UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

INFERENCIA ESTADÍSTICA APLICADA EN LA GENERACIÓN DE UNA PROPUESTA DE HORARIOS PARA LAS CARRERAS DEL DEPARTAMENTO DE MATEMÁTICAS

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

PRESENTA:
MIRIAM GABRIELA COLÍN NÚÑEZ

TUTOR
DR. ARRIGO COEN CORIA

CIUDAD UNIVERSITARIA, CD. MX., 2020

Datos de la Alumna:

Colín

Núñez

Miriam Gabriela

(Teléfono)

Universidad Nacional Autónoma de México

Facultad de Ciencias

Actuaría

(número de cuenta)

Datos del tutor:

Dr.

Arrigo

Coen

Coria

Datos del sinodal 1:

Datos del sinodal 2:

Datos del sinodal 3:

Datos del sinodal 4:

Datos del sinodal 5:

Datos del trabajo escrito:

Inferencia estadística aplicada en la generación de una propuesta de horarios para las carreras del departamento de matemáticas

(Número de Páginas)

2020

Dedicado a

Agradecimientos

¡Muchas gracias a todos!

Índice general

Índice de figuras

Índice de tablas

Códigos

Capítulo 1

Introducción

En este trabajo se hará un análisis estadístico de los datos recabados de las páginas de horarios de la Facultad de Ciencias de la UNAM. Se obtendrá un número estimado de alumnos, para cada materia y por cada hora, de las carreras del Departamento de Matemáticas. Se harán aleatoriamente esqueletos de horarios que se calificarán de acuerdo a ciertos criterios. Se resolverá el problema de asignación de horarios por medio del algoritmo genético. Con esto se desea disminuir el tiempo que se toma actualmente el hacer tanto los esqueletos de horarios como las asignaciones de grupos en la FC.

1.1. Motivación

Lo que motivó la realización de este trabajo es la aportación que se puede hacer a la FC, la cual nos parece de gran utilidad y para el beneficio de los futuros alumnos. Podremos tener una (posible) disminución del tiempo que toma realizar los esqueletos y la asignación de profesores en la FC.

Actualmente para hacer la asignación de horarios primero se reune el comité encargado de dicha tarea a realizar manualmente los esqueletos de los horarios. Éstos se dan a conocer a los profesores y ellos eligen diferentes opciones de materias y posibles horas en las cuales les gustaría impartir sus clases. Una vez que los profesores han hecho sus solicitudes, se vuelve a hacer una o varias juntas para la asignación final de los horarios que se hace de manera manual.

Se tienen dos tipos de profesores, los de tiempo completo (TC) y los de asignatura. Los profesores de TC, por contrato, deben de cubrir ciertas horas de clase por lo que al momento de hacer la asignación, éstos tienen prioridad sobre los profesores de asignatura. Finalmente se publican los horarios a los alumnos.

Una vez que los alumnos han elegido las materias que les gustaría tomar deben de ir con el profesor y él o ella les debe de firmar su tira de materias, si es que el cupo del salón lo permite. En caso de que el alumno no consiga la firma de la materia que desea, deberá buscar una segunda o tercera opción o incluso tener que meterla en algún semestre posterior.

La principal razón por la cual los profesores no firman las tiras de materias es porque el número de alumnos que desean inscribirse a su clase es mayor al número de lugares disponibles en el salón asignado. Es por ello que el trabajo que hemos realizado depende de la demanda de alumnos por materia y por horario.

1.2. Definición de conceptos

Las siguientes son las definiciones que se utilizarán a lo largo del trabajo:

Materia: Curso impartido en la Facultad de Ciencias por algún profesor.

Horario: Hora en la que se imparte alguna materia.

Esqueleto: Conjunto Materia-Horario.

Asignación: Conjunto Materia-Horario-Profesor.

Grupo: Clave con la que se identifica una asignación.

Turno Matutino: Comprende las clases impartidas de 7:00-14:00hrs incluyendo la clase de 14:00-15:00hrs.

Turno Vespertino: Comprende las clases impartidas de 15:00-21:00hrs incluyendo la clase de 21:00-22:00hrs.

1.3. Nomenclatura

m : Número de materias que se van a impartir

p : Número de profesores que van impartir alguna materia

i : Índice para profesores, $i \in \{1,2,3,\ldots,p\}$

j : Índice para materias, $j \in \{1,2,3,\ldots,m\}$

h : Índice para las horas del día, $h \in \{1,2,3\ldots,15\}$

$U_{j,i,h}$: Utilidad de que el profesor i imparta la materia j a la hora h

$x_{j,i,h}$: Variable binaria que vale 1 si la materia j es impartida por el profesor i a la hora h y cero en otro caso

$V_{j,i}$: Variable binaria que vale 1 si la materia j puede ser impartida por el profesor i y cero en otro caso

s : Semestre a simular

k : Número de semestres que se tienen como ventana de información

m_grande: Matriz en la que se guarda la información por semestres

r : Matriz *m_filtrada*, submatriz de *m_grande*

vec_sem_sig: Vector con los semestres que se van a simular

X_4 : Analizar presentación: Hacer varias pruebas con distintas combinaciones y elegir el mejor estilo/presentación

X_{14} : Revisar/Investigar al respecto del problema y resolverlo

num_sim : Número de simulaciones de la demanda de alumnos para s

E : Matriz de 15 renglones y tantas columnas como materias se tienen. En cada entrada se tiene la información del número de alumnos simulados en los grupos al crear *mat_esqueleto*.

D : Matriz de 15 renglones y tantas columnas como materias se tienen. En la entrada (i, j) se tiene la información de la demanda de alumnos para la i-ésima hora y la materia j.

bin_DUE : Matriz binaria de 15 renglones y tantas columnas como materias se tienen. Tiene un 1 en la entrada (i, j) si E_{ij} o D_{ij} tienen un valor distinto de cero. Tiene un cero e.o.c.

