T_n = T(X_1, \ldots, X_n)$ un estimador (por ejemplo $T_n = \overline{X_n}$) de θ (que en el caso de la esperanza es $\theta = E(X)$) y dados $\mathbf{x}_n = (x_1, \ldots, x_n)$ n datos, (es decir n números reales) el $p-valor$ es

$$\sup_{\theta \in A} P\Big(T(X_1, \ldots, X_n) \geq T_n(\mathbf{x}_n)\Big),$$

esto es, la probabilidad de que nuestro estadístico tome un valor tanto o más extremo que el que observamos (donde el que observamos es $T_n(\mathbf{x}_n)$).

Veamos con un ejemplo como se calcula y para que sirve,

Ejemplo 9.4. Sea $X_1, \ldots, X_n$ una muestra iid de $X \backsim N(\mu, 1)$, consideremos la prueba

$$\begin{cases} H_0 : \ \mu = \mu_0 \\ H_1 : \ \mu \neq \mu_0 \end{cases}$$

Si trabajamos a nivel $\alpha \in (0,1)$, vimos que la región crítica de la prueba es

$$RC = \left\{ |\overline{X_n} - \mu_0| \geq \frac{z_{1-\alpha/2}}{\sqrt{n}} \right\}. \tag{9.6}$$

Donde aquí hemos usado que $\sigma = 1$ y por lo tanto no necesitamos estimarlo con S_n. Si definimos $T(X_1, \ldots, X_n) = |\overline{X_n} - \mu|$ y tenemos n datos $\mathbf{x}_n$ (cuyo promedio denotamos $\overline{\mathbf{x}_n}$), el $p - valor$ es (observar que bajo H_0 $\mu = \mu_0$)

$$P_{H_0}\Big(|\overline{X_n} - \mu_0| \geq |\overline{\mathbf{x}_n} - \mu_0|\Big) = 1 - P_{H_0}\Big(|\overline{X_n} - \mu_0| \leq |\overline{\mathbf{x}_n} - \mu_0|\Big) = 1 - P_{H_0}\Big(\sqrt{n}|\overline{X_n} - \mu_0| \leq \sqrt{n}|\overline{\mathbf{x}_n} - \mu_0|\Big)$$

Bajo H_0 la variable $\sqrt{n}(\overline{X}_n - \mu_0)$ tiene distribución normal con media 0 y varianza 1 por lo tanto lo anterior es igual a

$$1 - \Phi\Big(\sqrt{n}|\overline{\mathbf{x}_n} - \mu_0|\Big) + \Phi\Big(-\sqrt{n}|\overline{\mathbf{x}_n} - \mu_0|\Big) = 2\Big(1 - \Phi(\sqrt{n}|\overline{\mathbf{x}_n} - \mu_0|)\Big).$$

Supongamos que este valor es menor que α (es decir p-valor menor que α), es decir

$$2\Big(1 - \Phi(\sqrt{n}|\overline{\mathbf{x}_n} - \mu_0|)\Big) < \alpha,$$

o lo que es lo mismo

$$1 - \frac{\alpha}{2} < \Phi\Big(\sqrt{n}|\overline{\mathbf{x}_n} - \mu_0|\Big).$$

Si aplicamos Φ^{-1} de ambos lados y recordamos que $z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ lo que obtuvimos no es otra cosa que

$$z_{1-\alpha/2} < \sqrt{n}|\overline{\mathbf{x}_n} - \mu_0|$$

y por lo tanto, si observamos la forma de la región crítica lo que llegamos es que si nuestras observaciones $\mathbf{x}_n$ son tal que el p-valor es menor que α entonces *estamos en la región crítica*, y por lo tanto se rechaza H_0. Este es un caso particular de una regla mnemotécnica que usualmente se usa, que dice que si el p-valor que obtenemos es menor que α entonces se rechaza H_0 a nivel α.
En el caso de una prueba unilateral para la media de datos normales con media μ y varianza σ^2, por ejemplo

$$\begin{cases} H_0 : \ \mu \leq \mu_0 \\ H_1 : \ \mu > \mu_0 \end{cases}$$

vimos que el estadístico que usamos es $\overline{X}_n$, y por lo tanto el p-valor es $P_{H_0}(\overline{X}_n > \overline{\mathbf{x}_n})$ donde nuevamente $\overline{\mathbf{x}_n}$ es el promedio de las n observaciones que obtuvimos. Bajo H_0 $\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu_0)$ tiene distribución normal con media 0 y varianza 1, por lo tanto el p valor es

$$P_{H_0}(\overline{X}_n > \overline{\mathbf{x}_n}) = P_{H_0}\Big(\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu_0) > \frac{\sqrt{n}}{\sigma}(\overline{\mathbf{x}_n} - \mu_0)\Big) = 1 - \Phi\Big(\frac{\sqrt{n}}{\sigma}(\overline{\mathbf{x}_n} - \mu_0)\Big),$$

si suponemos nuevamente que el p-valor es menor que α tenemos que

$$1 - \Phi\Big(\frac{\sqrt{n}}{\sigma}(\overline{\mathbf{x}_n} - \mu_0)\Big) < \alpha$$

o lo que es lo mismo

$$1 - \alpha < \Phi\Big(\frac{\sqrt{n}}{\sigma}(\overline{\mathbf{x}_n} - \mu_0)\Big).$$

Nuevamente aplicando Φ^{-1} y usando que $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$, lo que obtenemos es

$$z_{1-\alpha} < \frac{\sqrt{n}}{\sigma}(\overline{\mathbf{x}_n} - \mu_0), \quad \text{es decir} \quad \overline{\mathbf{x}_n} > \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha},$$

y por lo tanto los datos que tenemos están en la región crítica, es decir nuevamente se rechaza H_0.

En la práctica (y sobre todo para valores de n chicos) el p-valor, salvo para casos muy particulares, no se puede calcular exactamente, como hicimos en el caso de datos Normales. Lo que se hace es simular el estadístico $T(X_1, \ldots, X_n)$ bajo H_0 una cantidad grande de veces. Pares eso se simulan por ejemplo l (con l grande) copias iid de los n datos y se calculan $T_1, \ldots, T_l$ los l estadísticos en cada una de las copias, y entonces el p valor se estima (en el caso de que la región crítica sea de la forma $T(X_1, \ldots, X_n) > c$ para un cierto valor c que depende de α, n etc) por el promedio de las veces que los T_i supera $T(\mathbf{x}_n)$ donde $\mathbf{x}_n$ son los datos que originalmente teníamos.

Chapter 10

Pruebas de bondad de ajuste

10.1 Distancia de Kolmogorov

Supongamos que queremos ver que tan lejos esta la distribución empírica F_n de la verdadera distribución F_X, una forma de medir esto es calcular

$$D_n(F_X, F_n) = \sup_{x\in\mathbb{R}} \left|F_X(x) - F_n(x)\right|. \tag{10.1}$$

Observemos primero que D_n depende de la muestra $X_1, \ldots, X_n$ y por lo tanto es una variable aleatoria. No obstante, tiene una propiedad muy importante, la distribución de la variable aleatoria D_n *no* depende de quien sea F_X. Como $0 \le F_n(x) \le 1$ y $0 \le F_X(x) \le 1$ para todo x, y el valor absoluto de la resta de dos números en $[0,1]$ da un número en $[0,1]$ se tiene que $0 \le D_n \le 1$. En la práctica, así escrito, calcular D_n requiere calcular $|F_X(x) - F_n(x)|$ para infinitos valores de x y tomar el supremo, no obstante, como ya vimos en la sección anterior F_n es constante a trozos y la función F_X es no decreciente, es claro que el máximo entre $|F_X(x) - F_n(x)|$ se va a dar cuando x es un punto de la muestra (ver figura 10.1). En estos casos, como sabemos que $F_n(X_{(j)}) = j/n$ se tiene que el valor D_n, coincide con

$$\max_{i=1,\ldots,n} \max\left\{\left|F_X(X_{(i)}) - \frac{i}{n}\right|, \left|F_X(X_{(i)}) - \frac{i-1}{n}\right|\right\} \tag{10.2}$$

Un resultado muy importante, que no demostraremos, prueba que para valores de n *grandes* (es decir, cuando n tiende a infinito) la variable aleatoria $\sqrt{n}D_n$ *tiende* a una cierta variable que llamaremos $\mathcal{K}$ (cuya distribución no depende de quien sea F_X). Esto se traduce, desde un punto de vista práctico, en que para valores *grandes* de n, podemos aproximar $P(\sqrt{n}D_n \in [a,b])$ por $P(\mathcal{K} \in [a,b])$ para todo $a < b$. En particular, si $a = -\infty$, se sigue que $P(\sqrt{n}D_n \le b) \to P(\mathcal{K} \le b)$ o lo que es lo mismo, $F_{\sqrt{n}D_n}(b) \to F_{\mathcal{K}}(b)$ para todo b. La distribución $F_{\mathcal{K}}$ no es ninguna de las se dieron anteriormente (el gráfico de la funcion $F_{\mathcal{K}}$ se muestra en la figura 10.2 a la izquierda, mientras que la densidad (es decir su derivada) se muestra a la derecha), no obstante, existen tablas que dan el valor $F_{\mathcal{K}}(b)$ para valores de b. Este resultado nos será de utilidad mas

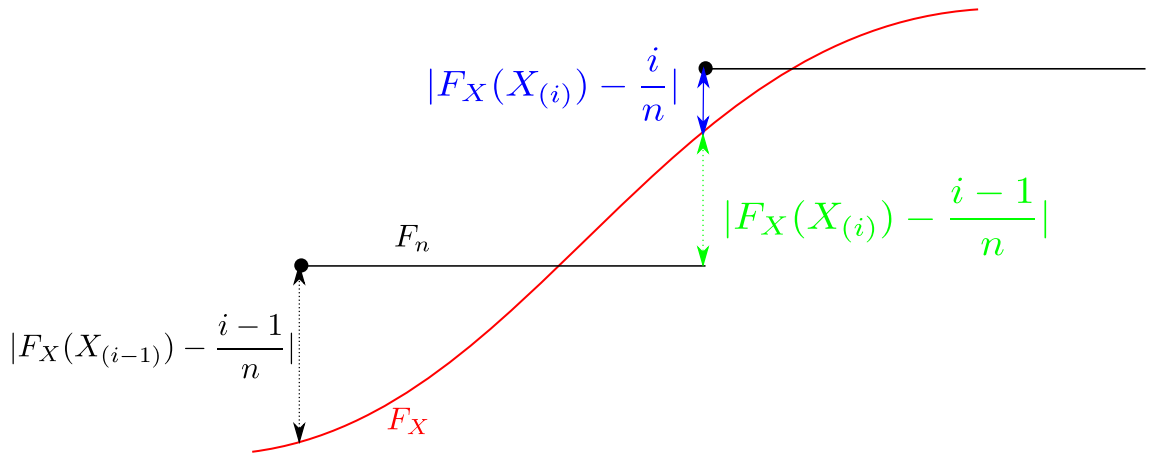


Figure 10.1: En rojo se grafica la función de distribución F_X en negro la distribución empírica F_n. En azul se muestra $\left|F_X(X_{(i)}) - \frac{i}{n}\right|$ y en verde $\left|F_X(X_{(i)}) - \frac{i-1}{n}\right|$

adelante, para *testear* si efectivamente la muestra $X_1, \ldots, X_n$ proviene de la distribución F_X o no.

En R

Generemos una muestra aleatoria de $n = 100$ uniformes en [-1,1], la llamamos a, si queremos hacer un plot de la función de distribución empírica, basta usar el comando plot.ecdf(a). Si queremos calcular la distancia D_n entre la distribución empírica de la muestra a, y una normal con media 0 y varianza 1, hacemos ks.test(a,"pnorm",0,1). Esto devuelve dos valores, por un lado D_n, y por otro lado $P(\mathcal{K} \geq \sqrt{n}D_n)$, o lo que es lo mismo, la probabilidad de que una variable con distribución $\mathcal{K}$ tome un valor mayor o igual que el $\sqrt{n}D_n$ que obtuvimos con nuestra muestra. En nuestro ejemplo, como la muestra a es de una uniforme es de esperarse que D_n sea grande mientras que si hacemos lo mismo pero comparando con la uniforme, el valor se reduce considerablemente como se ve:

```
a=runif(100,-1,1)
ks.test(a,"pnorm",0,1)

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  a
## D = 0.16476, p-value = 0.008775
## alternative hypothesis: two-sided

ks.test(a,"punif",-1,1)
```

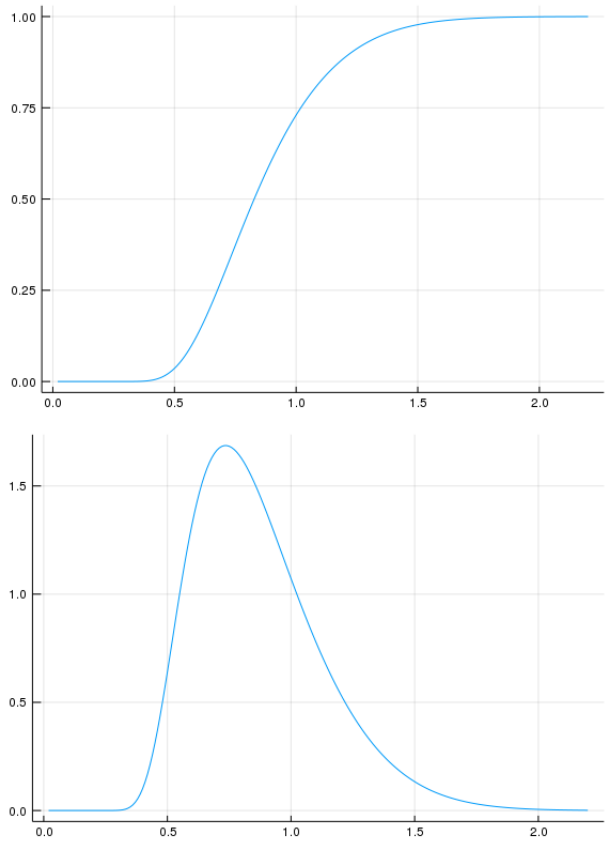


Figure 10.2: A la izquierda gráfico de la distribución de la variable aleatoria $\mathcal{K}$. A la derecha el gráfico de la densidad