1.4. Planteamiento del problema

En el problema de asignación de horarios se quiere asociar un profesor con una materia, un salón y un horario. Existen trabajos que han aborado este problema desde otro punto de vista, por ejemplo Yazdani, Naeri y Zeinali, en su artículo *Algorithms for university course scheduling problems* [9], proponen un modelo en el cual se toman 2 decisiones: la asignación de profesor por materia y el salón en el cual se va a impartir cada materia.

Con la función objetivo planteada en dicho modelo se desea maximizar la utilidad de que el profesor i imparta la materia j, más la utilidad de que el profesor i dé clases el día t, más la utilidad de la materia j por ser impartida en el día t.

Dos diferencias principales entre su modelo y el que proponemos en este trabajo son:

1) No tomamos en cuenta el día en el que se imparte la materia porque suponemos que todas las materias se imparten de lunes a viernes, a la misma hora, en el mismo salón.

2) Deseamos maximizar la utilidad de que el profesor i imparta la materia j a la hora h.

Los elementos que consideramos en nuestro modelo son:

- Esqueletos de horario: Matriz de 15 renglones con las horas (7-8, 8-9, ..., 21-22) y tantas columnas como materias. La entrada (i, j) contiene el número de grupos simulados de la i-ésima hora para la materia j.

- Función calificadora de esqueletos: Califica de acuerdo a qué tan bien o que tan mal se cubre la demanda de los alumnos esperados.

- Conjunto de materias: Nombres de las materias impartidas en la FC.

- Conjunto de profesores: Nombres de profesores de TC y de asignatura.

I) Variables de decisión:

$$x_{j,i,h} = \begin{cases} 1 & \text{si la materia } j \text{ es impartida por el profesor } i, \text{ a la hora } h \\ 0 & \text{e.o.c.} \end{cases}$$

II) Función objetivo: (se desea maximizar la utilidad)

$$\text{máx } z = \sum_{i=1}^{p}\sum_{j=1}^{m}\sum_{h=1}^{15} x_{j,i,h}U_{j,i,h} \text{ s. a}$$

III) Restricciones:

$$\begin{aligned}
\sum_{i=1}^{p}\sum_{h=1}^{15} x_{j,i,h} &= 1 \quad \forall\, j && (1.1)\\
\sum_{j=1}^{m} x_{j,i,h} &\leqslant 1 \quad \forall\, i,h && (1.2)\\
\sum_{h=1}^{15} x_{j,i,h} &\leqslant V_{i,j} \quad \forall\, i,j && (1.3)\\
x_{j,i,h},\, V_{j,i} &\in \{0,1\} \quad \forall\, j,i,h && (1.4)
\end{aligned}$$

Con las restricciones del tipo (1.1) aseguramos que todas las materias sean dadas. Con las del tipo (1.2) aseguramos que cada profesor no tenga más de un curso por hora. Con las del tipo (1.3) aseguramos que los profesores tengan asignadas materias que puedan impartir. Finalmente con las restricciones del tipo (1.4) se especifica que las variables utilizadas son binarias.

En el planteamiento se tienen dos tipos de restricciones: duras y suaves. Las restricciones duras son las que nos permiten tener soluciones factibles al cumplirlas en su totalidad y las restricciones suaves nos permiten evaluar la calidad de las diferentes soluciones. Usualmente las restricciones suaves están asociadas a preferencias y se cumplen en la medida de lo posible, pero no afectan la factibilidad de las soluciones.

El conjunto de soluciones se presenta por medio de la matriz *mat_asignaciones* la cual es una matriz de tres columnas y tantos renglones como grupos se hayan simulado. En el i-ésimo renglón se tiene la información de la i-ésima materia con su respectivo profesor y horario asignados. En la tabla 1.1 se muestra un ejemplo del resultado de la asignación.

Materia	Profesor	Horario
Inferencia Estadística	Margarita Elvira Chávez Cano	9-10
Estadística II	Jaime Vázquez Alamilla	10-11
Estadística Bayesiana	Ruth Selene Fuentes García	11-12
Estadística III	Lizbeth Naranjo Albarrán	13-14

Tabla 1.1: *Ejemplo de asignación: Esta tabla muestra un ejemplo de la matriz mat_asignaciones que tiene 3 columnas (Materia, Profesor, Horario).*

1.5. Objetivos

El primer objetivo del trabajo es hacer dos funciones que generen:

i) Esqueletos de horarios

ii) Una asignación de profesores por materia y por horario. La asignación debe cubrir la demanda de alumnos estimada para el semestre siguiente.

Los esqueletos de horarios son utilizados para simular una posible elección de materias y horarios de los profesores para finalmente hacer la asignación correspondiente a la hora, materia y profesor de cada grupo.

El segundo objetivo es disminuir el tiempo utilizado actualmente para la realización de la asignación de horarios.

1.6. Datos a analizar

Para poder realizar un análisis adecuado de los datos, hicimos dos divisiones de ellos. La primera fue con respecto al tipo de semestre, par o impar y la segunda con respecto al turno, matutino o vespertino.

Para explicar la elección de las divisiones tomamos la información de la materia *Probabilidad I*, desde el semestre 2015-1 hasta el 2020-1. Cabe aclarar que dicha materia en la carrera de Actuaría es una materia obligatoria de tercer semestre.

Veamos primero la división de acuerdo a semestres pares e impares.