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  a
## D = 0.10563, p-value = 0.2144
## alternative hypothesis: two-sided
```

10.2 Prueba de Kolmogorov-Smirnov

Supongamos que tenemos una muestra $X_1, \ldots, X_n$ de variables independientes e idénticamente distribuidas, de una variable aleatoria X cuya distribución es F_X (desconocida) y queremos contrastar la hipótesis nula

H_0, $F_X = F_0$ (donde F_0 es una distribución conocida, que elegimos nosotros, por ejemplo uniforme en $[0,1]$) conta la hipótesis alternativa H_1, $F_X \neq F_0$. Vamos a asumir que F_0 es continua, es decir este test no lo vamos a aplicar al caso de variables aleatorias discretas. Como vimos en la sección 7.4, si conocieramos F_X (en la práctica esto no es cierto en general) una forma de *medir* cuan distinta es F_X de F_0 podría ser calcular $\sup_{x\in\mathbb{R}} |F_X(x) - F_0(x)|$, claramente bajo H_0 esto da 0. Como no conocemos F_X un sustituto razonable es su distribución empírica F_n y por lo tanto estaríamos calculando $\sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)|$, observemos que, en el caso en que efectivamente estamos bajo H_0, esto no es otra cosa que la variable aleatoria D_n que ya estudiamos (donde en la equaciones (10.1) y (10.2) hay que sustituir F_X por F_0). Es razonable entonces que vamos a rechazar H_0 cuando D_n sea grande, es decir la región crítica será de la forma $RC = \{D_n > t_n\}$. Si fijamos $\alpha \in (0,1)$ y queremos que nuestra prueba tenga nivel α (esto es, $P_{H_0}(RC) = \alpha$) entonces buscamos un valor t_n que haga que, $P_{H_0}(D_n > t_n) = \alpha$. Si tomamos $t_n = t_{1-\alpha}/\sqrt{n}$ obtenemos $P_{H_0}(\sqrt{n}D_n > t_{1-\alpha}) = \alpha$. Si usamos que $P_{H_0}(\sqrt{n}D_n > t_{1-\alpha}) \to P(\mathcal{K} > t_{1-\alpha})$ obtenemos que $\alpha = P_{H_0}(\sqrt{n}D_n > t_{1-\alpha}) \approx P(\mathcal{K} > t_{1-\alpha})$. Por lo tanto para valores grandes de n el valor t_n a partir del cual rechazamos H_0 a nivel α se obtiene calculando $t_{1-\alpha}$ de la igualdad $\alpha = P(\mathcal{K} > t_{1-\alpha})$ (y luego usando que $t_n = t_{1-\alpha}/\sqrt{n}$). El valor $t_{1-\alpha}$ se calcula usando una tabla o un software. Una aproximación del valor $t_{1-\alpha}$ (cuando $n > 40$) está dada por la fórmula

$$t_{1-\alpha} \approx \sqrt{-\frac{1}{2}\log\left(\frac{\alpha}{2}\right)} \tag{10.3}$$

donde log es el logaritmo neperiano (en base e). Una cota bastante fina prueba que $P(\sqrt{n}D_n > t) \leq 2\exp(-2t^2)$, para n fijo. Si queremos obtener el valor $t_{1-\alpha}$ de la tabla del test tenemos que proceder de la siguiente forma. Supongamos que tenemos 17 datos (es decir no usamos (10.3)), y estamos trabajando a un nivel $\alpha = 0.05$, si vamos a la fila que corresponde a $n = 17$ de la tabla y la columna 0.05 vemos que el valor que nos da la tabla es 0.318. Este valor 0.318 corresponde al valor que llamamos t_n, a partir del cual rechazamos H_0. Si queremos trabajar con $\alpha = 0.01$ tendríamos que usar, para 17 datos, $t_n = 0.381$. Si tenemos mas de 40 datos se usa la aproximación 10.3.

Para ver que sucede con la potencia de la prueba (P_{H_1}(rechazar H_0)) hay que usar un resultado que dice que bajo H_1, $\sqrt{n}D_n$ tiende a $+\infty$. Por lo tanto para valores grandes de n, fijado α, $\sqrt{n}D_n$ va a superar cualquier valor crítico $t_{1-\alpha}$ y por lo tanto vamos a rechazar H_0 con probabilidad 1.

En R

Cuando introdujimos la distancia de Kolmogorov vimos que en R el comando ks.test calcula D_n y ademas calcula $P(\mathcal{K} > \sqrt{n}D_n)$. Veamos que este valor es de mucha utilidad desde un punto de vista práctico, a la hora de realizar el test que mencionamos antes. Como dijimos, rechazaremos H_0 si el valor $\sqrt{n}D_n$ supera $t_{1-\alpha}$ que vimos que era el valor que verificaba $\alpha = P(\mathcal{K} > t_{1-\alpha})$. Si vemos en la figura 10.3, esto es lo mismo que decir que el área encerrada por el gráfico de la densidad de $\mathcal{K}$ desde $t_{1-\alpha}$ en adelante es α. De forma análoga $P(\mathcal{K} > \sqrt{n}D_n)$ es el área encerrada por la gráfica de $\mathcal{K}$, desde $\sqrt{n}D_n$ en adelante, por lo tanto si $P(\mathcal{K} > \sqrt{n}D_n) < \alpha$, $\sqrt{n}D_n$ tiene que ser *mayor* que $t_{1-\alpha}$, y recordemos que esto sucede si rechazamos H_0. El p valor (asintótico, es decir para valor de n grades) de la prueba es $P(\mathcal{K} > \sqrt{n}D_n)$, nuevamente si p-valor$< \alpha$

rechazo H_0 a nivel α.

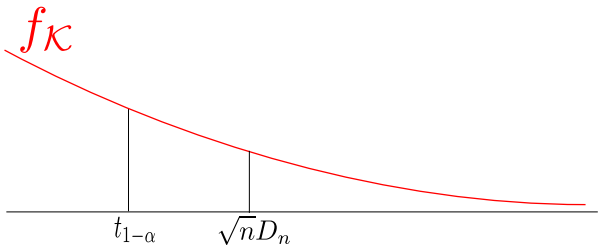


Figure 10.3: En rojo se grafica la densidad de la variable $\mathcal{K}$, si el área desde $t_{1-\alpha}$ en adelante es α, y el área desde $\sqrt{n}D_n$ (el cual es por definición el p-valor) es menor que α esto implica que $\sqrt{n}D_n > t_{1-\alpha}$ y por lo tanto se rechaza H_0.

10.2.1 Dos muestras

Si en lugar de tener una sola muestra, tenemos dos muestras, es decir, tenemos variables $X_1,\ldots,X_n$ independientes, y todas con la misma distribución F_X, y $Y_1,\ldots,Y_m$ independientes y con la misma distribución F_Y, y queremos contrastar H_0 $F_X = F_Y$ contra H_1 $F_X \neq F_Y$, es razonable proceder como antes y calcular $D_{n,m} = \sup_x |F_n^X(x) - F_m^Y(x)|$, donde F_n^X y F_m^Y son las distribuciones empíricas de $X_1,\ldots,X_n$ y $Y_1,\ldots,Y_m$ respectivamente (es importante observar que al igual que antes, para calcular $D_{n,m}$ basta calcularlo en puntos de la muestra). Es claro que la región crítica será de la forma $RC = \{D_{n,m} > t_{n,m}\}$. Para calcular $t_{n,m}$ hacemos uso del siguiente resultado:

$$P\left(\sqrt{\frac{nm}{n+m}}D_{n,m} > t\right) \to P(\mathcal{K} > t)$$

cuando n y m tienden a infinito, para todo t. Si quisiéramos hacer una prueba de hipótesis a nivel α tenemos que proceder igual que para el caso de una muestra: hallar $t_{1-\alpha}$ tal que $\alpha = P(\mathcal{K} > t_{1-\alpha})$, el cual se halla igual que antes a partir de una tabla de la distribución de $\mathcal{K}$, mediante algún software, o usando la fórmula aproximada (10.3) si $n > 40$. De forma totalmente análoga a como hicimos antes, rechazaremos H_0 si

$$\sqrt{\frac{nm}{n+m}}D_{n,m} > t_{1-\alpha}.$$

En R

En este caso el mismo comando ks.test nos sirve, si x e y son las dos muestras, por ejemplo x=runif(100), y=rnorm(150), hacemos ks.test(x,y). Esto devuelve entre otras cosas el p valor $P(\mathcal{K} > \sqrt{nm/(n+m)}D_{n,m})$, y, al igual que antes, rechazamos H_0 (a nivel α) si este valor es menor que α.

```r
x=runif(100)
y=rnorm(150)
ks.test(x,y)


##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x and y
## D = 0.52, p-value = 1.621e-14
## alternative hypothesis: two-sided
```

10.3 Prueba de Lilliefors

En la prueba de Kolgomorov-Smirnov la distribución F_0 esta fijada de antemano, por ejemplo F_0 puede ser la normal con media 0 y varianza 1, o una exponencial de parametro $\lambda = 2$. Si lo que queremos es testear si los datos vienen de una distribución normal, pero no conocemos la media o la varianza. O si provienen de una distribución exponencial pero no conocemos el parámetro, etc, se usa la prueba de Lilliefors. La idea es muy simple, si la distribución F_0 depende de parámetros, los estimamos por medio de la muestra, esto nos da una distribución $\hat{F}_0$ (en el caso de $N(\mu, \sigma^2)$, usamos $N(\overline{X}_n, S_n^2)$), y luego calculamos el estadístico D_n de la prueba de Kolmogorov Smirnov (sustituyendo F_0 por $\hat{F}_0$). Por lo tanto la prueba que planteamos es $H_0 : F_X = \hat{F}_0$ contra $H_1 : F_X \neq \hat{F}_0$. Lamentablemente en el caso en que no se conocen los parámetros exactos de F_0 la distribución de D_n es mas complicada (y depende de quien sea F_0), por lo tanto no se puede usar la misma tabla que para el test de Kolmogorov-Smirnov. Para la distribución normal hay una tabla, para las exponenciales otra etc. En R se puede usar el paquete KScorrect que tiene varias funciones para cada caso.

10.4 Prueba χ^2 de Pearson

Datos categóricos

Supongamos que tenemos un experimento cuyo resultado esta dividido en k categorías distintas $A_1, A_2, \ldots, A_k$ (son las únicas posibles), como por ejemplo los posibles resultados de tirar un dado, o el color que sale al extraer (con reposición) una bola de una urna que contiene bolas de k colores distintos. Denotemos la probabilidad de que resulte A_i como p_i^0. El superíndice ^{0} lo usamos para indicar que estas probabilidades corresponden a la hipótesis nula que queremos verificar si es cierta o no. El resultado de repetir n veces un experimento es una variable aleatoria multinomial, es decir tenemos n-úplas de *letras* A_i que representan la categoría a la que perteneció el experimento i. Por ejemplo podemos haber obtenido $A_1A_1A_1...A_1$, es decir n veces el resultado perteneció a la categoría A_1, otra posibilidad es $A_1A_2A_3A_1 \ldots A_1$, entre otras (la cantidad de posibilidades distintas es k^n). En las pruebas de bondad de ajuste para datos categóricos lo que se tienen son

los resultados de las n repeticiones del experimento, y se desea testear si efectivamente la probabilidad p_i de cada una de las categorías es p_i^0 o no, es decir queremos testear

$$\begin{cases} H_0: \ p_i = p_i^0 & \forall\, i = 1, \ldots, k \\ H_1: \ p_i \neq p_i^0 & \text{para algún } i \end{cases}$$

Si en nuestra muestra de n realizaciones del experimento la categoría A_i salió n_i veces, una estimación de p_i es su frecuencia n_i/n. Por lo tanto a partir de la muestra podemos hacer la siguiente tabla,

nro. esperado	nro. observado
np_1^0	n_1
np_2^0	n_2
$\vdots$	
np_k^0	n_k

Observemos que si estamos bajo H_0 ambas columnas deberian contener valores similares. Consideremos el estadístico de prueba

$$T_n = \sum_{i=1}^{k} \frac{(np_i^0 - n_i)^2}{np_i^0}. \tag{10.4}$$

Y la región crítica es $\{T_n > h_n\}$, donde h_n es un valor que tenemos que hallar para que la prueba tenga nivel α. Para eso usamos que bajo H_0 $P(T_n > h)$ tiende cuando n tiende a infinito (k siempre está fijo) a $P(\chi^2_{k-1} > h)$ donde χ^2_{k-1} es una variable aleatoria con distribución chi-cuadrado con $k-1$ grados de libertad (ver 4.8 y 4.11). Tenemos que hallar h tal que $P(\chi^2_{k-1} > h) = \alpha$. Esto se puede hacer en R mediante el comando $\mathsf{qchisq}(1-\alpha, k-1)$. Esta prueba tiene, asintóticamente cuando k está fijo y $n \to \infty$, potencia 1, esto implica en particular que el error de tipo II tiende a 0. Veamos un ejemplo concreto:

Ejemplo 10.1. Se dispone de datos genéticos de determinada población referidos a dos alelos que llamaremos A y a. Se quiere someter a prueba la hipótesis de que la proporción del alelo a es del 30%. En una muestra de 200 personas los distintos genotipos se reparten según la tabla siguiente:

genotipo	nro. observado
aa	10
Aa	119
AA	71

Denotemos p_a a la probabilidad de alelo a, por la información que se nos da $p_a = 0.3$ por lo tanto la probabilidad del genotipo aa es 0.3^2 mientras que la del $Aa = aA$ es $2 \times 0.7 \times 0.3$ y finalmente la del AA es 0.7^2. En este caso $k = 3$. Los números esperados son $0.3^2 \times 200 = 18$ para el genotipo aa, 84 para el Aa y 98

para el AA. El estadístico T_{200} nos queda

$$T_{200} = \frac{(10-18)^2}{18} + \frac{(84-119)^2}{84} + \frac{(71-98)^2}{98} \approx 25.6.$$

Este valor supera el que arroja el comando $\texttt{qchisq}(0.95,2) = 6$ y por lo tanto se rechaza H_0 a nivel 5%.

Datos continuos

Veamos con un ejemplo como se construye el test y la región crítica.

Ejemplo 10.2. Se tiene una muestra que registra la altura en centímetros de 100 niños de siete años de determinada población. Se intenta ver si es razonable afirmar que la misma corresponde a una variable aleatoria normal con media 118 centímetros y desvío estándar $\sigma = 5$ centímetros. Para ello, se resumen en una tabla las observaciones:

altura en cm.	n. observado (o_i)
menos de 111	7
entre 111 y 115	17
entre 115 y 119	29
entre 119 y 122	30
más de 122	17

Suponer que los datos tienen la distribución de una variable $X \sim N(118,25)$ significa, por ejemplo, que la probabilidad de que la altura de un niño sea menor que 111 es $P(X < 111)$, análogamente la probabilidad de que la altura de un niño este entre 111 y 115 es $P(111 \leq X \leq 115)$, etc. Observemos que estos valores los podemos calcular a partir de la tabla de la distribución normal, o usando la función qnorm. Esto nos dice que, si dicha hipótesis fuese cierta, deberíamos tener, entre las 100 alturas, *aproximadamente*:

- $100 \times P(X < 111) = 100 \times 0.0808$ alturas menores que 111.
- $100 \times P(111 \leq X \leq 115) = 100 \times 0.1935$ alturas entre 111 y 115.
- $100 \times P(115 < X \leq 119) = 100 \times 0.305$ alturas entre 115 y 119.
- $100 \times P(119 < X \leq 122) = 100 \times 0.2088$ alturas entre 119 y 122.
- $100 \times P(X > 122) = 100 \times 0.2119$ alturas mayores que 122

Por lo tanto una idea intuitiva para testear si los datos de la tabla provienen efectivamente de la distribución $N(118,25)$ es compararlos con las frecuencias anteriores. Por ejemplo considerar

$$(100 \times 0.0808 - 7)^2 + (100 \times 0.1935 - 17)^2 + (100 \times 0.305 - 29)^2 + (100 \times 0.2088 - 30)^2 + (100 \times 0.2119 - 17)^2.$$

Donde las diferencias, al igual que en el caso de la definición de la varianza, se elevan al cuadrado para que los valores más grandes *pesen más* que los más chicos. Para darle mayor peso a las diferencias correspondientes a intervalos de probabilidad pequeña vamos a dividir cada sumando por la *frecuencia teórica* del intervalo, es decir consideraremos

$$\frac{(100\times 0.0808-7)^2}{100\times 0.0808}+\frac{(100\times 0.1935-17)^2}{100\times 0.1935}+\frac{(100\times 0.305-29)^2}{100\times 0.305}+ \\ \frac{(100\times 0.2088-30)^2}{100\times 0.2088}+\frac{(100\times 0.2119-17)^2}{100\times 0.2119}=5.315. \quad (10.5)$$

Observemos que si llamamos I_1 al intervalo $(-\infty,111]$, I_2 al intervalo $(111,115]$, I_3 al intervalo $(115,119]$, I_4 al intervalo $(119,122]$ e I_5 al intervalo $(122,+\infty]$ y n al número de observaciones (en nuestro caso $n=100$) lo que calculamos en (10.5) no es otra cosa que

$$T_n=\sum_{i=1}^{5}\frac{\left(n\times P(X\in I_i)-[\text{nro de observaciones en } I_i]\right)^2}{n\times P(X\in I_i)}. \quad (10.6)$$

La región crítica, es decir los valores de $X_1,\ldots,X_n$ que hacen que rechazemos la hipótesis nula de que la muestra tiene distribución $N(118,25)$ serán aquellos que hacen que T_n sea *grande*, por lo tanto es razonable plantear una región crítica de la forma $RC=\{T_n>h\}$ donde h es un valor que se determina en función del nivel de confianza $\alpha\in(0,1)$ en el que estemos trabajando. Es decir, queremos que encontrar h tal que $\alpha=P_{H_0}(T_n>h)$. Se puede demostrar que si los datos efectivamente provienen de la distribución $N(118,25)$ entonces

$$P_{H_0}(T_n>h)\to P(\chi^2_4>h)$$

donde χ^2_4 fue definida en (4.11), con $k=4$. Por lo tanto para determinar h tenemos que plantear

$$\alpha=P(\chi^2_4>h).$$

Por último determinamos el valor de h a partir de la tabla de la distribución χ^2_4. Por ejemplo si estamos trabajando a nivel $\alpha=0.05$ (95% de confianza) y $k=4$ la tabla nos dice que $h=9.488$ (esto se puede calcular también con el comando qchisq(0.95,4), que arroja el valor 9.487729), como $5.315<9.48$ no rechazamos la hipótesis de que los datos provengan de una variable con distribución $N(118,25)$.

La prueba que explicamos a partir de un ejemplo se conoce como Prueba chi-cuadrado de Pearson, veamos ahora la explicación general. Supondremos primero que tenemos k intervalos $I_1,\ldots,I_k$ que forman una partición de todos los números reales, es decir $I_1=(-\infty,a_1]$, $I_2=(a_1,a_2]$, $I_3=(a_2,a_3]$ hasta $I_k=(a_{k-1},+\infty)$, con $a_1<a_2<a_3<\ldots<a_{k-1}$. Vamos a suponer que tenemos una variable X de la cual queremos saber si tiene

una cierta distribución conocida, que denotaremos F_0 o no, es decir, la prueba que planteamos es

$$\begin{cases} H_0: & F_X = F_0 \\ H_1: & F_X \neq F_0 \end{cases} \tag{10.7}$$

Vamos a llamar p_i a la probabilidad de que X pertenezca al intervalo I_i si es cierto H_0 es decir $p_i = P_{H_0}(X \in I_i)$. Asumiremos que los intervalos son tomados de forma tal que $p_i > 0$ para todo i. Es importante observar además que como estamos suponiendo que F_0 la conocemos (por ejemplo es $N(118,25)$ los valores p_i se pueden calcular. Construimos ahora el *estimador* que usaremos, que será la generalización del T_n que definimos antes:

$$T_n = \sum_{i=1}^{k} \frac{\left(n \times p_i - [\text{nro de observaciones en } I_i]\right)^2}{n \times p_i}. \tag{10.8}$$

Nuevamente la región crítica es $RC = \{T_n > h\}$. Por último, para determinar h de modo tal que RC tenga nivel α usamos el resultado

$$P_{H_0}(T_n > h) \to P(\chi^2_{k-1} > h),$$

del cual se sigue que h es el valor (obtenido mediante el uso de la tabla de la distribución χ^2_{k-1}) que deja probabilidad α a la derecha de h, es decir:

$$\alpha = P(\chi^2_{k-1} > h).$$

Test χ^2 con parámetros estimados

Al igual que como hicimos con la prueba de Lilliefors para el caso de que en que la prueba (10.7) F_0 dependa de, por ejemplo r parámetros y estos se estimen por el método de máxima verosimilitud, el estadístico que se usa es el mismo T_n que antes, pero la distribución límite en el caso en que se hayan usado $k > r+1$ intervalos es una χ^2 con $k-1-r$ grádos de libertad.

En R

Para este test existe la función chisq.test, le tenemos que dar dos vectores: x=c(7,17,29,30,17), y por otro lado el vector de probabilidades p=c(0.0808,0.1935,0.305,0.2088,0.2119). Y hacemos chisq.test(x,p=p), esto nos devuelve

```
x=c(7,17,29,30,17)
p=c(0.0808,0.1935,0.305,0.2088,0.2119)
chisq.test(x,p=p)


##
##  Chi-squared test for given probabilities
##
```

```
## data:  x
## X-squared = 5.3155, df = 4, p-value = 0.2564
```

el valor del estadístico, en este caso devuelve X-squared=5.3155, los grados de libertad, en nuestro caso 4, y el p-valor, que para esta prueba da 0.2564. Como es mayor que α no rechazamos (ver la explicación de la implementación en R del test de Kolmogorov-Smirnov, para una explicación detallada del p-valor).

Chapter 11

Test de aleatoriedad

11.1 Introducción

Para calcular intervalos de confianza, hacer pruebas de hipótesis o pruebas de bondad de ajuste hemos asumido que los datos son independientes e idénticamente distribuidos. Sin dicha hipótesis muchos de los cálculos que hemos hecho hasta ahora no son válidos. En éste capítulo veremos primero algunos test que permiten verificar cuándo una muestra $X_1, \ldots, X_n$ cumple esas hipótesis. Veremos sólamente el test de Spearman, no obstante existen otros como el test de correlación de Pearson y el test de Rachas, que no daremos, pero cuya explicación se encuentra en el apéndice.

El segundo test, de Spearman, es similar, en el sentido de que permite testear si existe una relación monótona en la muestra (una tendencia creciente o decreciente) o entre dos muestras (al crecer una crece la otra). Bajo la hipótesis de que la muestra es iid (o si son 2 muestras, que son independientes), dicha relación no debería existir. Tanto el test de rachas como Spearman son formas de testear dependencia en la muestra o entre las muestras.

Al final del capítulo veremos la prueba χ^2 de independencia para datos categóricos, esto quiere decir que tenemos por ejemplo una población de n individuos a los cuales hemos medido ciertas características que representamos por dos variables aleatorias X e Y. Por ejemplo X puede ser el género del individuo (hombre o mujer), e Y si es fumador, si fumó y ya no lo hace, o si nunca fumó, y queremos determinar si hay dependencia o no entre sus hábitos como fumador o no fumador y su género. Se llaman categóricos porque la X representa una característica o categoría, y la Y otra, y se quiere ver si son independientes.

11.2 Test de Spearman

11.2.1 Test de Spearman de una muestra

La aleatoriedad esta relacionado con la presencia o no de comportamiento monótono en la muestra (es decir, si es creciente o no) ya que si es iid entonces no existe un comportamiento monótono. Esto se puede ver de manera intuitiva de la siguiente forma, si tenemos dos datos X_1 y X_2 de modo que son iid, el suceso $X_1 < X_2$ tiene la misma probabilidad que el suceso $X_2 < X_1$ ya que por ser iid son *intercambiables*. Eso implica en particular que cualquier ordenación de esos datos es igualmente probable, y lo mismo pasa si tenemos 3 o mas, es decir no deberíamos tener un comportamiento monótono en la muestra. Por lo tanto si construimos un estadístico que bajo H_0 (iid) da un valor próximo a 0, y que si hay comportamiento monótono toma un valor *grande*, podríamos decir que rechazamos H_0 si el estadístico supera un valor critico que tenemos que hallar. El estadístico de Spearman lo que hace es *medir* si hay o no un comportamiento monótono en la muestra. Por lo tanto bajo H_0 no deberíamos rechazar (ya que da valores próximos a 0). El problema que tiene es que una muestra puede *no ser* iid (es decir estamos bajo H_1), pero no tener comportamiento monótono, en cuyo caso el estadístico igual va a dar un valor próximo a 0, y por lo tanto no rechazaríamos H_0, y estaríamos decidiendo erroneamente que es iid cuando en realidad *no* lo es. Es decir la potencia que tiene este test no es muy alta en ciertos casos, o lo que es lo mismo, el error de tipo II tiene una probabilidad que puede no ser chica.

El estadístico de Spearman calcula el coeficiente de correlación de Pearson 6.3 entre los rangos de la muestra (la posición que ocupa cada dato en la muestra, ver 11.1) y el vector ordenado $(1,2,\ldots,n)$. Como el coeficiente de correlación de Pearson mide dependencia lineal, si el coeficiente de Pearson entre los rangos y el vector ordenado $(1,2,\ldots,n)$ es próximo a 1 o -1, lo que estamos diciendo es que la muestra original es monótona y por lo tanto no es iid.

Veamos como se construye el estadístico que usaremos. A partir de una muestra $X_1,\ldots,X_n$ ordenamos los datos y obtenemos $X_{(1)} \leq \cdots \leq X_{(n)}$. Por ejemplo supongamos que tenemos las siguientes mediciones de temperatura (en grados celsius),

1	22,67	6	17,88
2	21,66	7	23,17
3	16,31	8	24,85
4	15,95	9	15,17
5	15,15	10	23,19

los datos ordenados son

$X_{(1)} = 15.15$ $X_{(2)} = 15.17$ $X_{(3)} = 15.95$ $X_{(4)} = 16.31$ $X_{(5)} = 17.88$ $X_{(6)} = 21.66$ $X_{(7)} = 22.67$ $X_{(8)} = 23.17$ $X_{(9)} = 23.19$ $X_{(10)} = 24.85$

Construimos los estadísticos de rangos $R_1, \ldots, R_n$ donde

$$R_i = \sum_{j=1}^{n} \mathbb{I}_{\{X_j \leq X_i\}}. \tag{11.1}$$

Es decir, para cada dato X_i, R_i cuenta la cantidad de datos menores o iguales que X_i. En lo que sigue vamos a asumir que no hay datos repetidos en la muestra. Recordar que esto pasa si la variable X es absolutamente continua, es decir tiene densidad.

Como $X_1 = 22.67 = X_{(7)}$ tenemos que $R_1 = 7$, como $X_2 = 21.66 = X_{(6)}$ $R_2 = 6$ y así sucesivamente, obtenemos que

$$R_1 = 7, R_2 = 6, R_3 = 4, R_4 = 3, R_5 = 1, R_6 = 5, R_7 = 8, R_8 = 10, R_9 = 2, R_{10} = 9$$

Se puede demostrar aunque no lo haremos que el coeficiente de correlación de Pearson entre $R_1, \ldots, R_n$ y $(1, \ldots, n)$ se puede calcular como

$$\rho_s = 1 - \frac{6D}{n(n^2-1)},$$

con $D = \sum_{i=1}^{n} (R_i - i)^2$. En el caso de nuestro ejemplos es

$$\rho_s = 1 - \frac{6D}{10 \times 99} \tag{11.2}$$

y

$$D = (7-1)^2 + (6-2)^2 + (4-3)^2 + (3-4)^2 + (1-5)^2 + (5-6)^2 + (8-7)^2 + (10-8)^2 + (2-9)^2 + (9-10)^2 = 126$$

Es decir, $\rho_1 = 0.2364$.
Se sabe, aunque no lo veremos aquí, que ρ_s es una variable aleatoria discreta que tiene una distribución simétrica y toma valores entre -1 y 1. Se puede demostrar además que $E(\rho_s) = 0$ y $\text{Var}(\rho_s) = 1/(n-1)$. Definimos la región crítica $RC = \{(X_1, \ldots, X_n) : |\rho_s| > c\}$, tenemos que determinar (al igual que como hacíamos en el test de rachas con $\hat{R}_n$) el valor de c, que dependerá del nivel de significación de la prueba. Para valores de n menores o iguales que 10 existen tablas con la distribución de ρ_s que nos dan, para diferentes valores k_i de ρ_s entre 0 y 1 (en el caso de que el estadístico de un valor negativo usamos la simetría de ρ_s) la probabilidad de obtener un valor mayor o igual que k_i, es decir $P(\rho_s \geq k_i)$. En el caso por ejemplo de $n = 5$ la tabla es la siguiente

n	R	P
5	1.000	0.008
	0.900	0.042
	0.800	0.067
	0.700	0.117
	0.600	0.175
	0.500	0.225
	0.400	0.258
	0.300	0.342
	0.200	0.392
	0.100	0.475
	0.000	0.525

Esta tabla nos dice en particular que si $n = 5$ $P(\rho_s \geq 1) = 0.008$, $P(\rho_s \geq 0.900) = 0.042$, $P(\rho_s \geq 0.800) = 0.067$ etc. Si tomáramos $\alpha = 0.1$ y tuviéramos 5 datos, la región crítica es entonces $RC = [-1,-0.9] \cup [0.9,1]$. rechazamos a nivel α si nuestro estadístico ρ_s calculado mediante (11.2) nos da mayor o igual que 0.900 o menos que -0.900. En el caso de $n = 10$ y $\alpha = 0.1$ la región crítica es

$$RC = [-1;-0.564] \cup [0.564;1]$$

Como obtuvimos $\rho_s = 0.2364$ no estamos en la región crítica y por lo tanto no se rechaza la hipótesis de que los datos son $i.i.d$.

Ejercicio 11.1.

- Usando la tabla de ρ_s demostrar que la región crítica para $\alpha = 0.10$ y $n = 7$ datos es

$$RC = [-1;-0.714] \cup [0.714;1]$$

- Usando la tabla de ρ_s demostrar que la región crítica para $\alpha = 0.05$ y $n = 9$ datos es

$$RC = [-1;0.683] \cup [0.683;1]$$

De $n = 11$ a $n = 30$ la forma de la tabla cambia, veamos como es para algunos valores de n

n	0.100	0.050	0.025	0.010	0.005	0.001
11	0.427	0.536	0.618	0.709	0.764	0.855
12	0.406	0.503	0.587	0.678	0.734	0.825
30	0.241	0.307	0.363	0.426	0.467	0.548

Supongamos que $n = 12$, esta tabla nos dice que $P(\rho_s \geq 0.406) = 0.1$, $P(\rho_s \geq 0.503) = 0.05$,..., $P(\rho_s \geq 0.825) = 0.001$. Por lo tanto nuestra región crítica si $\alpha = 0.1$ es $RC = [-1;-0.503] \cup [0.503,1]$. Si fuese $n = 30$ y $\alpha = 0.1$ es $RC = [-1;-0.307] \cup [0.307;1]$ mientras que si $n = 30$ y $\alpha = 0.05$ la región crítica es

$RC = [-1;-0.363] \cup [0.363;1]$. Obsérvese que para el mismo α a medida que n crece la longitud de la región crítica crece, ¿por qué?.

Para valores grandes de n ($n > 30$) usamos que, si la muestra $X_1, \ldots, X_n$ es iid,

$$P\left(\sqrt{n-1}\rho_s \leq t\right) \to \Phi(t)$$

por lo tanto vamos a rechazar que la muestra es iid si

$$\left|\sqrt{n-1}\rho_s\right| > z_{1-\alpha/2}.$$

En R

Para realizar el test de Spearman de 1 muestra primero tenemos que crear un vector con los datos, en nuestro ejemplo anterior creamos t=c(22.67,21.66,16.31,15.95,15.15,17.88,23.17,24.85,15.17,23.19). Luego ejecutamos el comando cor.test(t,sort(t),method="spearman"), eso da el valor del estadístico ρ_s, en nuestro caso da 0.2363636 el valor de D (que en R se llama S), en nuestro caso 126. Y el p-valor para el test, que da 0.5139. Como ya sabíamos, no se rechaza H_0.

```
t=c(22.67,21.66,16.31,15.95,15.15,17.88,23.17,24.85,15.17,23.19)
cor.test(t,sort(t),method="spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  t and sort(t)
## S = 126, p-value = 0.5139
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2363636
```

11.2.2 Test de Spearman de dos muestras

El coeficiente de correlación de Pearson (ver 6.3) mide en cierto modo si existe o no una relación lineal entre las variables X e Y, en el caso de máxima correlación una es una función lineal de la otra. El coeficiente de correlación se Spearman lo que hace es medir si existe o no una relación mónotona (lineal o no) entre las variables (y al igual que en el caso de 1 muestra, si son independientes dicha relación no debería existir). Esto significa que si tenemos dos muestras $X_1, \ldots, X_n$ de X e $Y_1, \ldots, Y_n$ de Y, valores grandes de X_i se corresponden con valores grandes de Y_i. Y valores chicos de X_i con valores chicos de Y_i. Si pasa eso lo que sucede es que

los estadísticos de rangos de los X_i, definidos como en 11.1, están en relación lineal con los estadísticos de rango de los Y_i. Es decir calculamos para $i = 1, \ldots, n$

$$R_i^X = \sum_{j=1}^{n} \mathbb{I}_{\{X_j \leq X_i\}} \qquad R_i^Y = \sum_{j=1}^{n} \mathbb{I}_{\{X_j \leq X_i\}} \tag{11.3}$$

y luego calculamos el coeficiente de correlación de Pearson estimado (definido en (7.4)) $\hat{\rho}(R^X, R^Y)$ de los vectores $(R_1^X, \ldots, R_n^X)$ y $(R_1^Y, \ldots, R_n^Y)$. Se puede ver que esa cuenta es exactamente igual a calcular

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n}(R_i^X - R_i^Y)^2}{n(n^2-1)}. \tag{11.4}$$

Bajo la hipótesis nula de X e Y son independientes el estadístico ρ_s no depende de las distribuciones de los datos (ya que cualquiera de las $n!$ posibles permutaciones de los rangos es igualmente probable). Nuevamente la región crítica es $RC = \{|\rho_s| > t_n\}$. Para valores de n grandes, se puede probar nuevamente que $P_{H_0}(\sqrt{n-1}\rho_s > t) \to \Phi(t)$ y por lo tanto vamos a rechazar H_0 si

$$\left|\sqrt{n-1}\rho_s\right| > z_{1-\alpha/2}.$$

En R

Para realizar el test de Spearman de 2 muestra se procede igual al de una muestra, creamos dos vectores, u y t con los datos. Ahora tenemos que ejecutar el comando `cor.test(t,u,method="spearman")` obtenemos el valor del estadístico, y el p-valor para el test.

11.3 Prueba χ^2 de independencia, cuadro de contingencia

Como dijimos al comienzo del capítulo lo que queremos es determinar si existe o no relación de dependencia entre ciertas características de los datos. En este contexto lo que tenemos son n pares $(X_1, Y_1), \ldots (X_n, Y_n)$ donde las variables X_i toman una cantidad finita de valores $c_1, \ldots, c_r$ y las variables Y_j toman otra cantidad finita $d_1, \ldots, d_2$. Queremos determinar si las características c son independientes o no de las d. La hipótesis nula H_0 que planteamos es que son independientes, y la alternativa que no. A modo de ejemplo c_1 y c_2 es el género de la persona (hombre o mujer) y d_1 es si fuma actualmente, d_2 es si fumaba y ya no lo hace, y d_3 si nunca fumó. Si tenemos 402 individuos de los cuales 193 son hombres y 209 son mujeres, primero se construye la tabla siguiente, que se denomina tabla de contingencia,

$Y\|X$	Hombre	Mujer	Total
Nunca fumó	149	148	297
Fue fumador y no fuma	13	24	37
Fuma actualmente	31	37	68
Total	193	209	402

Esta tabla nos dice por ejemplo que 13 de los 193 hombres fueron fumadores pero no fuman actualmente, y que 37 de las 209 mujeres fuma actualmente, mientras que 149 nunca fumó. Bajo la hipótesis de independencia entre los hábitos como fumador o no fumador y el género debería verificarse que, por ejemplo 149/402 que es la probabilidad de que nunca haya fumado *y* sea hombre sea aproximadamente 297/402 que es la probabilidad de que nunca haya fumado multiplicado por 193/402 que es la probabilidad de que sea hombre, es decir la probabilidad de la intersección es el producto de las probabilidades. Y eso para cualquier otra entrada de la matriz 3×2 anterior. El estadístico que se plantea tiene en cuenta estas diferencias de la siguiente forma: se calculan 6 sumandos uno por cada entrada de la matriz:

$$\frac{\left(\frac{149}{402}-\frac{193}{402}\frac{297}{402}\right)^2}{\frac{193}{402}\frac{297}{402}}+\frac{\left(\frac{148}{402}-\frac{209}{402}\frac{297}{402}\right)^2}{\frac{209}{402}\frac{297}{402}}+\frac{\left(\frac{13}{402}-\frac{37}{402}\frac{193}{402}\right)^2}{\frac{37}{402}\frac{193}{402}}+\frac{\left(\frac{24}{402}-\frac{37}{402}\frac{209}{402}\right)^2}{\frac{37}{402}\frac{209}{402}}+\frac{\left(\frac{31}{402}-\frac{68}{402}\frac{193}{402}\right)^2}{\frac{68}{402}\frac{193}{402}}+\frac{\left(\frac{37}{402}-\frac{68}{402}\frac{209}{402}\right)^2}{\frac{68}{402}\frac{209}{402}}$$

Este cálculo da 0.00789. Observemos que lo que se hizo es a cada entrada de la matriz 3×2 anterior se divide entre el total, este valor lo llamamos o_{ij} , con $i = 1,2,3$ filas y $j = 1,2$ columnas, se le resta el valor

$$e_{ij} = \frac{(\text{ total de la fila } i)\times(\text{total de la columna } j)}{(\text{total de datos})^2}$$

se eleva al cuadrado, es decir tomamos $(o_{ij}-e_{ij})^2$ y dividimos entre e_{ij}. Luego se suman los 6 términos. Esto da el estadístico

$$T_n = \sum_{i=1}^{\text{nro filas}}\sum_{j=1}^{\text{nro col}}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

Por lo tanto vamos a rechazar la hipótesis nula de que son independientes si T_n supera un valor crítico t_n. Es decir la región crítica es $RC = \{T_n > t_n\}$; para hallar t_n tenemos que usar un resultado teórico que dice que, si denotamos n al total de datos (en el ejemplo $n = 402$), R al numero de filas de la matriz (en nuestro ejemplo $R = 3$) y C al número de columnas (en el ejemplo $C = 2$), bajo la hipótesis H_0 de que X e Y son independientes, para todo $t > 0$,

$$P_{H_0}(nT_n > t) \to P(\chi^2_{(R-1)(C-1)} > t).$$

Es decir la distribución asintótica (para valores de *n grandes*) de nT_n es una χ^2 con $(R-1)\times(C-1)$ grados de libertad. Finalmente fijado $\alpha \in (0,1)$, hacemos $P(\chi^2_{(R-1)(C-1)} > t) = \alpha$, y despejamos t de una tabla o usando R. En nuestro ejemplo tenemos que calcular t tal que $P(\chi^2_2 > t) = 0.05$. En R si hacemos qchisq(0.95,2) da 5.99146. Por lo tanto el valor $t_n = 5.99146/402 = 0.01490$. Como $0.00789 < 0.01490$ no estamos en la región crítica y por lo tanto no se rechaza H_0 a nivel 0.05.

En R

En R tenemos que ingresar la matriz como tabla, y luego usar el comando chisq.test, esto da el estadístico nT_n y el p-valor.