En la figura 1.1 vemos una gráfica en la que la línea azul representa el número de alumnos de los semestres impares y la línea roja representa el número de alumnos de los semestres pares. Observamos que en todo momento el número de alumnos de los semestres impares es mayor al número de alumnos de los semestres pares. Ésto nos interesa porque al momento de simular debemos tomar en cuenta que el número de alumnos totales de semestres impares debe de ser siempre mayor al número total de alumnos de los semestres pares.

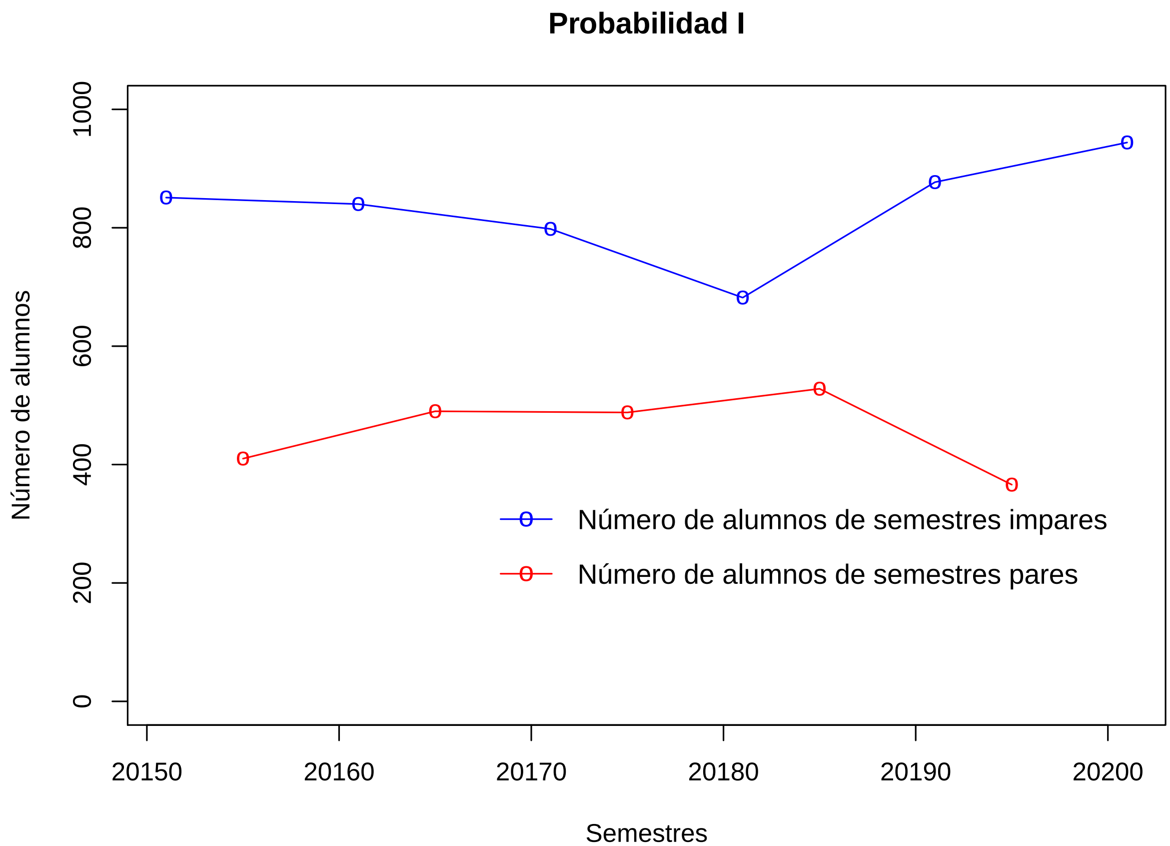


Figura 1.1: *Número de alumnos totales por semestres pares e impares (Probabilidad I): Observamos que el número de alumnos de semestres impares es siempre mayor al número de alumnos de semestres pares.*

Continuando con los datos de *Probabilidad I*, obtuvimos la gráfica de la figura 1.2 que contiene dos histogramas, las barras rojas representan el número de alumnos por grupo de semestres pares y las barras azules representan el número de alumnos por grupo de semestres impares.

Las líneas que se encuentran sobre los histogramas son densidades estimadas que se ajustan a los histogramas. Algunos datos que se pueden obtener de esas densidades son por ejemplo que alrededor del 20 % de los grupos de los semestres pares tienen aproximadamente de 60 a 70 alumnos y que alrededor del 3 % de los grupos de los semestres impares tienen entre 150 y 180 alumnos.