```r
datos=as.table(matrix(c(149,13,31,148,24,37),ncol=2))
chisq.test(datos)
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos
## X-squared = 3.1713, df = 2, p-value = 0.2048
```

Chapter 12

Regresión lineal

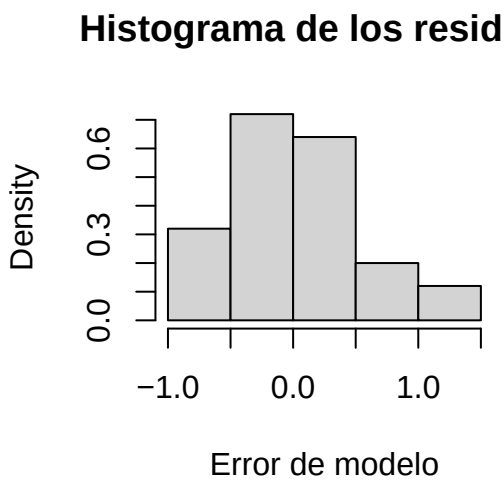
12.1 Mínimos cuadrados

En muchos problemas surge la necesidad de entender la dependencia entre dos o más variables. Por ejemplo, si sabemos que X es una variable aleatoria muy relacionada a Y, podemos postular un modelo lineal de la forma $Y = aX + b$, donde a y b son constantes que en general desconocemos y queremos estimar. En dicho caso, se ve una dependencia absoluta entre las variables X e Y (o sea, sabiendo el valor de X y las constantes a y b podemos calcular sin error el valor de Y). Sin embargo, en general, el efecto de la variable X sobre la Y no sigue este modelo perfectamente lineal sino que aparece un error ε (el cual es desconocido, y puede deberse a errores de medición). Es decir estamos ante un modelo que podemos escribir como $Y = aX + b + \varepsilon$.

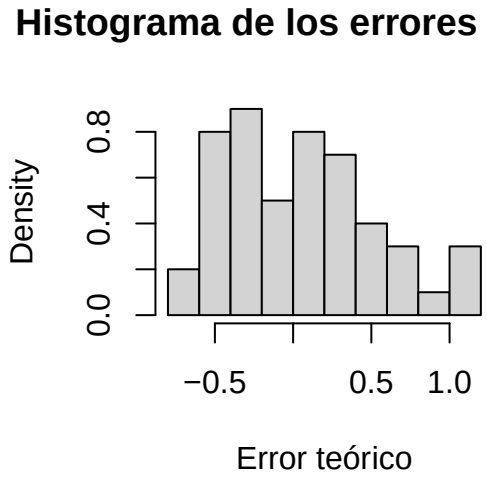
Vamos a estudiar el caso en el que tenemos valores de X que los pensamos como no aleatorios, cada uno de los cuales tiene asociado un valor de la variable Y, esto se conoce como *modelo de efectos fijos* (en contraposición con el *modelo de efectos aleatorios*, donde la X se considera aleatoria). Este nombre se debe a que el efecto de la variable X sobre la Y se puede considerar fijo, por ejemplo X puede ser el género de un individuo, su edad, su peso, mientras que Y puede ser su salario en $. Suponer efectos fijos tiene que ver con el diseño del experimento en si, con como son los datos, etc, muchas veces es una hipótesis poco razonable, y hay que pensar las X como aleatorias con cierta distribución (en estos casos se asume generalmente que X y ε son independientes). En nuestro caso la X es fija y el error ε lo supondremos con distribución normal con media 0 y varianza σ^2. En general las constantes a y b no se conocen y lo que se dispone es de una muestra $(X_1, Y_1), \ldots, (X_n, Y_n)$ de pares que se asumen independientes e idénticamente distribuidos (con la misma distribución del par (X, Y), y a partir de ellos escribimos entonces que $Y_i = aX_i + b + \varepsilon_i$ para todo $i = 1, \ldots, n$. Más adelante vamos a proponer un estimador $\hat{a}$ de a, y $\hat{b}$ de b basados en dicha muestra. En este caso, el valor que "predice" nuestro modelo con los parámestros estimados, para un dato X_i es $\hat{Y}_i = \hat{a}X_i + \hat{b}$. Observar que aquí el error ε_i no aparece, ya que en general dicho error no se conoce, y sólo se dispone de los pares (X_i, Y_i). La diferencia entre el verdadero valor Y_i y el valor estimado $\hat{Y}_i$ se conoce como residuo i-ésimo.

En la figura 12.1 se muestra el gráfico de los pares (X_i, Y_i) con $i = 1, \ldots, 20$.

```
hist(y - a_hat*x - b_hat, main = "Histograma de los residuos",
     freq = FALSE, xlab = "Error de modelo")
```




```
hist(e, main = "Histograma de los errores teóricos",
     freq = FALSE, xlab = "Error teórico")
```



Matricialmente podemos escribir $\mathcal{Y} = \mathcal{X}(b,a)^T + e$, donde $\mathcal{X}(b,a)^T$ denota el producto de la matrix $\mathcal{X}$ por el vector columna $(b,a)^T$, que es el transpuesto del vector fila (b,a), multiplicado por la matriz con n

filas y 2 columnas $\mathcal{X}$, dada por

$$\mathcal{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_n \end{pmatrix}.$$

El vector $\mathcal{Y} = (Y_1, \ldots, Y_n)^T$, y el vetor $e = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Observar que si llamamos $\mathbf{1}_n$ a la columna 1 de la matriz $\mathcal{X}$ (formado por el vector de n unos), y c_2 al segundo vector de la matriz $\mathcal{X}$, es decir $c_1 = (X_1, \ldots, X_n)$ el producto $\mathcal{X}(b,a)^T$ da el vector $b\mathbf{1}_n + ac_2$. Es decir, $\mathcal{X}(b,a)^T$ es un vector que pertenece al subespacio espacio vectorial de dimensión 2, incluido en $\mathbb{R}^n$, generado por los vectores $\mathbf{1}_n$ y c_2. Estos vectores asi como el espacio generado por las columnas de $\mathcal{X}$ se respresentan en la Figura (12.1)

Una manera de estimar los parámetros a y b es hallar el par $(\hat{a}_n, \hat{b}_n)$ que minimicen la expresión

$$S(a,b) = \sum_{i=1}^{n} (Y_i - aX_i - b)^2 \tag{12.1}$$

El cuadrado en la expresión $(Y_i - aX_i - b)^2$ lo que hace es darle más peso a valores grandes de $(Y_i - aX_i - b)$. Por otra parte observar que la expresión $(Y_i - aX_i - b)$ lo que calcula es la distancia en vertical entre la recta $ax + b$ y el dato Y_i. La suma en 12.1 se conoce como *suma de los cuadrados de los residuos*. Para resolver el problema de minimización 12.1 vamos a proceder primero razonando de manera geométrica, y luego resolviendo el problema de minimización en dos variables.

Para resolver el problema de minimización de $S(a,b)$, primero hay que notar que $S(a,b)$ es la norma ecuclideana, al cuadrado, del vector cuyas coordenadas son $Y_1 - aX_i - b$. Este vector es, como vimos $b\mathbf{1}_n + ac_2$. Por lo tanto queremos el vector, llamemosle v, en el plano representado en la Figura 12.1, que minimiza la norma al cuadrado de $\mathcal{Y} - v$. Se puede ver fácilmente, usando el Teorema de Pitágoras, que el vector v que queremos es la proyección ortogonal del vector $\mathcal{Y}$ sobre dicho plano. Es decir $Y - v$ es un vector perpendicular al plano, es decir tiene que cumplir que es perpendicular a todos los vectores del plano. En particular es ortogonal a c_2 y a $\mathbf{1}_n$. En forma matricial esto se plantea como el problema de hallar $v = \mathcal{X}(b,a)^T$ tal que $\mathcal{X}^T(\mathcal{Y} - \mathcal{X}(b,a)^T) = 0$, es decir

$$\begin{pmatrix} 1 & \ldots & 1 \\ X_1 & \ldots & X_n \end{pmatrix} \left(\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} - \begin{pmatrix} b & aX_1 \\ b & aX_2 \\ \vdots & \\ b & aX_n \end{pmatrix} \right) = 0$$

De la ecuación $\mathcal{X}^T(\mathcal{Y} - \mathcal{X}(b,a)^T) = 0$ obtenemos $\mathcal{X}^T\mathcal{Y} = \mathcal{X}^T\mathcal{X}(b,a)^T$. Si ahora multiplicamos ambos lados de esta igualdad por la inversa de la matriz $\mathcal{X}^T\mathcal{X}$ (la cual supondremos que es invertible), obtenemos que

$$(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathcal{Y} = (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathcal{X}(b,a)^T$$

El producto matricial $(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathcal{X}$ a la derecha de la ecuación anterior da la identidad, ya que estamos multiplicando una matriz por su inversa. Finalmente obtuvimos que el vector (a,b) se puede escribir como

$(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathcal{Y}$. Para hallar este producto, veamos el producto matricial que escribimos antes. Observemos que $\mathcal{X}^T\mathcal{Y}$ es un vector de dos dimensiones, igual a $(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i X_i)$. Mientras que la matriz $(\mathcal{X}^T\mathcal{X})$ es una matriz 2x2 dada por

$$\begin{pmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{pmatrix}.$$

Se deja como ejercicio verificar que la inversa de dicha matriz es

$$\frac{1}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2}\begin{pmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{pmatrix}.$$

Observar que esta verificación se hace simplemente multiplicando esta matriz por la anterior, y viendo que dicho producto da la matrix identidad 2x2. Si multiplicamos esta matriz por el vector $\mathcal{Y}$ obtenemos que

$$\hat{a} = \frac{-\sum_{i=1}^n Y_i \sum_{i=1}^n X_i + n\sum_{i=1}^n Y_i X_i}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2}$$

si multiplicamos en esta expresión el numerador y el denominador por $(1/n)$ obtenemos que

$$\hat{a} = \frac{\sum_{i=1}^n Y_i X_i - \sum_{i=1}^n Y_i \sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2 - (1/n)(\sum_{i=1}^n X_i)^2}.$$

Se deja como ejercicio verificar que

$$\hat{b} = \frac{(\sum_{i=1}^n X_i^2)(\sum_{i=1}^n Y_i) - (\sum_{i=1}^n X_i Y_i)(\sum_{i=1}^n X_i)}{\sum_{i=1}^n X_i^2 - (1/n)(\sum_{i=1}^n X_i)^2}.$$

12.2 Cálculo de $\hat{a}$ y $\hat{b}$ mediante derivadas

Lo que hacemos es derivar $S(a,b)$ respecto de a y respecto de b e igual a cero. Esto nos da

$$\frac{\partial}{\partial a}S(a,b) = -2\sum_{i=1}^n (Y_i - aX_i - b)X_i = 0$$

y

$$\frac{\partial}{\partial b}S(a,b) = -2\sum_{i=1}^n (Y_i - aX_i - b) = 0$$

De la segunda ecuación si despejamos b, obtenemos

$$\hat{b}_n = \frac{1}{n}\sum_{i=1}^n (Y_i - aX_i) = \overline{Y_n} - a\overline{X_n} \tag{12.2}$$

donde $\overline{Y_n} = (1/n)\sum_{i=1}^n Y_i$ y $\overline{X_n} = (1/n)\sum_{i=1}^n X_i$. Si sustituimos este valor en la primera ecuación nos da

$$\hat{a}_n = \frac{\sum_{i=1}^n X_i Y_i - (1/n)(\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)}{\sum_{i=1}^n X_i^2 - (1/n)(\sum_{i=1}^n X_i)^2}$$

Luego este valor se inserta en la ecuación 12.2 y se obtiene

$$\hat{b}_n = \overline{Y_n} - \hat{a}_n \overline{X_n}.$$

Calculando la matriz Hessiana se puede ver que la misma es definida positiva y por lo tanto $S(a,b)$ posee un mínimo global que es el hallamos antes.

Veamos una interpretación del estimador $\hat{a}_n$ para el caso en que X es aleatoria. Primero observemos que, usando las propiedades de la covarianza y que ε es independiente de X, $\text{cov}(X,Y)/\text{var}(X) = \text{cov}(X, aX+b+\varepsilon)/\text{var}(X) = a\text{cov}(X,X)/\text{Var}(X) = a$. Por otra parte recordemos que $\text{cov}(X,Y) = E[(X - E(X))(Y - E(Y))]$, y un estimador de dicho valor es

$$\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n) = \frac{1}{n}\sum_{i=1}^n X_i Y_i - \overline{X}_n \overline{Y}_n$$

Este valor es exactamente el mismo que el numerador de $\hat{a}_n$ dividido por n (verificarlo!). Por otra parte el denominador de $\hat{a}_n$ divido n es un estimador de la varianza de X. Finalmente obtuvimos que $\hat{a}_n$ es una estimación de $\text{cov}(X,Y)/\text{var}(X)$ que ya vimos que es igual a a. Veamos que pasa con $\hat{b}_n$. Sabemos que $\overline{Y}_n \to E(Y)$ y $\overline{X}_n \to E(X)$ por la ley de los grandes números, por otra parte de la ecuación $Y = aX + b + \varepsilon$ usando que $E(\varepsilon) = 0$. Obtenemos que $E(Y) = aE(X) + b$. Por lo tanto $b_n \to E(Y) - aE(X) = b$.

12.2.1 En R

Si hacemos un summary de la regresión que mostramos en la Figura 12.1 obtenemos

```
summary(regresion)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86102 -0.35635 -0.06449  0.28149  1.13252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.05056    0.07028   29.18   <2e-16 ***
## x            1.03857    0.06246   16.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.495 on 48 degrees of freedom
## Multiple R-squared:  0.8521,Adjusted R-squared:  0.849
## F-statistic: 276.5 on 1 and 48 DF,  p-value: < 2.2e-16
```

Esto devuelve primero un summary de los residuos (recordar que los residuos son $Y_i - \hat{Y}_i$ con $\hat{Y} = \hat{a}X_i + \hat{b}$).

Y luego la estimación de los coeficientes en la columna estimate $\hat{b}$ (intercept) $\hat{a}$ (llamado x). Más adelante veremos que son los otros valores que devuelve.

12.3 Significación del Modelo

Veamos una prueba de hipótesis para contrastar la existencia o no de una relación lineal entre la X y la Y, para el caso de efectos fijos y *errores normales independientes, todos ellos con la misma varianza* $\sigma^2 < \infty$ y esperanza 0. En términos del modelo esto se traduce en testear $H_0 : a = 0$, contra $H_1 : a \neq 0$. Si se rechaza H_0 a un cierto nivel α estamos diciendo que hay indicios de que exista una relación lineal entre las variables explicativas X y la Y. Denotemos $\hat{Y}_i = \hat{a}_n X_i + \hat{b}_n$, observar que este es el valor de la estimación de Y_i que da el modelo con los parámetros estimados.

Para eso se usa el siguiente resultado, que no demostraremos: bajo H_0,

$$\Gamma_n = \frac{\hat{a}_n}{\sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{(n-2)\sum_{i=1}^n (X_i - \overline{X}_n)^2}}} \sim T_{n-2}, \tag{12.3}$$

donde T_{n-2} es la distribución T de student con $n-2$ grados de libertad. Una vez calculado el valor del estadístico Γ_n, se rechaza H_0 a nivel α, si $|\Gamma_n| > t_{1-\alpha/2}(n-2)$ donde $t_{1-\alpha/2}(n-2)$ es es valor que hace que $P(T_{n-2} > t_{1-\alpha/2}(n-2)) = \alpha/2$.

El valor de T_{n-2} es el que aparece en la segunda fila de la columna t-value del summary mientras que el p-valor es el que se obtiene en la segunda fila de la columna $P(>|t|)$. Observamos en el summary anterior que el p valor para la prueba $H_0 : a = 0$ contra $H_1 : a \neq 0$ da muy pequeño y por lo tanto se rechaza H_0 a nivel 0.05 por ejemplo (esto quiere decir que hay una clara relación lineal entre las variables, como se mostraba en la figura 12.1).

12.4 Coeficiente de Determinación

El coeficiente de determinación, conocido como R^2, da a grandes rasgos una idea de que tanto explica el modelo estimado, construido con los parámetros $\hat{a}_n$ y $\hat{b}_n$. Para eso vamos a introducir la suma de cuadrados totales:

$$SCT = \sum_{i=1}^n (Y_i - \overline{Y}_n)^2,$$

y la suma de cuadrados de la regresión.

$$SC_{Reg} = \sum_{i=1}^n (\hat{Y}_i - \overline{Y}_n)^2.$$

Observar que $(1/n)SCT$ es un estimador de la varianza de Y. Finalmente la suma de cuadrados de los errores es

$$SC_{Err} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2,$$

donde hemos usado nuevamente la notación $\hat{Y}_i = \hat{a}_n X_i + \hat{b}_n$, que corresponden a los valores ajustados por el modelo con coeficientes estimados. El valor $Y_i - \hat{Y}_i$ se conoce como la variación no explicada por el modelo lineal (es el error que cometemos al estimar Y_i usando el modelo con los parámetros estimados) que hemos ajustado, mientras que $\hat{Y}_i - \overline{Y}_n$ es la variación explicada. Otra forma de entender esto es la siguiente, si aproximamos las Y_i con el modelo mas simple (es decir el modelo constante) es razonable tomar la constante $\overline{Y}_n$. Por lo tanto la cantidad $\hat{Y}_i - \overline{Y}_n$ mide cuan lejos esta nuestro modelo estimado del modelo constante, mientras que $Y_i - \overline{Y}_n$ mide cuan lejos están los verdaderos datos del modelo constante. Se puede demostrar que $SCT = SC_{Reg} + SC_{Err}$, es decir la variación total es la suma de la variación explicada mas las variaciones en el error (variación no explicada en la regresión). El R^2 no es otra cosa que

$$R^2 = \frac{SC_{Reg}}{SCT} = 1 - \frac{SC_{Err}}{SCT}$$

Observar que dicho cociente es mayor o igual que 0 (es cociente de cantidades que son positivas) y menor o igual que 1 (esto se sigue fácilmente de la igualdad $SCT = SC_{Reg} + SC_{Err}$), y si nuestro modelo estimado estima perfectamente (sin error) los parámetros, dicho cociente vale uno ya que $Y_i = \hat{Y}_i$. Es decir el R^2 es el cociente entre la variación explicada por el modelo, y la variación total. Cuanto mayor es el valor de R^2 mejor es el ajuste de la regresión lineal a los datos. En el summary que vimos antes, este valor se muestra como Multiple R-squared. Se puede demostrar, aunque no lo haremos, que el coeficiente de determinación R^2 es igual al cuadrado del coeficiente de correlación empírico $(\hat{\rho}(\hat{Y}, Y))^2$ calculado con la ecuación (7.4) donde $\hat{Y} = \hat{Y}_1, \ldots, \hat{Y}_n$ son los valores estimados por el modelo (para eso se usa que $\overline{\hat{Y}} = \overline{Y}_n$). Esta observación da una interpretación más del R^2 como medida de ajuste del modelo lineal a los datos si tenemos en cuenta que el coeficiente de correlación para el caso de variables independientes es 0 (que correspondería a decir que el modelo estimado $\hat{Y}$ es independiente de Y y por lo tanto el ajuste lineal no sirve) y que es 1 o -1 si una variable es un múltiplo de la otra (que en nuestro caso implica $\hat{Y} = Y$ y por lo tanto el ajuste es perfecto).

12.5 Ejemplo en R: Datos reales

A modo de ejemplo vamos a usar el dataset llamado airquality que viene ya en R. El cual consiste en la medición de 4 variables atmosféricas (temperatura, radiaci'on solar, nivel de ozono y velocidad del viento) en distintos días y lugares de la ciudad de Nueva York, tomados en 1973. Para obtener más información sobre los mismos se puede ejecutar.

```
?airquality
```

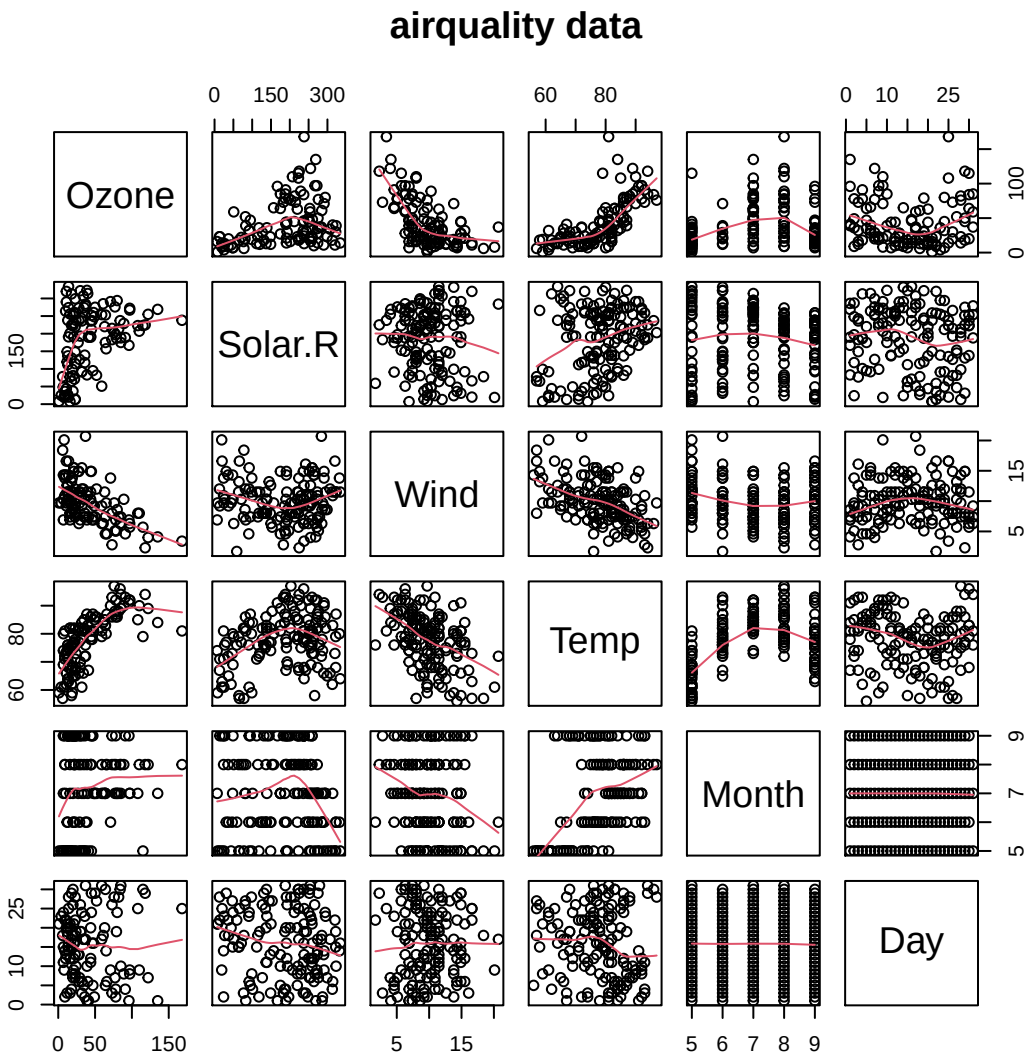
Es importante tener en cuenta las unidades en las que fueron medidas cada una de las variables a considerar. Para este ejemplo trabajaremos con el viento (que está en millas por hora según arroja el comando

anterior) y la temperatura (en grados Farenheit). Dado que estos datos contienen algunas filas con datos faltantes (que se representan como NA en la tabla) lo que haremos para simplificar el análisis es quedarnos con aquellas que no. Esto se hace mediante

```
datos<-airquality[complete.cases(airquality),]
```

Ahora trabajaremos con el objeto datos que consiste en una lista de 111 observaciones de 6 variables. Antes de hacer la regresión en si, es bueno tener una idea de como son los datos, a lo mejor hay que pre procesarlos (por ejemplo pasarlos a una escala logarítima, exponencial etc). Para visualizarlos una forma práctica es ejecutar

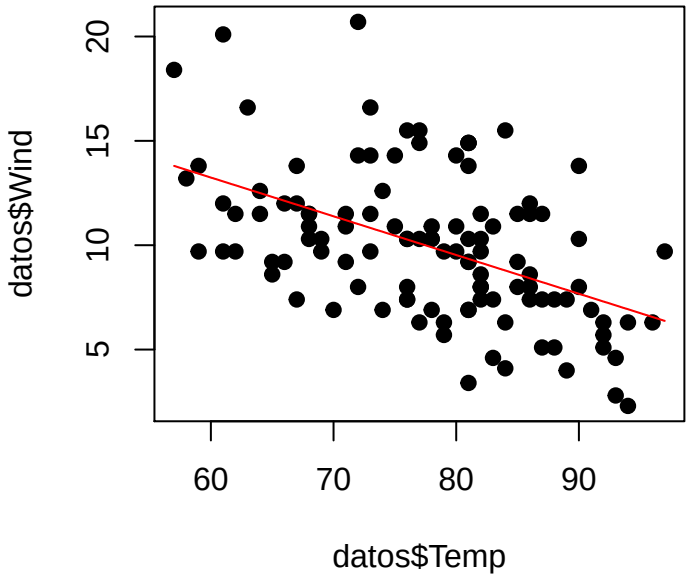
```
require(graphics)
pairs(airquality, panel = panel.smooth, main = "airquality data")
```



Esto grafica los puntos de a pares, y nos puede dar una idea de si existe o no una relación lineal entre las variables. Si queremos por ejemplo hacer una regresión lineal de la variable viento respecto de la temperatura escribimos simplemente

```
lineal<-lm(Wind~Temp,datos)
plot(datos$Temp, datos$Wind, pch = 19)
ext1 = min(datos$Temp)
ext2 = max(datos$Temp)
eje_x = seq(ext1, ext2, length.out = 10000)
```

```
a_hat = lineal$coefficients[2]
b_hat = lineal$coefficients[1]
lines(eje_x, a_hat*eje_x + b_hat, col = "red")
```




```
summary(lineal)
```

```
##
## Call:
## lm(formula = Wind ~ Temp, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9443 -2.3584 -0.3005  1.6136  9.6852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.37877    2.43137  10.027  < 2e-16 ***
## Temp        -0.18561    0.03102  -5.983 2.84e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.101 on 109 degrees of freedom
## Multiple R-squared:  0.2472,Adjusted R-squared:  0.2403
## F-statistic: 35.79 on 1 and 109 DF,  p-value: 2.842e-08
```

Esto devuelve varios valores de interés. Lo primero es el summary de los residuos $\hat{Y}_i - Y_i$ (los cuartiles). Observar que el valor esperado de los residuos es 0 por lo tanto la mediana debería dar próximo a 0 si se cumplen las hipótesis de nuestro modelo. Los valores que se obtienen de $\hat{a}_n$ y $\hat{b}_n$ están en la columna Intercept

y son -0.18 y 24.3 respectivamente. Esto significa que la pendiente de la recta de regresión es negativa. Es decir un aumento de la temperatura *disminuye* el valor medio del viento ya que el valor. Además se sigue que aumentar en una unidad la temperatura (en este caso 1 grado Farenheit) produce una diminución de 0.18 millas por hora, en la velocidad media de los vientos. No es lo mismo si hubiera sido -1.8 por ejemplo. Por otro lado el valor 24,3 es el coeficiente $\hat{b}_n$ que nos da el valor de la intersección de la recta en 0 (nos estaría diciendo que nuestro modelo, no el real sino el que hallamos, estima la velocidad media de los vientos en 24.3 si la temperatura es 0). Los valores 2.43137 y 0.03102 son estimaciones de la raiz de la varianza de $\hat{b}_n$ y $\hat{a}_n$ respectivamente (las fórmulas para obtener estos valores estan en el apéndice). La última columna da el p-valor para la pruebas $H_0 : b_n = 0$ contra $H_1 : b_n \neq 0$ (primera fila de dicha columna) y para la prueba $H_0 : a_n = 0$ contra $H_1 : a_n \neq 0$ (segunda fila). El primero de dichos valores da casi 0, menor a 2^{-16}), con lo cual rechazamos a nivel por ejemplo 0.05 la hipótesis de que $b_n = 0$ (y de hecho se rechaza a casi cualquier nivel razonable). El otro p valor es pequeño también, por lo tanto también rechazamos $H_0 : a_n = 0$, esto último nos dice que hay una relación lineal entre las variables viento y temperatura. Finalmente el valor que obtenemos para el R^2 es el que se denomina Multiple R-squared y da 0.2472. Esto nos dice que la variación explicada por el modelo es aproximadamente 1/4 de la variación total, con lo cual el ajuste no es muy bueno.

12.6 Ejemplos en R: Casos simulados

La figura 12.1 nos mostraba un ejemplo ideal de modelo lineal univariado. La relación $Y = aX + b + e$ que buscamos en nuestros datos es exactamente la relación que existe entre las variables X_i y las variables Y_i. Como además el valor de σ es chico en relación a la cantidad de datos que disponemos, podemos esperar que las estimaciones $\hat{a}$ y $\hat{b}$ sean buenas.

Sin embargo, si los datos no siguen exactamente el modelo impuesto, no podemos esperar que el modelo estimado que obtenemos sea fiable. A continuación, veremos a través de ejemplos, el impacto que pueden tener algunas hipótesis que no se verifiquen. Recordemos que nuestro modelo es

$$\begin{aligned} Y_i &= aX_i + b + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma^2), \text{iid} \\ a, b &\in \mathbb{R} \end{aligned}$$

Comencemos estudiando qué sucede si tenemos pocos datos y σ es muy grande como en 12.2

```
summary(regresion)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.357  -8.444   2.154   9.114  17.531
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.356      1.402   1.680   0.0995 .
## x              2.324      1.317   1.765   0.0839 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.828 on 48 degrees of freedom
## Multiple R-squared:  0.06094,Adjusted R-squared:  0.04137
## F-statistic: 3.115 on 1 and 48 DF,  p-value: 0.08394
```

Está claro que, si disponemos de pocos datos, o si las X_i toman valores cercanos, un valor grande de σ puede oscurecer la relación lineal existente entre las X_i y las Y_i.

Veamos qué sucede si la relación entre X_i y Y_i no es lineal como en 12.3

```
summary(regnolin)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3704 -1.1779 -0.1220  0.7782  3.3668
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2364     0.1948  16.616   <2e-16 ***
## x             0.3538     0.1840   1.923   0.0604 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.354 on 48 degrees of freedom
## Multiple R-squared:  0.07154,Adjusted R-squared:  0.05219
## F-statistic: 3.698 on 1 and 48 DF,  p-value: 0.06041
```

En teoría, para el rango en el que se encuentran las X_i y su relación cuadrática con las Y_i, debería suceder que $\hat{a} = 0$. Sin embargo, esto va a depender de dónde se encuentren las X_i, y de los valores de e_i.

Veamos ahora el caso de los *outliers*, o sea, datos atípicos o erroneos en nuestra base de datos como en 12.4

```
summary(regconoutlier)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8385 -1.7402 -0.3724  1.7124  4.9530
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.57317    0.36688   4.288 8.66e-05 ***
## x           -1.00323    0.02587 -38.785  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.57 on 48 degrees of freedom
## Multiple R-squared:  0.9691,Adjusted R-squared:  0.9684
## F-statistic:  1504 on 1 and 48 DF,  p-value: < 2.2e-16
```

Podemos ver que el peso en el modelo del outlier es mucho mayor al del resto de los datos. Esto se debe a que, en la ecuación 12.1, el sumando del error cuadrático del outlier es mucho más sensible a las elecciones de $\hat{a}$ y $\hat{b}$ que el resto de los sumandos.

Por último, hay que estudiar qué sucede para casos no normales. Vale observar que, para nuestro modelo teórico, se tiene que

$$Y_i|X_i, \sigma \sim N(aX_i + b, \sigma^2),$$

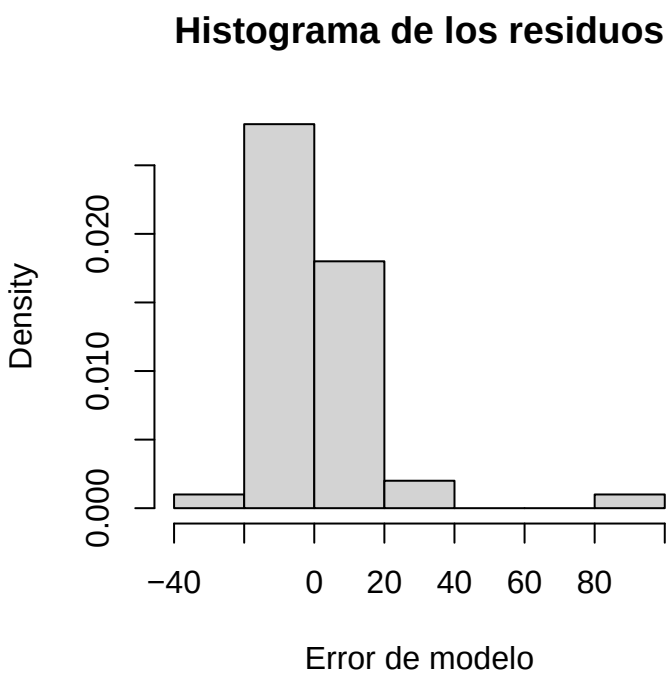
o sea, las Y_i son variables aleatorias normales condicionadas a X_i y σ. ¿Pero qué sucede si se cambia la relación normal por otro tipo de relación?

Podemos ver en 12.5 que los errores $Y_i - X_i$ tienen media cero, pero su tercer momento no es cero (indicando que la variable aleatoria no es simétrica respecto a su media). Esto juega un factor importante, ya que el modelo lineal será sensible a estas asimetrías (de manera similar al caso de los outliers en la figura 12.4).