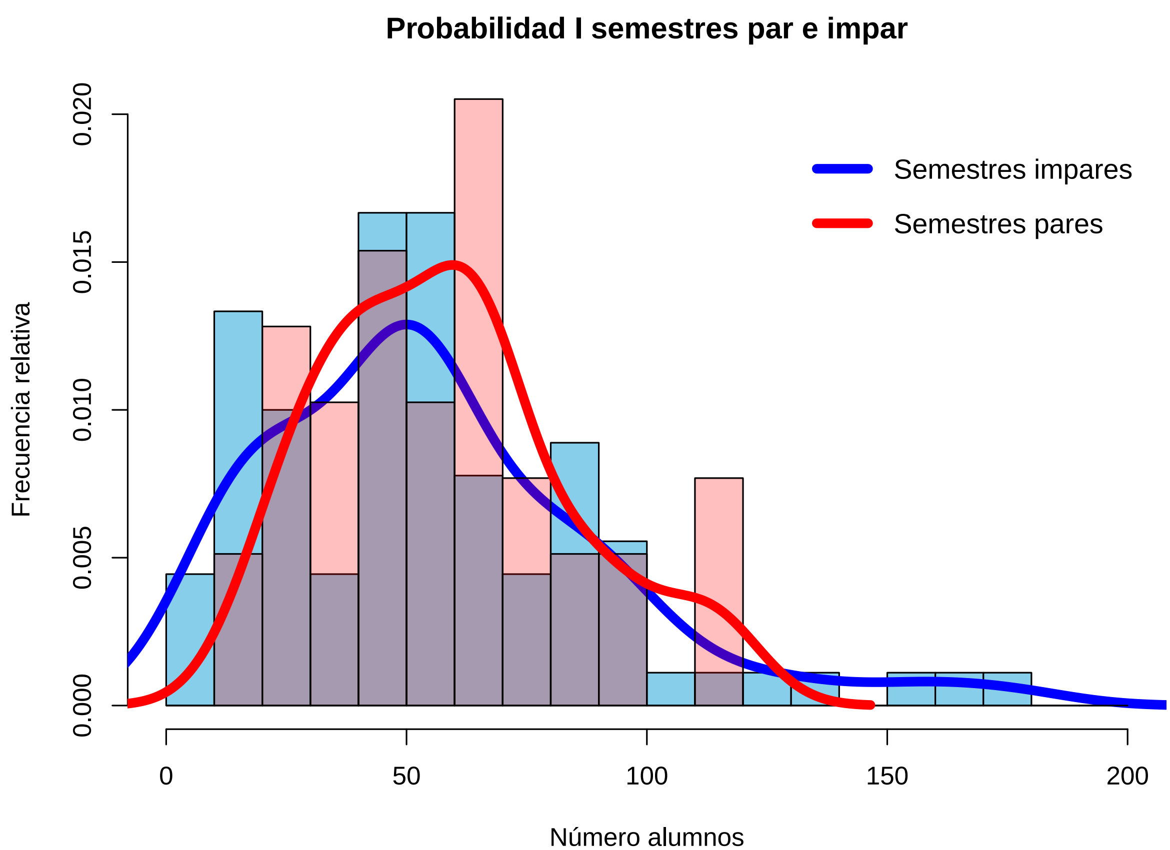


Figura 1.2: *Histograma del número de alumnos por semestre (Probabilidad I): Los datos se dividen en semestres pares e impares. Las densidades ajustadas son muy parecidas.*

Ahora veamos la segunda división de acuerdo a los turnos matutino y vespertino.

En la gráfica de la figura 1.3 la línea azul representa el número de alumnos del turno matutino y la línea roja representa el número de alumnos del turno vespertino. Se puede observar que en todo momento el número de alumnos del turno matutino es mayor al número de alumnos del turno vespertino. Ésto impacta en el hecho de que por semestres la varianza en el turno matutino es mucho mayor que en el turno vespertino. Lo cual indica que en el turno vespertino se tiene prácticamente el mismo número de alumnos sin importar si la materia pertenece a un semestre par o impar, a diferencia de lo que ocurre en el turno matutino en donde si influye el hecho de que la materia corresponda a un semestre par o impar.

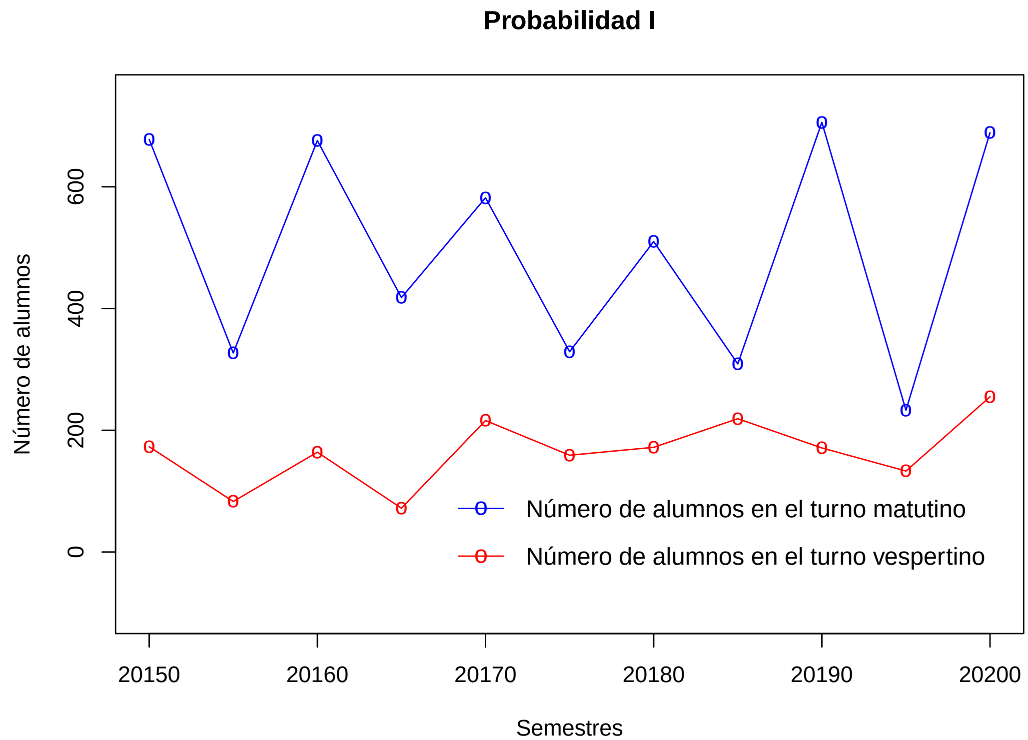


Figura 1.3: *Número de alumnos por turno (Probabilidad I): Los datos se dividen por turno matutino y vespertino. El número de alumnos del turno matutino es siempre mayor al número de alumnos del turno vespertino.*

Al igual que en la primera división se obtuvo una gráfica que contiene dos histogramas, sobre los cuales se tienen 2 líneas con densidades estimadas que se ajustan a los histogramas. Dicha gráfica se observa en la figura 1.4 en la cual podemos ver que las barras rojas representan el número de alumnos del turno vespertino y las barras azules representan el número de alumnos del turno matutino.