```
summary(regresion)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.260  -8.656  -2.227   2.770  82.977
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.9234     4.0961  -0.470 0.640789
## x             2.9777     0.7126   4.179 0.000123 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.31 on 48 degrees of freedom
## Multiple R-squared:  0.2668,Adjusted R-squared:  0.2515
## F-statistic: 17.46 on 1 and 48 DF,  p-value: 0.0001232
```

```
hist(y - a_hat*x - b_hat, main = "Histograma de los residuos",
     freq = FALSE, xlab = "Error de modelo")
```



12.7 Análisis de varianza

El análisis de varianza (comunmente conocido como ANOVA por sus siglas en inglés), es un caso particular del modelo lineal que estudiamos antes, pero en el caso en que las X_i son 0 o 1. Veamos por medio de un ejemplo por qué merece ser estudiado en particular. Supongamos que se quiere comparar el rendimiento de un cultivo (medido en kilogramos por hectárea) en tres tipos de suelo (arenoso, arcilloso, limoso) teniendo 10 parcelas de cada uno. La variable rendimiento Y depende del tipo de suelo usado, y de la parcela en cuestión, por lo tanto tenemos $Y_{1,1}, \ldots, Y_{1,10}$ rendimientos para el tipo de suelo arenoso, para cada una de las 10 parcelas, análogamente tenemos $Y_{2,1}, \ldots, Y_{2,10}$ para el tipo de suelo arcilloso y finalmente $Y_{3,1}, \ldots, Y_{3,10}$ para el tipo de suelo limoso. Supondremos que podemos modelar las $Y_{i,1}, \ldots, Y_{i,10}$ como $Y_{i,j} = \mu + \alpha_i + e_{i,j}$ para $j = 1, \ldots, 10$, donde μ es una constante. Supondremos que los errores e_i son independientes, todos ellos con media 0 y varianza σ^2. La hipótesis que podríamos querer contrastar es $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ o no. Si *no* se rechaza H_0 tenemos indicios de que el rendimiento medio, en cada tipo de suelo, es el mismo, ya que estamos suponiendo $E(e_{i,j}) = 0$ y por lo tanto $E(Y_{i,j}) = \mu$. Es decir *no* hay diferencias en el rendimiento medio, según el tipo de suelo, si se rechaza hay evidencia estadística de que el suelo influye en el rendimiento medio. Otra aplicación similar e importante, es estudiar el efecto de ciertos medicamentos sobre alguna variable medible Y (presión arterial, nivel de colesterol, etc). En este caso rechazar H_0 estaría significando que alguno de los medicamentos influye sobre el valor promedio de la variable Y que se estudia. Es importante mencionar

que en lo que haremos la hipótesis de que *los errores tienen distribución normal y son independientes*, es crucial Aquí veremos unicamente el análisis de varianza de una vía o factor, por ejemplo en el caso del suelo, no consideramos otras características que puedan influir en el rendimiento, más que el tipo de suelo. Consideremos el modelo

$$Y_{i,j} = \mu + \alpha_i + e_{i,j} \quad con \quad i = 1, \ldots, k \quad y \quad j = 1, \ldots, n_i$$

Esto significa que del factor $i = 1, \ldots, k$ (tipo de suelo, tipo de medicamento, etc), tenemos n_i datos, por lo tanto en total tenemos $n = \sum_{i=1}^{k} n_i$ datos. La hipótesis que queremos contrastar es $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ contra la alternativa H_1 : para algún i, $\alpha_i \neq 0$. Para definir el estadístico que usaremos para realizar el test vamos a introducir algo de notación, denotamos:

$$\overline{Y_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j} \quad y \quad \overline{Y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{i,j}$$

Para el test se usa el estadístico

$$F_{k-1,n-k} = \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i (\overline{Y_i} - \overline{Y})^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{i,j} - \overline{Y_i})^2}. \tag{12.4}$$

Se puede demostrar que bajo H_0 y asumiendo que los errores son iid con distribucion $N(0, \sigma^2)$, este estadístico tiene una distribución conocida como "distribución F de Snedecor" o simplemente "F" la cual depende de n y k, observar que es mayor o igual que 0 ya que es cociente de cantidades que son positivas. Para calcular en R, $P(F_{k-1,n-k} < t)$ se escribe `pf(t,k-1,n-k)`, y si queremos el valor t_α de modo que $P(F_{k-1,n-k} > t_\alpha) = \alpha$ (es decir el cuantil $1 - \alpha$) escribimos `qf(1-α,k-1,n-k)`. Por lo tanto rechazamos H_0 a nivel α si el valor calculado $F_{k-1,n-k}$ en (12.4) supera el valor `qf(1-α,k-1,n-k)`. El siguiente ejemplo muestra que en realidad en R basta ingresar los datos y usar el comando `aov`.

12.8 Ejemplo en R

Consideremos el ejemplo descrito en la sección anterior, correspondientes al rendimiento de 3 tipos de suelos. Para eso tenemos la tabla:

arenoso	arcilloso	limoso
6	17	13
10	15	16
8	3	9
6	11	12
14	14	15
17	12	16
9	12	17
11	8	13
7	10	18
11	13	14

En R para crear dicha tabla se escribe

```
arenoso<-c(6,10,8,6,14,17,9,11,7,11)
arcilloso<-c(17,15,3,11,14,12,12,8,10,13)
limoso<-c(13,16,9,12,15,16,17,13,18,14)
datos<-data.frame(cbind(arenoso,arcilloso,limoso))
stackeddata<-stack(datos)
```

Este último comando transforma los datos en dos columnas, una con los valores (denominada values) de rendimiento dados por la tabla, y la otra con el tipo de suelo al que corresponde dicho valor, denominada ind. Ahora ejecutamos

```
summary(aov(values~ind,data=stackeddata))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## ind          2   99.2   49.60   4.245  0.025 *
## Residuals   27  315.5   11.69
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

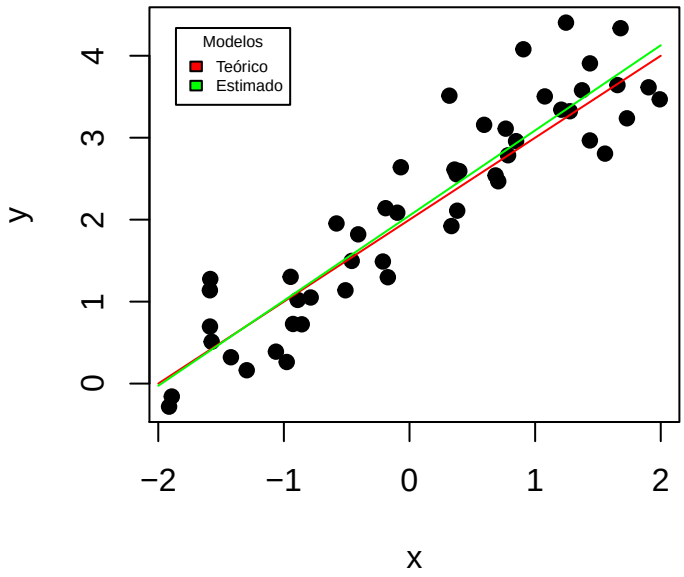
Obtuvimos una tabla con varios valores, entre ellos el valor del estadístico 12.4, que en este caso es 4.245, por lo tanto para realizar el test a nivel por ejemplo $\alpha = 0.05$, tenemos que comparar dicho valor con el que devuelve el comando qf(0.95,2,27):

```
qf(0.95,2,27)
```

```
## [1] 3.354131
```

Como $4.245 > 3.354$ se rechaza H_0 a nivel $\alpha = 0.05$. Esto significa que a nivel $\alpha = 0.05$ hay indicios de que el tipo de suelo influye sobre el rendimiento medio en los mismos.
El summary también devuelve el p-valor en la columna Pr(>F). En este caso 0.025, lo cual significa que $P_{H_0}(F_{2,27} > 4.245) = 0.025$. Por lo tanto como $0.025 < 0.05$ nuevamente rechazamos H_0.

```
n = 50
x = runif(n, -2, 2)
a = 1
b = 2
e = rnorm(n, 0, 1/2)
y = a*x + b + e
plot(x, y, pch = 19)
eje_x = seq(-2,2, length.out = 10000)
lines(eje_x, a*eje_x + b, col = "red")
regresion = lm(y ~ x)
a_hat = regresion$coefficients[2]
b_hat = regresion$coefficients[1]
lines(eje_x, a_hat*eje_x + b_hat, col = "green")
legend("topleft", inset=.05, title="Modelos", c("Teórico", "Estimado"),
       fill= c("red", "green"), cex = 0.5)
```



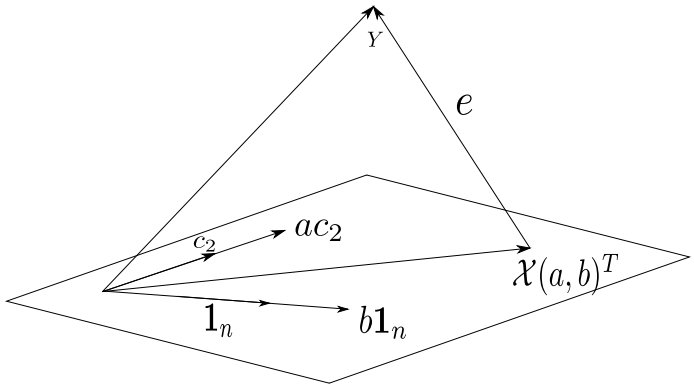


Figure 12.1: Observar que el vector *e* no es necesariamente ortogonal al plano generado por $\mathbf{1}_n$ y c_2

```
  n = 50
  x = runif(n, -2, 2)
  a = 1
  b = 2
  e = rnorm(n, 0, 10)
  y = a*x + b + e
  plot(x, y, pch = 19)
  eje_x = seq(-2,2, length.out = 10000)
  lines(eje_x, a*eje_x + b, col = "red")
  regresion = lm(y ~ x)
a_hat = regresion$coefficients[2]
b_hat = regresion$coefficients[1]
lines(eje_x, a_hat*eje_x + b_hat, col = "green")
legend("topleft", inset=.05, title="Modelos", c("Teórico", "Estimado"),
       fill= c("red", "green"), cex = 0.5)
```

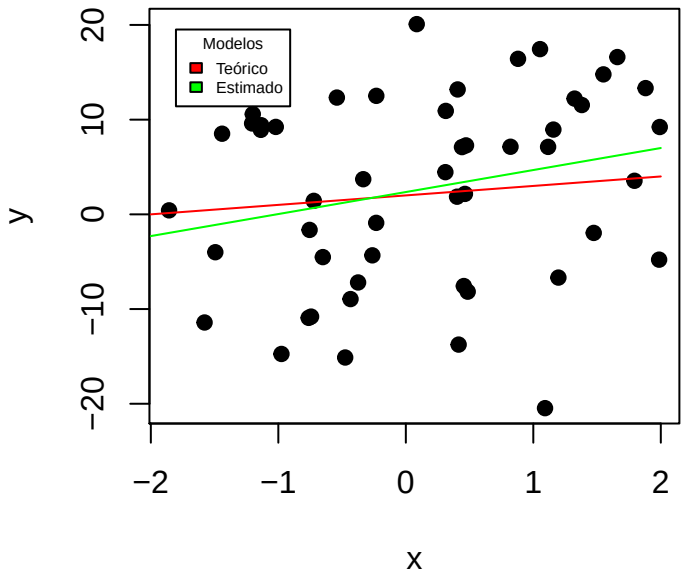


Figure 12.2: El valor σ del desvio de los errores es grande.

```
n = 50
x = runif(n, -2, 2)
a = 1
b = 2
e = rnorm(n, 0, 1)
y = a*x^2 + b + e
plot(x, y, pch = 19)
eje_x = seq(-2,2, length.out = 10000)
lines(eje_x, a*eje_x^2 + b, col = "red")
regnolin = lm(y ~ x)
a_hat = regnolin$coefficients[2]
b_hat = regnolin$coefficients[1]
lines(eje_x, a_hat*eje_x + b_hat, col = "green")
legend("topleft", inset=.05, title="Modelos", c("Teórico", "Estimado"),
       fill= c("red", "green"), cex = 0.5)
```

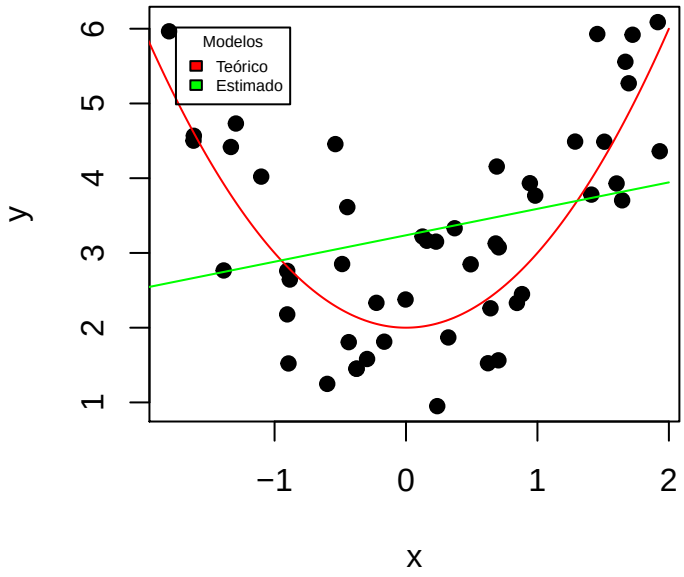


Figure 12.3: Hay una relación no lineal entre las variables

```r
n = 50
x = runif(n, -2, 2)
a = 1
b = 2
e = rnorm(n, 0, 1)
y = a*x + b + e
x[1] = 100
y[1] = -100 # El primer par (X_i, Y_i) es un outlier y no respeta el modelo lineal
plot(x, y, pch = 19)
eje_x = seq(-2,100, length.out = 10000)
lines(eje_x, a*eje_x + b, col = "red")
regconoutlier= lm(y ~ x)
a_hat = regconoutlier$coefficients[2]
b_hat = regconoutlier$coefficients[1]
lines(eje_x, a_hat*eje_x + b_hat, col = "green")
legend("topleft", inset=.05, title="Modelos", c("Teórico", "Estimado"),
       fill= c("red", "green"), cex = 0.5)
```

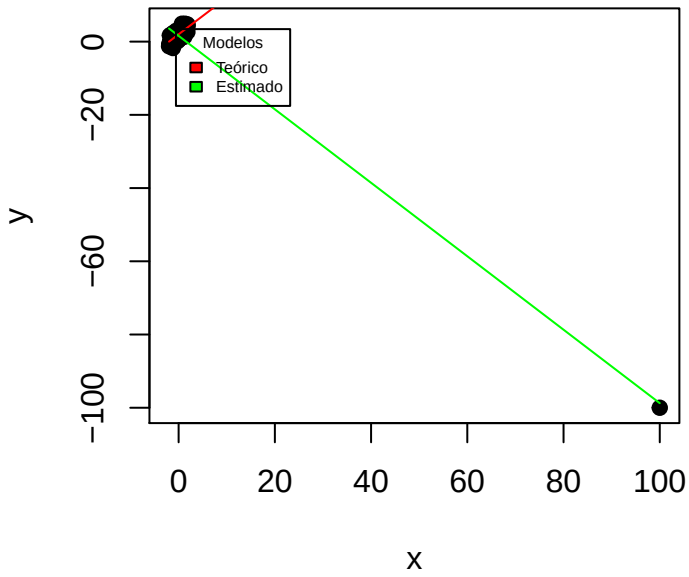


Figure 12.4: Modelo con 1 outlier

```
n = 50
x = runif(n, 0, 10)
a = 2
b = 0.1
y = rexp(n, (a*x + b)^(-1))
plot(x, y, pch = 19)
eje_x = seq(0,10, length.out = 1000)
lines(eje_x, a*eje_x + b, col = "red")
regresion = lm(y ~ x)
a_hat = regresion$coefficients[2]
b_hat = regresion$coefficients[1]
lines(eje_x, a_hat*eje_x + b_hat, col = "green")
legend("topleft", inset=.05, title="Modelos", c("Teórico", "Estimado"),
       fill= c("red", "green"), cex = 0.5)
```

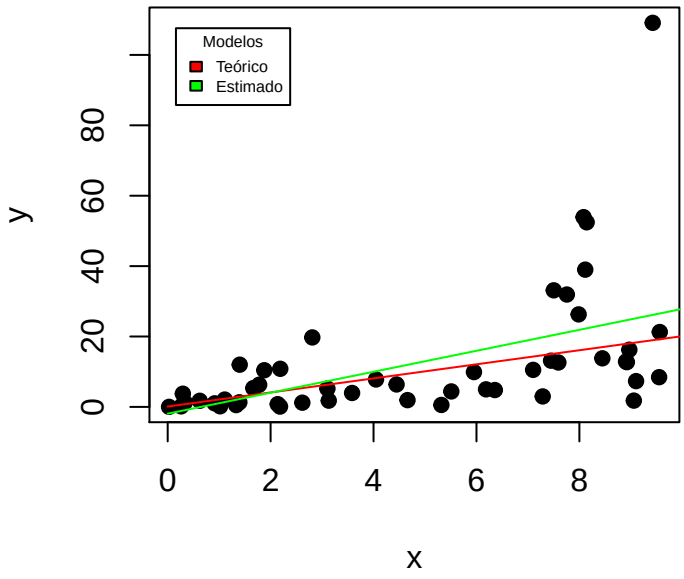


Figure 12.5: No se cumple la hipótesis de que los errores sean normales

Part III

Apéndice

Chapter 13

Apéndice

13.1 Intervalos de confianza

Veamos algunos casos particulares en los cuales se puede calcular un intervalo de confianza exacto, no aproximado.

13.1.1 Intervalos de confianza para datos normales, σ conocido.

Sea $X \backsim N(\mu, \sigma^2)$ con $0 < \sigma^2 < \infty$, conocido. Buscamos un intervalo I_n de la forma $\left[\overline{X_n} - k, \overline{X_n} + k\right]$. Debemos hallar k tal que $P(\mu \in I_n) = 1 - \alpha$, entonces

$$1 - \alpha \ = P\left(\overline{X_n} - k \leq \mu \leq \overline{X_n} + k\right) = P\left(\mu - k \leq \overline{X_n} \leq \mu + k\right).$$

Observemos que, por la parte 1) del Teorema (7.2) sabemos que $\overline{X_n} \backsim N\left(\mu, \frac{\sigma^2}{n}\right)$. Primero retaremos μ, y luego dividimos entre $\sigma/\sqrt{n}$, es decir

$$P\left(\mu - k \leq \overline{X_n} \leq \mu + k\right) = P\left(\mu - k - \mu \leq \overline{X_n} - \mu \leq \mu + k - \mu\right) = P\left(\frac{\mu - k - \mu}{\sigma/\sqrt{n}} \leq \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \leq \frac{\mu + k - \mu}{\sigma/\sqrt{n}}\right)$$

y de (4.8) tenemos que

$$\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

con lo cual, de (4.5)

$$P\left(\frac{-k}{\sigma/\sqrt{n}} \leq \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \leq \frac{k}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{k}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{-k}{\sigma/\sqrt{n}}\right). \tag{13.1}$$

Si usamos que $\Phi(-t) = 1 - \Phi(t)$ tenemos que 13.1 es

$$P\left(\frac{-k}{\sigma/\sqrt{n}} \leq \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \leq \frac{k}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{k}{\sigma/\sqrt{n}}\right) - \left[1 - \Phi\left(\frac{k}{\sigma/\sqrt{n}}\right)\right] = 2\Phi\left(\frac{\sqrt{n}k}{\sigma}\right) - 1 = 1 - \alpha.$$

Por lo tanto obtuvimos que

$$1 - \alpha/2 = \Phi\left(\frac{\sqrt{n}k}{\sigma}\right).$$

Si aplicamos Φ^{-1} a ambos lados de la desigualdad anterior obtenemos que $\frac{\sqrt{n}k}{\sigma} = \Phi^{-1}(1 - \alpha/2)$, es decir

$$k = \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1 - \alpha/2).$$

Notación: Denotaremos, para $\alpha \in (0,1)$, $z_\alpha = \Phi^{-1}(\alpha)$, este valor se calcula a partir de la tabla de Φ o con R, mediante el comando $\mathsf{qnorm}(\alpha)$, con esta notación el intervalo de confianza del ejemplo anterior es

$$\left[\overline{X_n} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \; ; \; \overline{X_n} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right]. \tag{13.2}$$

Es importante destacar que el intervalo (13.2) es exacto, no es para valores de n grandes. Por otro lado, observemos que la longitud del mismo es $2\frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$ que tiende a 0 cuando n tiende a infinito. Esto nos dice que a medida que tenemos más datos podemos determinar, con probabilidad $1 - \alpha$ con más exactitud (en un intervalo mas chico) donde estará μ.