Notamos que en este caso las densidades son completamente diferentes. Algunos datos que se pueden obtener de dichas densidades son por ejemplo que alrededor del 20 % de los grupos del turno vespertino tienen aproximadamente entre 10 y 20 alumnos y un poco más del 10 % de los grupos del turno matutino tienen entre 80 y 90 alumnos.

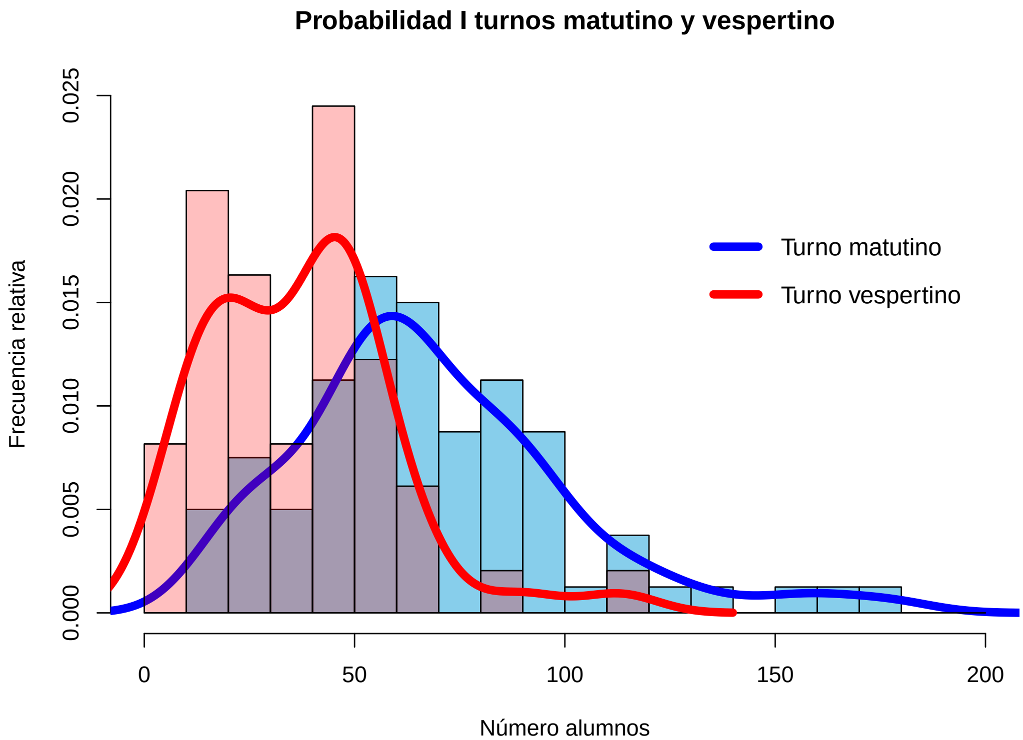


Figura 1.4: *Histograma del número de alumnos por turno (Probabilidad I): Los datos se dividen por turno matutino y vespertino. Las densidades ajustadas son muy diferentes.*

Con los resultados observados se obtuvieron los grupos de datos G_1, G_2, G_3, G_4, para hacer los análisis estadísticos, los cuales se definen en la tabla 1.2.

Sem. \ Turno	**Matutino**	**Vespertino**
Impar	G_1	G_2
Par	G_3	G_4

Tabla 1.2: *Grupos de datos: Obtuvimos 4 grupos con las combinaciones de los turnos (matutino y vespertino) con los tipos de semestres (pares o impares).*

Capítulo 2

Extracción de datos

La fuente de información de donde obtuvimos los datos utilizados son las páginas de los horarios de la Facultad de Ciencias de la UNAM. Cada página contiene toda la posible información de los grupos de una materia, un semestre y una carrera. Sólo tomamos en cuenta la información de las carreras del Departamento de Matemáticas, las cuales son: Actuaría, Ciencias de la Computación, Matemáticas y Matemáticas Aplicadas.

La información que se puede extraer de las páginas mencionadas es: nombre de profesores, nombre de ayudantes, salón, horario, plan, carrera, año y tipo de semestre (ej. *2018-2* implica que es el semestre par del 2018), materia, semestre de la materia. (ej. *Quinto Semestre*), tipo de materia (obligatoria u optativa) y fechas, horarios y salones de exámenes finales. En la figura 2.1 se muestra un ejemplo de dichas páginas.

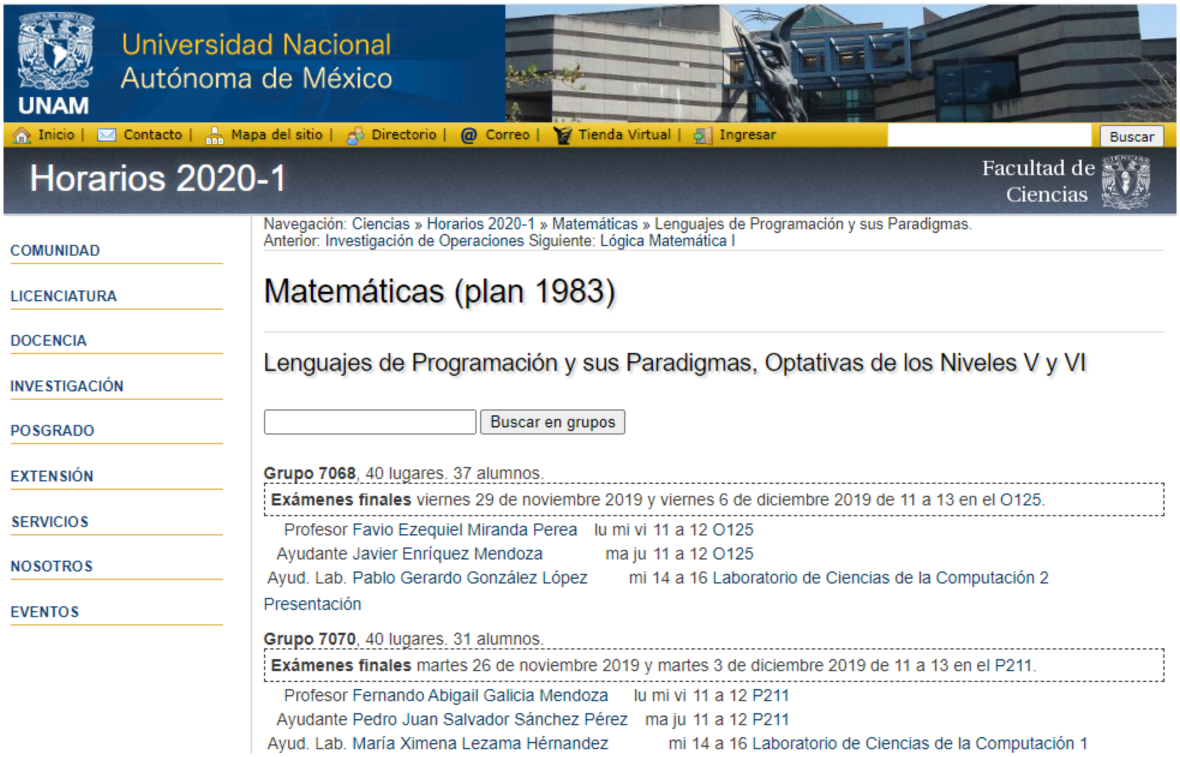


Figura 2.1: *Página de horarios de la Facultad de Ciencias: Muestra la información de los horarios de la materia "Lenguajes de Programación y sus Paradigmas", de la carrera de Matemáticas, plan 1983, del semestre 2020-1.*

2.1. Extracción de información con la aplicación SelectorGadget

Para extraer los datos de las páginas de la FC utilizamos una aplicación de Google Chrome llamada *SelectorGadget*, la cual permite seleccionar la información deseada y ésta arroja una sección del código CSS de la página web el cual se introduce en R para que se descargue la información deseada.

En la figura 2.2 se puede ver un ejemplo del funcionamiento de la aplicación. El ícono de la aplicación es una lupa, el cual se encuentra señalado por la flecha roja. Una vez presionado el ícono seleccionamos el nombre de un profesor (en color verde), la aplicación automáticamente seleccionó todas las entradas que coindicidían (en color amarillo). Debido a que se seleccionó más información de la deseada entonces dimos click sobre un salón y el nombre de un ayudante (en color rojo). En el cuadro de texto, la aplicación arroja la sección del código CSS correspondiente. Junto al código mencionado, se puede observar que se seleccionaron 15 entradas correspondientes a los nombres de los profesores de la materia de *Probabilidad I*, de la carrera de Actuaría, plan 2015.

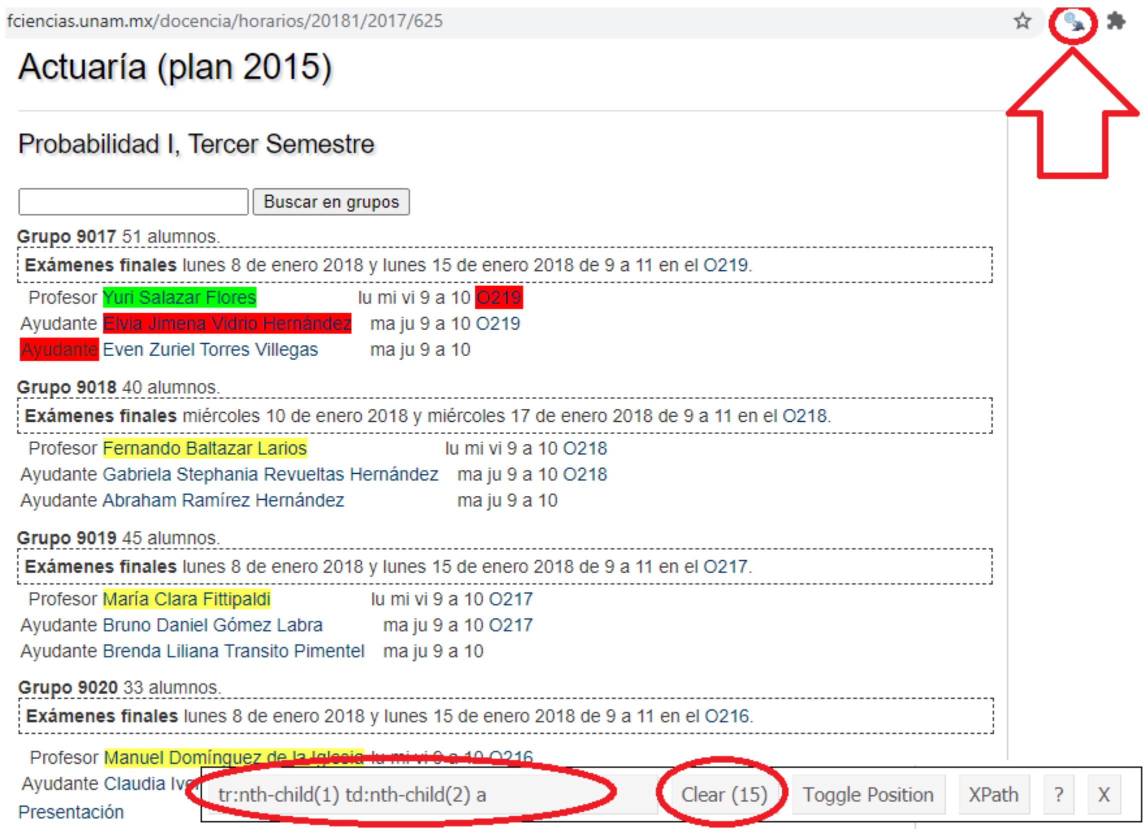


Figura 2.2: *Aplicación SelectorGadget: Se muestra cómo se ve una página de internet al usar la aplicación mientras se selecciona la información que deseamos extraer.*