Ejemplo 13.1. Veamos un ejemplo de como se calcula. La siguiente muestra corresponde a los niveles de sodio en sangre de 10 pacientes de una clínica (medidos en milimoles por litro).

139.13	136.67	140.25	140.57	137.71	142.38	142.56	139.92	140.65	140.35

Asumiendo que los datos provienen de una distribución normal $N(\mu, \sigma^2)$ con $\sigma = 2$, dar un intervalo de confianza 90% para la media μ.

Observemos primero que como estamos hablando de intervalo de confianza 90% entonces $\alpha = 0.1$ es decir $1 - \alpha/2 = 0.95$. Para calcular el intervalo definido en (13.2) tenemos que calcular $\overline{X_n}$ y $z_{0.95}$.

$$\overline{X_n} = \frac{1}{10}\left(139,13 + 136,67 + 140,25 + 140,57 + 137,71 + 142,38 + 142,56 + 139,92 + 140,65 + 140,35\right) \approx 140$$

$z_{0.95}$ es por definición $\Phi^{-1}(0.95)$, es decir es el valor que hace que el área hasta $z_{0.95}$ sea 0.95. Si vemos una tabla de la distribución normal, tenemos que $P\big(N(0,1) < 1.64\big) = 0.9494$ y $P\big(N(0,1) < 1.65\big) = 0.9505$ por lo tanto el $z_{0.95}$ que queremos está entre 1.64 y 1.65 (si hacemos $\mathsf{qnorm}(0.95)$ nos da 1.644854), tomemos

1.65, el intervalo es

$$\Big[140-\frac{2}{\sqrt{10}}1.65\ ;\ 140+\frac{2}{\sqrt{10}}1.65\Big]=\big[138.956\ ;141.043\big].$$

En R esto se escribe como (suponiendo que los datos están en un vector que llamamos x),

$$\Big[\mathsf{mean(x)}-\frac{2}{\sqrt{10}}\mathsf{qnorm(0.95)}\ ;\ \mathsf{mean(x)}+\frac{2}{\sqrt{10}}\mathsf{qnorm(0.95)}\Big]=\big[138.956\ ;\ 141.043\big].$$

13.1.2 Intervalos de confianza para datos normales, σ desconocido.

Veamos otro intervalo de confianza, para datos normales, pero el caso en que desconocemos σ. Para eso sea $X \backsim N(\mu,\sigma^2)$ con $0<\sigma^2<\infty$, desconocido. Dado que desconocemos σ pero sabemos por lo dicho en la sección anterior que S_n es un estimador del mismo, lo que se hace es sustituir S_n por σ. Buscamos un intervalo I_n de la forma

$$\big[\overline{X_n}-kS_n\ ;\ \overline{X_n}+kS_n\big].$$

Tenemos que hallar k tal que $P(\mu\in I_n)=1-\alpha$, por lo tanto

$$P(\mu\in I_n)=P\Big(\mu-kS_n\le\overline{X_n}\le\mu+kS_n\Big)=P\Big(-k\le\frac{\overline{X_n}-\mu}{S_n}\le k\Big).$$

Si multiplicamos todo por $\sqrt{n}$ tenemos que

$$P(\mu\in I_n)=P\Big(-k\sqrt{n}\le\sqrt{n}\frac{\overline{X_n}-\mu}{S_n}\le k\sqrt{n}\Big).$$

Por el Teorema (7.2) 4) sabemos que $\sqrt{n}\frac{\overline{X_n}-\mu}{S_n}$ tiene distribución T_{n-1} de Student con $n-1$ grados de libertad. Denotemos $F_{T_{n-1}}(x)$ la función de distribución de T_{n-1} evaluada en x (es la función que juega el papel de Φ) es decir

$$P(T_{n-1}\le x)=F_{T_{n-1}}(x).$$

Observemos que

$$P(\mu\in I_n)=P\Big(-k\sqrt{n}\le\sqrt{n}\frac{\overline{X_n}-\mu}{S_n}\le k\sqrt{n}\Big)=F_{T_{n-1}}\Big(k\sqrt{n}\Big)-F_{T_{n-1}}\Big(-k\sqrt{n}\Big).$$

Si usamos la propiedad (4.10) de las variables con distribución T de Student tenemos que

$$\begin{aligned}P(\mu\in I_n)=&F_{T_{n-1}}(\sqrt{n}k)-F_{T_{n-1}}(-\sqrt{n}k)\\=&2F_{T_{n-1}}(\sqrt{n}k)-1=1-\alpha,\end{aligned}$$

de donde

$$k = \frac{F^{-1}_{T_{n-1}}(1-\alpha/2)}{\sqrt{n}} = \frac{t_{1-\alpha/2}(n-1)}{\sqrt{n}},$$

donde usamos la notación $F^{-1}_{T_{n-1}}(p) = t_p(n-1)$ siendo $n-1$ los grados de libertad. Por lo tanto el intervalo de confianza para μ al nivel $1-\alpha$ es

$$I_n = \left[\overline{X_n} - \frac{S_n}{\sqrt{n}} t_{1-\alpha/2}(n-1); \overline{X_n} + \frac{S_n}{\sqrt{n}} t_{1-\alpha/2}(n-1)\right].$$

La longitud de dicho intervalo es

$$2\frac{S_n}{\sqrt{n}} t_{1-\alpha/2}(n-1)$$

que nuevamente tiende a 0 cuando n tiende a infinito (ya que $t_{1-\alpha/2}(n-1) \to z_{1-\alpha/2}$ y $S_n \to \sigma$).

Ejemplo 13.2. Veamos, para los datos del ejemplo anterior, y $\alpha = 0.1$ como queda el intervalo de confianza. Si usamos la fórmula para S_n nos da que $S_n^2 = 3.36074$ y por lo tanto $S_n = 1.8332$. Nos resta calcular $t_{0.95}(9)$, esto se puede hacer de dos maneras: con una tabla análoga a la que usabamos para Φ. En la tabla de la distribución t de Student tenemos en general, en la primer columna los posibles grados de libertad, por lo tanto nos fijamos en la fila 9. Las diferentes columnas corresponden a los diferentes α. Observemos que si bien trabajamos con $\alpha = 0.1$, lo que queremos es el valor $t_{0.95}(9)$ es decir

$$P(T_9 < t_{0.95}(9)) = 0.95,$$

que, uso de tabla mediante, nos da $t_{0.95}(9) = 1.83$. En R basta usar el comando `qt(0.95,9)` y nos da 1.833113. El intervalo nos queda entonces

$$\left[140 - \frac{1.8332}{\sqrt{10}} 1.83 \; ; \; 140 + \frac{1.8332}{\sqrt{10}} 1.83\right] = [138.94 \; ; 141.06]$$

Observemos que este intervalo es levemente más largo que el que obtuvimos antes. Lo cual es muy razonable si pensamos que en este caso teníamos menos información (desconociamos σ).

13.2 Pruebas de Hipótesis para datos normales

13.2.1 Pruebas de hipótesis unilaterales, datos normales, σ conocido

En el caso de que $X_1, \ldots, X_n$ tengan distribución $N(\mu, \sigma^2)$ con σ conocido, las pruebas de hipótesis

$$\begin{cases} H_0: & \mu & \leq \mu_0 \\ H_1: & \mu & > \mu_0 \end{cases} \qquad \begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & > \mu_0 \end{cases} \qquad \begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & = \mu_1 \end{cases}$$

con $\mu_1 > \mu_0$ tienen región crítica

$$RC = \left\{ \overline{X_n} > \mu_0 + \frac{z_{1-\alpha}\sigma}{\sqrt{n}} \right\}.$$

Para demostrar esto tenemos que hallar $\delta > 0$ tal que $P_{H_0}(\overline{X}_n > \mu_0 + \delta) = \alpha$. Usaremos que, por la parte 1) del Teorema (7.2) bajo H_0 (en este caso la esperanza de las X_i es $\mu_0 = \mu$), $\overline{X_n} \backsim N\left(\mu, \frac{\sigma^2}{n}\right)$. Por lo tanto

$$\alpha = P_{H_0}\Big(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} > \frac{\delta}{\sigma/\sqrt{n}}\Big) = 1 - \Phi\Big(\frac{\delta}{\sigma/\sqrt{n}}\Big)$$

o lo que es lo mismo

$$1 - \alpha = \Phi\Big(\frac{\delta}{\sigma/\sqrt{n}}\Big).$$

Si aplicamos Φ^{-1} a ambos lados de la igualdad anterior obtenemos

$$\frac{\delta}{\sigma/\sqrt{n}} = \Phi^{-1}(1-\alpha) = z_{1-\alpha},$$

de donde se sigue que

$$\delta = z_{1-\alpha}\frac{\sigma}{\sqrt{n}}.$$

Análogamente, las pruebas de hipótesis

$$\begin{cases} H_0: & \mu & \geq \mu_0 \\ H_1: & \mu & < \mu_0 \end{cases} \qquad \begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & < \mu_0 \end{cases} \qquad \begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & = \mu_1 \end{cases}$$

donde $\mu_1 < \mu_0$, tienen región crítica

$$RC = \left\{ \overline{X_n} < \mu_0 - \frac{z_{1-\alpha}\sigma}{\sqrt{n}} \right\}.$$

Ejemplo 13.3. Consideremos los datos del ejemplo (13.1) correspondientes al nivel de sodio en sangre de 10 personas (observemos que aquí estamos asumiendo que los datos tienen distribución normal con σ conocido). Vamos a considerar que el nivel es crítico si es mayor o igual que $\mu_0 = 143$. Trabajaremos con $\alpha = 0.05$ Plantemos la prueba de hipótesis

$$\begin{cases} H_0: & \mu \geq 143 \\ H_1: & \mu < 143 \end{cases}$$

Planteamos la región crítica

$$RC = \left\{ \overline{X_n} < 143 - 2\frac{z_{0.95}}{\sqrt{10}} \right\}.$$

Donde, como en intervalos de confianza, $z_{0.95}$ denota el valor de t (que se obtiene mediante la tabla de la

distribución normal o con el comando qnorm) tal que $\Phi(t) = 0.95$. Verificar que $z_{0.95} = 1.65$, entonces

$$RC = \left\{\overline{X_n} < 141.93\right\}.$$

Como $\overline{X_n} = 140 < 141.93$ estamos en la región crítica, por lo tanto rechazamos H_0.

13.2.2 Pruebas de hipótesis unilaterales, datos normales, σ desconocido

En el caso de que $X_1, \ldots, X_n$ tengan distribución $N(\mu, \sigma^2)$ con σ desconocido, las pruebas de hipótesis

$$\begin{cases} H_0: & \mu & \leq \mu_0 \\ H_1: & \mu & > \mu_0 \end{cases} \qquad \begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & > \mu_0 \end{cases} \qquad \begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & = \mu_1 \end{cases}$$

con $\mu_1 > \mu_0$ tienen región crítica

$$RC = \left\{\overline{X_n} > \mu_0 + \frac{t_{1-\alpha}(n-1)S_n}{\sqrt{n}}\right\}.$$

Por otro lado, las pruebas de hipótesis

$$\begin{cases} H_0: & \mu & \geq \mu_0 \\ H_1: & \mu & < \mu_0 \end{cases} \qquad \begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & < \mu_0 \end{cases} \qquad \begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & = \mu_1 \end{cases}$$

Con $\mu_1 < \mu_0$. Tienen región crítica

$$RC = \left\{\overline{X_n} < \mu_0 - \frac{t_{1-\alpha}(n-1)}{\sqrt{n}}S_n\right\}.$$

Ejemplo 13.4. La siguiente muestra registra los niveles en sangre de colesterol *LDL*, medidos en mg/dl, correspondientes a diez pacientes de una clínica.

134.3	133.7	129.5	131.2	128.4	125.4	135.0	130.4	133.0	132.7

Clínicamente, se consideran altos los niveles de colesterol que están por encima de 130 mg/dl. Asumiendo que los datos tienen distribución normal con media μ y desviación estándar σ (que por el momento suponemos desconocida), implemente (con nivel $\alpha = 0.05$) la siguiente prueba de hipótesis para decidir si los niveles de colesterol son razonables:

$$\begin{cases} H_0: & \mu & \geq 130 \\ H_1: & \mu & < 130 \end{cases}$$

Observemos que como desconocemos σ tenemos que estimarlo mediante S_n, haciendo cuentas $\overline{X_n} = 131.36$,

$S_n^2 = 8.9493$. Por otro lado, recordemos que $t_{0.95}(9) \approx 1.83$, entonces

$$RC = \left\{ \overline{X_n} < 130 - 1.83\frac{2.991}{\sqrt{10}} \right\} = \left\{ \overline{X_n} < 128.27 \right\}$$

Como $131.36 > 128.27$ no estamos en la región crítica y por lo tanto no se rechaza H_0.

13.2.3 Pruebas de hipótesis bilaterales, datos normales, σ conocido

En el caso de que $X_1, \ldots, X_n$ tengan distribución $N(\mu, \sigma^2)$ con σ conocido, la prueba de hipótesis:

$$\begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & \neq \mu_0 \end{cases}$$

tiene región crítica

$$RC = \left\{ \left| \frac{\sqrt{n}(\overline{X_n} - \mu_0)}{\sigma} \right| \geq z_{1-\alpha/2} \right\}$$

Ejemplo 13.5. Cada luciérnaga tiene un modo peculiar de centellear; un destello corto de luz es seguido por un período de reposo. Los siguientes datos corresponden a los períodos de reposo entre centelleos (medidos en segundos) para una muestra de 11 luciérnagas:

4.05	3.95	3.74	3.33	3.94	4.04	3.73	3.75	3.88	3.50	3.59

Se asume que los datos corresponden a una distribución normal de valor esperado μ y desvio $\sigma = 0.06$. Realice una prueba de hipótesis al nivel 0.1 para decidir entre las siguientes hipótesis

$$\begin{cases} H_0: & \mu & = 4 \\ H_1: & \mu & \neq 4 \end{cases}$$

$\overline{X_n} = 3.773$, $\sigma = 0.06$. Haciendo cuentas tenemos que

$$\frac{\sqrt{n}(\overline{X_n} - \mu_0)}{\sigma} = \frac{\sqrt{11}(3.773 - 4)}{0.06} = -12.714$$

y al tomar valor absoluto tenemos que $|-12.714| = 12.714$ que es mayor que $z_{0.95} = 1.65$ y por lo tanto estamos en la región crítica y se rechaza H_0.

13.2.4 Pruebas de hipótesis bilaterales, datos normales, σ desconocido

En el caso de que $X_1, \ldots, X_n$ tengan distribución $N(\mu, \sigma^2)$ con σ desconocido, la pruebas de hipótesis

$$\begin{cases} H_0: & \mu & = \mu_0 \\ H_1: & \mu & \neq \mu_0 \end{cases}$$

tiene región crítica

$$RC = \left\{ \left| \frac{\sqrt{n}(\overline{X_n} - \mu_0)}{S_n} \right| \geq t_{1-\alpha/2}(n-1) \right\}$$

13.2.5 Pruebas de hipótesis para proporciones

Supongamos que queremos saber si una moneda está balanceada o no. Se tira 100 veces y obtenemos 54 caras, debemos tomar una decisión entre

$$\begin{cases} H_0: & p = 1/2 \quad \text{donde } p = P(\text{cara}) \\ H_1: & p \neq 1/2. \end{cases}$$

Es razonable pensar que rechazaremos H_0 si el promedio de veces que salió cara es mucho menor que $1/2$ o mucho mayor. Es decir planteamos una región crítica de nivel α de la forma

$$RC = \left\{ \left| \frac{\sqrt{n}(\overline{X_n} - \mu_0)}{\sigma} \right| \geq z_{1-\alpha/2} \right\},$$

Observemos que, bajo H_0 estamos en el caso en que $\mu = p = 1/2$ por lo tanto $\sigma^2 = p(1-p)$. En el caso de la moneda, supongamos que trabajamos con $\alpha = 0.05$, $\overline{X_n} = 54/100$, $z_{1-\alpha/2} = z_{0.975} = 1.96$. Haciendo cuentas

$$\frac{\sqrt{n}(\overline{X_n} - \mu_0)}{\sigma} = \frac{10(54/100 - 1/2)}{1/2} = 0.8.$$

Como $0.8 < 1.96$ no estamos en la región crítica y por lo tanto no se rechaza H_0.

13.3 Más test de Aleatoriedad

13.3.1 Test de Pearson

El test de correlación de Pearson que usa como estadístico una función de (7.4). Si bien lo que mide es si existe o no una relación lineal entre las variables (ya que en el caso de máxima correlación una variable es una combinación lineal de la otra), bajo la hipótesis de independencia dicha relación no existe. Por lo tanto

es una forma indirecta de medir dependencia.