2.2. Estructura de las páginas web

El estar haciendo la búsqueda de información, de manera manual, de los horarios para cada materia, cada carrera y cada uno de sus planes de estudio, requiere de mucho tiempo de trabajo. Decidimos buscar la existencia de alguna estructura en las páginas web de los horarios de la FC para poder realizar la búsqueda de la información de una manera automática y mucho más rápida. Observamos que la estructura que siguen todas las páginas de internet es la siguiente:

http://www.fciencias.unam.mx/docencia/horarios/a/b/c

Se tiene una raíz común para todas las páginas y al final se tienen tres números los cuales representan:

a = año y tipo de semestre (par o impar), $a \in \{20081, 20082, 20091, 20092, 20101, 20102, \ldots, 20201\}$

b = clave de plan de estudio

c = número de materia, $c \in \{1, 1000\}$.

Para la realización de este trabajo tomamos en cuenta sólo los planes de estudio vigentes al día de hoy, los cuales son todos los planes mostrados en la tabla 2.1, salvo el plan 1972 de

Actuaría. Dicha tabla muestra los planes de estudio de cada carrera con su clave correspondiente.

PLAN	CLAVE
Actuaría	
1972	214
2000	119
2006	1176
2015	2017
Ciencias de la Computación	
1994	218
2013	1556
Matemáticas	
1983	217
Matemáticas Aplicadas	
2017	2055

Tabla 2.1: *Planes de estudio por carrera con clave: La clave de cada plan de estudios se sutituye en **b** en la estructura de las páginas web.*

Una vez identificada la estructura de las páginas web pudimos realizar la búsqueda de información de manera automatizada. Originalmente decidimos que $c \in \{1,2,3,\ldots,10000\}$. Hicimos una función que genera una matriz llamada *mat_posibles_url*, en la cual se guardan, entre otros datos, las URL's de las páginas de los horarios de la FC. La función sólo guarda las URL's que si existen. Al obtener dicha matriz, observamos que el valor máximo que toma c es 991, por lo que redujimos su conjunto de posibles valores y definimos $c \in \{1,\ldots,1000\}$.

Las categorías de las columnas de la matriz *mat_posibles_url* están descritas en la tabla 2.2.

Col.	Nombre	Explicación	Posibles valores
1	Semestre	Semestre al que pertenece la materia (Año y semestre)	$1^o, 2^o, ..., 8^o$
2	Plan	Año en el que se implementó un nuevo plan de estudios	1983, 1994, 2000, 2006, 2013, 2015, 2017
3	Materia	Clave del curso impartido	$\mathbb{N}$
4	URL	Nombres de las páginas de los horarios de FC	Páginas web de FC
5	Num. Grupos	Número de grupos que hay en cada página de internet	$\mathbb{N}$

Tabla 2.2: *Descripción de las columnas de la matriz mat_posibles_url: La matriz contiene información de cada URL.*

2.3. Tipos de grupos de las páginas web de la Facultad de Ciencias

Al inicio encontramos tres tipos de grupos dentro de las páginas de horarios de la FC. Cada uno con información similar, pero hicimos la separación de acuerdo a sus diferencias. Cabe mencionar que en este trabajo consideramos como semestre actual al semestre $2020-1$. En todos los grupos se puede encontrar la información del nombre de profesor, nombre del o de los ayudantes, salón, horario y el número de alumnos inscritos en el grupo.

a) En el grupo **A** se tienen las páginas correspondientes al semestre actual. Este grupo tiene la información del número de lugares diponibles por salón, pero no contiene la información de los exámenes finales, porque se considera que el semestres aún está en curso y aún no termina. En la figura 2.3 podemos ver un ejemplo de este tipo de grupo.

Figura 2.3: *Tipo de grupo A: Correspondiente al semestre en curso que aún no finaliza.*

b) En el grupo **B** se tienen las páginas correspondientes a semestres entre el $2018-2$ y el semestre anterior al actual, con respecto al año en curso. En este tipo de grupos se tiene información del número de lugares disponibles por salón y la información de los exámenes finales, porque son semestres que ya finalizaron. En la figura 2.4 encontramos un ejemplo de este tipo de grupo.

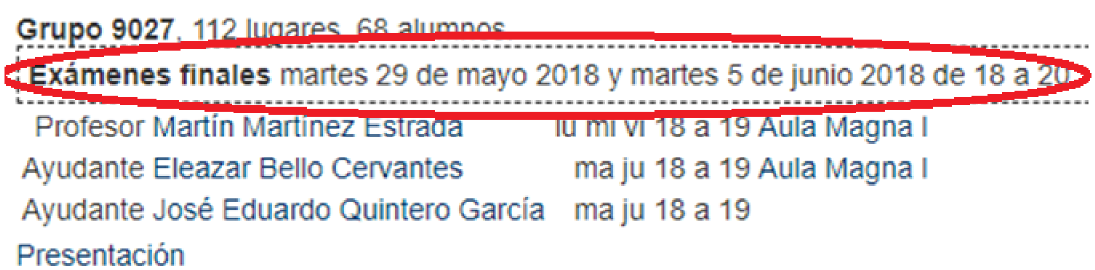


Figura 2.4: *Tipo de grupo B: Correspondiente a semestres ya finalizados, posteriores al semestre 2018-2.*

c) En el grupo **C** se tienen las páginas correspondientes a semestres anteriores al $2018-1$, incluyéndolo. Este tipo de grupos tiene información de los exámenes finales, pero no contiene la información del número de lugares disponibles por salón. En la figura 2.5 podemos ver un ejemplo.