Como dijimos antes, el coeficiente de correlación de Pearson (ver 6.3) es un número entre -1 y 1 que mide si existe o no una relación lineal entre las variables X e Y. Bajo la hipótesis de que son independientes, dicha correlación debería dar 0. Por lo tanto testar correlación (dependencia lineal) es una forma indirecta de medir grado de dependencia, pero hay que tener presente que las variables podrían ser dependientes con una dependencia que no es lineal. En esos casos el test *no* rechaza. Esto implica que la potencia de la prueba en muchos casos puede no ser alta, (y por lo tanto el error de tipo II, β, no ser bajo).
Vamos a explicar como es el estadístico, cual es la región crítica y como se calcula el p-valor. Para eso supongamos que tenemos dos muestras $X_1,\ldots,X_n$ con la misma distribución que una variable X e $Y_1,\ldots,Y_n$ de Y (por ahora no supondremos que las muestras son independientes). Primero calculamos el coeficiente de correlación estimado $\hat{\rho}$, según la formula 7.4. Luego el estadístico de prueba es

$$T = \hat{\rho}\sqrt{\frac{n-2}{1-\hat{\rho}^2}} \tag{13.3}$$

Se puede demostrar que, bajo la hipótesis nula de que $\rho(X,Y) = 0$, para valores de n *grandes*

$$P_{H_0}(T > t) \approx P(T_{n-2} > t), \tag{13.4}$$

donde T_{n-2} es una variable aleatoria con distribución T de Student con $n-2$ grados de libertad (ver 4.14). La región crítica es de la forma $RC = \{|T| > t_n\}$ y t_n se calcula usando (13.4), es decir fijado $\alpha \in (0,1)$, $P_{H_0}(|T| > t) \approx P(|T_{n-2}| > t) = \alpha$. Recordemos que la distribución T_{n-2} es simétrica respecto del origen y hay que tomar $t = t_{1-\alpha/2}$, el cuantil $1-\alpha/2$ de la distribución T_{n-2}. Por lo tanto rechazamos H_0 si $|T| > t_{1-\alpha/2}$.

En R

Para realizar el test de correlación de Pearson creamos dos vectores, u y t con los datos, luego ejecutamos el comando `cor.test(t,u,method="pearson")`, esto nos da el valor del estadístico, el número df de grados de libertad $(n-2)$ y el p-valor para el test. Como siempre, rechazamos a nivel α si el p-valor es menor que α.

13.3.2 Test de Rachas

El test de rachas se basa en estudiar la forma en que la muestra está ordenada, es decir, las variables crecen o decrecen.

Supongamos que tenemos 10 mediciones de temperatura hechas a la misma hora, durante 20 días, es decir tenemos $X_1,\ldots,X_{10}$ es claro que si los datos son crecientes, no van a ser independientes, ya que saber el valor de la temperatura en el día i-ésimo nos aporta información respecto de lo que pasará con la temperatura el día siguiente, nos dice que será mayor. Lo mismo diríamos si los valores son decrecientes o si alternan (crece:$X_1 < X_2$, decrece: $X_3 < X_2$, crece: $X_3 < X_4$, etc). Esta idea intuitiva que relaciona el crecimiento-decrecimiento de los datos es la que usaremos para testear la alteatoriedad de la muestra. Lo que hacemos es, dados los datos $X_1,\ldots,X_n$ contruir $n-1$ variables $Y_1,\ldots,Y_{n-1}$ donde la variable Y_i es 1 si

$X_i < X_{i+1}$ y 0 en otro caso. Por ejemplo, si las X_i son mediciones (en grados celcius) de temperatura dadas por la siguiente tabla:

1	22,67	6	17,88
2	21,66	7	23,17
3	16,31	8	24,85
4	15,95	9	15,17
5	15,15	10	23,19

Las variables $Y_1, \ldots, Y_9$ son

1	0	6	1
2	0	7	1
3	0	8	0
4	0	9	1
5	1		

En general, supongamos que tenemos variables aleatorias $X_1, X_2, \ldots, X_n$ y queremos testear si son i.i.d, para eso definimos las variables $Y_1, \ldots, Y_{n-1}$ de la siguiente forma

$$Y_i = \begin{cases} 1 \text{ si } X_i < X_{i+1} \\ 0 \text{ en otro caso} \end{cases}$$

Lo que haremos es estudiar el número de rachas totales de $Y_1, \ldots, Y_{n-1}$ o lo que es lo mismo

$$\hat{R}_n = 1 + \sum_{i=1}^{n-2} \mathbb{I}_{\{Y_i \neq Y_{i+1}\}}.$$

Observemos que en general $\hat{R}_n \leq n - 1$.
En el caso del ejemplo anterior, verificar que $n = 10$ y $\hat{R}_n = 4$. Si n es chico ($n \leq 25$) la distribución de $\hat{R}_n$ esta tabulada, y rechazamos la hipótesis de ser i.i.d., a nivel α, si el valor observado $\hat{R}_n$ cumple que $|\hat{R}_n| > R_{1-\alpha/2}$ donde $R_{1-\alpha/2}$ se obtiene de la tabla. La tabla nos da, para diferentes valores de n la probabilidad de que se obtengan k rachas, con k de 0 a $n-1$ veamos en el ejemplo. En nuestro caso vamos en la tabla a $n = 10$ y obtenemos

n	$\hat{R}_n$	Left tail P	$\hat{R}_n$	Right tail P
10	1	0,0000		
	2	0,0003	9	0,0278
	3	0,0079	8	0,1671
	4	0,0633	7	0,4524
	5	0,2427	6	0,7573

Esta tabla hay que interpretarla de la siguiente manera, si $\hat{R}_{10}$ denota la variable aleatoria que indica el número de rachas posibles cuando tenemos $n = 10$ datos (que es un valor como dijimos, que puede ser $1, 2, 3, \ldots, 9$),

tenemos que la columna *Left tail* indica las probabilidades hasta ese valor de $\hat{R}_{10}$, y por lo tanto va creciendo a medida que $\hat{R}_{10}$ decrece, es decir

$$P(\hat{R}_{10} \leq 1) = 0.0000,\ P(\hat{R}_{10} \leq 2) = 0.0003,\ P(\hat{R}_{10} \leq 3) = 0.0079,\ P(\hat{R}_{10} \leq 4) = 0.0633,\ P(\hat{R}_{10} \leq 5) = 0.2427.$$

La columna *Right tail* nos da la probabilidad de obtener valores mayores o iguales que el valor, por lo tanto va creciendo a medida que $\hat{R}_{10}$ decrece, es decir

$$P(\hat{R}_{10} \geq 6) = 0.7573,\ P(\hat{R}_{10} \geq 7) = 0.4524,\ P(\hat{R}_{10} \geq 8) = 0.1671 \text{ y } P(\hat{R}_{10} \geq 9) = 0.0278.$$

Supongamos que nuestro nivel de confianza es $\alpha = 0.1$. Vamos a rechazar H_0 (que los datos son *i.i.d*) si obtenemos valores de $\hat{R}_{10}$ muy grandes, o muy chicos. Dado que $P(\hat{R}_{10} \leq 3) = 0.079 < \alpha/2$ y $P(\hat{R}_{10} \leq 4) = 0.0633 > \alpha/2$, la región crítica incluye los valores de $\hat{R}_n = 1, 2, 3$. Por otro lado $P(\hat{R}_{10} \geq 9) = 0.0278 < \alpha/2$ mientras que $P(\hat{R}_{10} \geq 8) = 0.1671 > \alpha/2$. Finalmente, en virtud de esto, la región crítica es

$$RC = [1, 3] \cup [9]$$

como obtuvimos $\hat{R}_{10} = 4$ no estamos en la región crítica y por lo tanto no se rechaza la hipótesis (nula) de que los datos son *iid*.

Ejercicio 13.6.

- Usando la tabla de $\hat{R}_n$ demostrar que la región crítica para $\alpha = 0.10$ y $n = 15$ datos es

$$RC = [1.5] \cup [14.15]$$

- Usando la tabla de $\hat{R}_n$ demostrar que la región crítica para $\alpha = 0.10$ y $n = 23$ datos es

$$RC = [1.11] \cup [19.22]$$

  mientras que si tomamos $\alpha = 0.05$ la región crítica (para $n = 23$) es

$$RC = [1.10] \cup [20.22]$$

Para valores grandes de n ($n > 25$) usamos que

$$P\left(\frac{\hat{R}_n - (2n-1)/3}{\sqrt{\frac{16n-29}{90}}} \leq t \right) \to \Phi(t)$$

por lo tanto vamos a rechazar H_0 si

$$\left| \frac{\hat{R}_n - (2n-1)/3}{\sqrt{\frac{16n-29}{90}}} \right| > z_{1-\alpha/2}$$

13.4 El modelo lineal, el caso general

Consideremos primero el modelo lineal de efectos fijos $Y_i = aX_i + b + e_i$ con $\text{var}(e_i) = \sigma^2$ y e_i iid con distribución normal con media 0. Se deja como ejercicio verificar que los estimadores $\hat{a}_n$ y $\hat{b}_n$ verifican $E(\hat{a}_n) = a$, $E(\hat{b}_n) = b$,

$$\text{Var}(\hat{a}_n) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}$$

$$\text{Var}(\hat{b_n}) = \sigma^2 \Big[\frac{1}{n} + \frac{(\overline{X}_n)^2}{\sum_{i=1}^n (X_i - \overline{X}_n)^2} \Big]$$

De donde se deduce que $Y_i \sim N(a + bX_i, \sigma^2)$ y

$$\hat{a}_n \sim N\Big(a, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}\Big)$$

Observar que $Y_i = \hat{a}_n X_i + \hat{b}_n + \hat{e}_i$ donde $\hat{e}_i = Y_i - \hat{Y}_i$, como $E(\hat{a}_n) = a$ y $E(\hat{b}_n) = b$ se sigue que $E(\hat{e}_i) = 0$. Un estimador de la varianza de $\hat{e}_i$ es entonces

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}.$$

13.4.1 El modelo lineal general, efectos fijos

En el modelo lineal general la variable Y depende no solamente de una variable X sino de un conjunto k de variables $X_1, \ldots, X_k$ (las cuales se asumen independientes de los errores) pero la dependencia es lineal. Esto es, existen constantes $\theta_1, \ldots, \theta_p$ tal que $Y = \theta_1 X_1 + \theta_p X_p + e$ donde e es una variable aleatoria independiente de X que en general se asume con media 0 y varianza σ^2. El objetivo es entonces estimar $\theta_1, \ldots, \theta_p$ a partir de una muestra de $Y_1, \ldots, Y_n$ de variables independientes, con la misma distribución que Y, y n muestras independientes $X_{11}, \ldots, X_{n1}$ de X_1, n muestras independientes $X_{21}, \ldots, X_{2,n}$ de X_2, hasta $X_{1k}, \ldots, X_{nk}$ n muestras independientes de X_k. Al igual que antes los errores $e_1, \ldots, e_k$ no son conocidos. Por lo tanto podemos plantear de forma matricial:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1k} \\ \vdots & \vdots & \ddots & \\ X_{n1} & X_{n2} & \ldots & X_{nk} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_k \end{pmatrix}$$

La independencia es por filas, pero X_{11} no necesariamente es independientes de X_{12} por ejemplo. Para simplificar el análisis supondremos que

$$\mathcal{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ \vdots & \vdots & \ddots & \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix},$$

que se llama matriz de diseño, constante y conocida.

Ejemplo 13.7. Modelo lineal simple: $Y = \alpha + \beta\mathcal{X} + e$, tomamos $(Y_1, X_1), \dots, (Y_n, X_n)$ y $\theta = (\alpha, \beta)$, y como matriz de diseño la matriz

$$\mathcal{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_n \end{pmatrix},$$

lo que se busca es entonces ajustar una recta a los datos.

Ejemplo 13.8. Ajuste de un polinomio de grado k: De forma análoga al ejemplo anterior, si $Y = \alpha + \beta_1 X_1 + \beta_2 X_2^2 + \dots + \beta_k X_k^k + e$, planteamos la matriz de diseño

$$\mathcal{X} = \begin{pmatrix} 1 & X_1 & X_1^2 & \dots & X_1^k \\ \vdots & \vdots & & & \\ 1 & X_n & X_n^2 & \dots & X_n^k \end{pmatrix}.$$

13.5 Hipótesis del modelo

Algunas hipótesis que usaremos son las siguientes:

1) $Rango(\mathcal{X}) = k$.

2) Los errores tienen media 0, $E(e_i) = 0$ para todo i.

3) Homocedasticidad: $Var(e_i) = \sigma^2$ para todo i.

3') $cov(e_i, e_j) = 0$ para todo $i \neq j$.

4) el vector e de errores tiene distribución $N(0, \sigma^2 I)$ en este caso se cumplen $2), 3)$ y $3')$

Para estimar $\theta \in \mathbb{R}^k$ buscamos hallar $\hat{\theta} \in \mathbb{R}^k$ donde se realiza

$$\min_{\theta \in \mathbb{R}^k} \|Y - \mathcal{X}\theta\|.$$

Observación 13.9. *Se deja como ejerccio verificar que bajo la hipótesis 1 se cumple que* $\hat{\theta} = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'Y$ *es solución del problema de minimización anterior. Observar que la matriz* $(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'$ *es la proyección*

ortogonal del vector Y sobre el espacio vectorial (de dimensión k) general por las columnas de $\mathcal{X}$, ver figura 13.1. La norma del vector u es la suma de cuadrados de la regresión mientras que $\cos(\alpha)$ *es es el coeficiente de determinación* R^2.

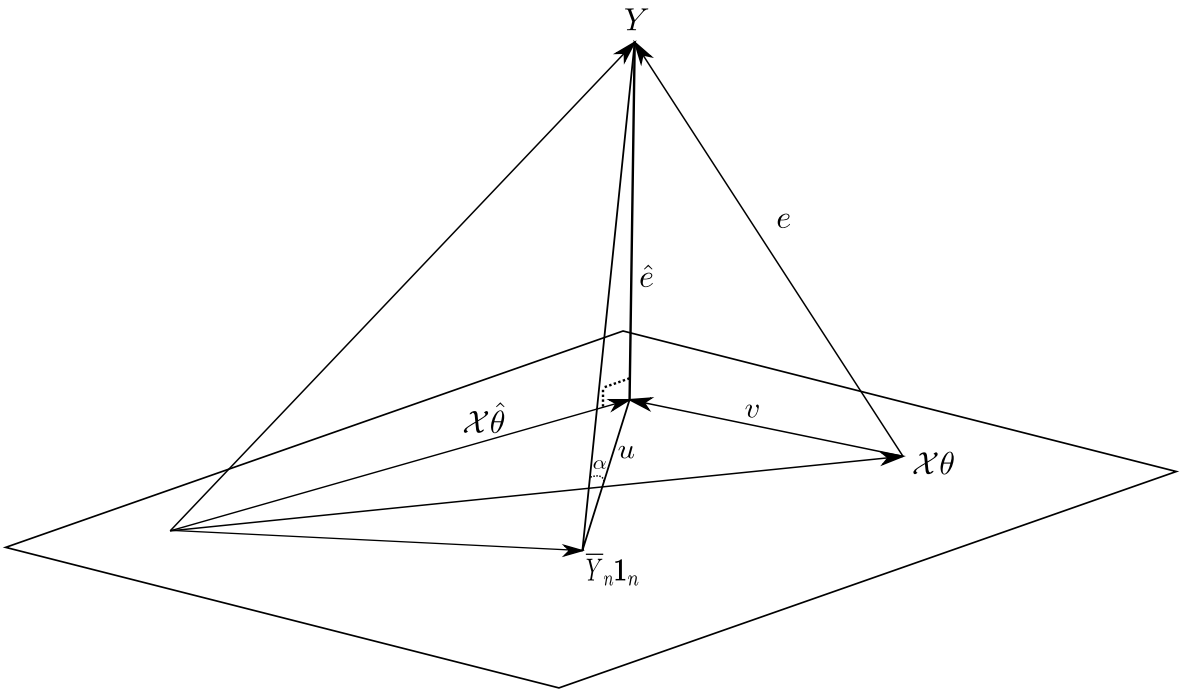


Figure 13.1: $\mathcal{X}\hat{\theta}$ es la proyección ortogonal de Y sobre el espacio generado por las columnas de $\mathcal{X}$. $\mathbf{1}_n = (1,1,\ldots,1) \in \mathbb{R}^n$. $R^2 = \cos(\alpha)$. Observar que $\|u\|^2 = SC_{Reg}$

Teorema 13.10. *a) Bajo las hipótesis* 1) *y* 2), $\hat{\theta}$ *es insesgado, es decir* $E(\hat{\theta}) = \theta$.

b) Bajo las hipótesis 1),2) *y* 3), $\Sigma_{\hat{\theta}} = \sigma^2(\mathcal{X}'\mathcal{X})^{-1}$.

Teorema 13.11. *Bajo los supuestos 1) a 4)*

a) $\dfrac{n\hat{\sigma}^2}{\sigma^2} = \dfrac{\|Y - \mathcal{X}\hat{\theta}\|^2}{\sigma^2} \backsim \chi^2_{(n-k)}$

b) $s^2 = \dfrac{n\hat{\sigma}^2}{n-k} = \dfrac{\|Y - \mathcal{X}\hat{\theta}\|}{n-k}$ *es insesgado (de donde* $\hat{\sigma}^2$ *es asintóticamente insesgado).*

c) $\dfrac{\|\mathcal{X}(\hat{\theta} - \theta)\|^2}{ks^2} \backsim F(k, n-k)$

d) $\dfrac{\lambda_1(\hat{\theta}_1 - \theta_1) + \lambda(\hat{\theta}_2 - \theta_2) + \cdots + \lambda_n(\hat{\theta}_n - \theta_n)}{s\sqrt{\lambda'(\mathcal{X}'\mathcal{X})^{-1}\lambda}} \backsim t_{n-k} \quad \forall \lambda \in \mathbb{R}^n$

Index

Bibliography

[FPP] D. Freedmand, R. Pisani, and R. Purves, *Estadística. 2ed*, Antoni Bosch, Editor., 1993.

[G] R. Grimaldi, *Matemáticas discreta y combinatoria. 3ed.*, Addison Wesley, 1997.

[S] B. Shahbaba, *Bioestatistics with r*, Springer, 2012.