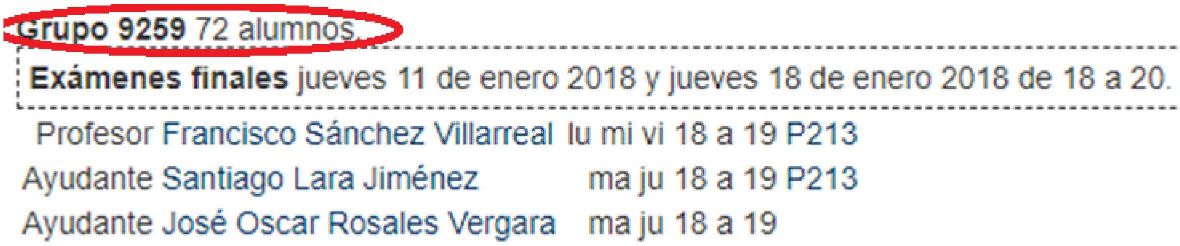


Figura 2.5: *Tipo de grupo C: Correspondiente a semestres ya finalizados, anteriores al semestre 2018-1, incluyéndolo.*

2.4. Limpieza de base de datos

Se puede encontrar que, en general, cuando uno realiza la limpieza de datos se hace el 80 % del análisis de los datos, ya que es en ese momento en donde se encuentran los diferentes problemas que se pueden presentar, como posibles errores en los datos, información incompleta, o valores poco comunes de acuerdo al comportamiento observado.

En nuestro caso al limpiar los datos nos encontramos con problemas de falta de información, problemas de información repetida, problemas al extraer información, entre otros, cada uno de éstos se desglosan en las siguientes subsecciones.

2.4.1. Problemas de falta de información

Encontramos diferentes tipos de páginas que tenían grupos sin información e incluso páginas sin información alguna. Para guardar la información consideramos sólo los grupos que al menos tenían: nombre de profesor, número de alumnos inscritos y horario. A continuación se muestran varios ejemplos con los diferentes casos encontrados.

- En la figura 2.6 vemos un ejemplo de páginas en las cuales se tiene el nombre de la materia, pero no hay información de algún grupo: `http://www.fciencias.unam.mx/docencia/horarios/20081/1556/803`

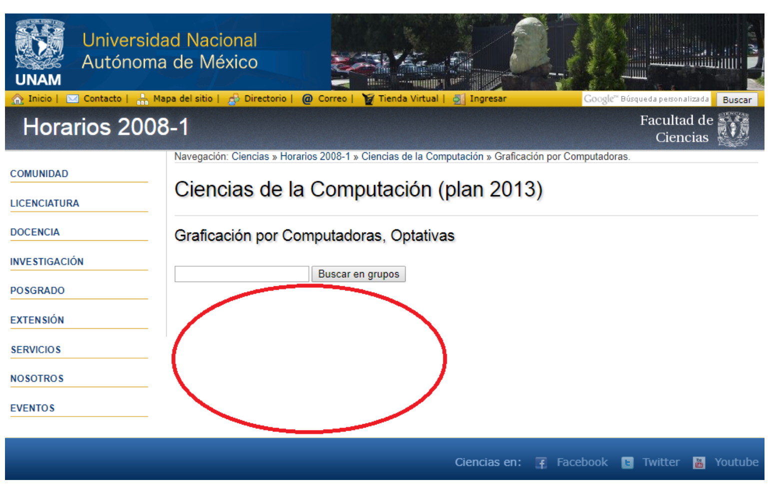


Figura 2.6: *Ejemplo de página web en blanco: En este tipo de páginas no encontramos información de los grupos para la materia.*

- En la figura 2.7 encontramos un ejemplo de páginas que no tienen información del salón: `http://www.fciencias.unam.mx/docencia/horarios/20081/119/4`

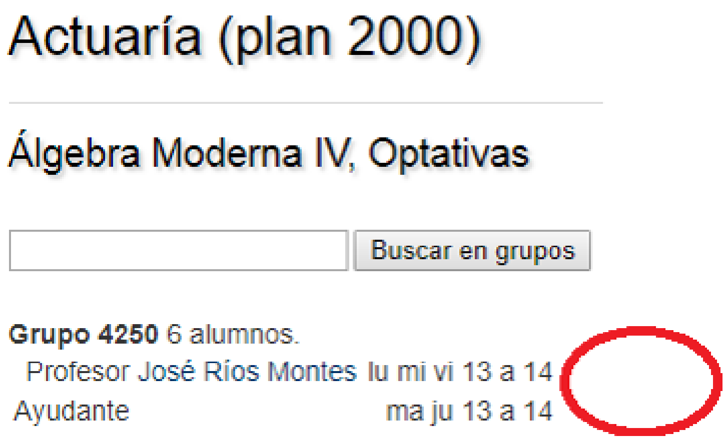


Figura 2.7: *Ejemplo de grupo sin información de salón: En este tipo páginas no se muestra el salón en el que se imparte la clase.*

- En la figura 2.8 tenemos un ejemplo de páginas que tienen grupos sin información del número de alumnos inscritos en el grupo: `http://www.fciencias.unam.mx/docencia/horarios/20112/119/630`

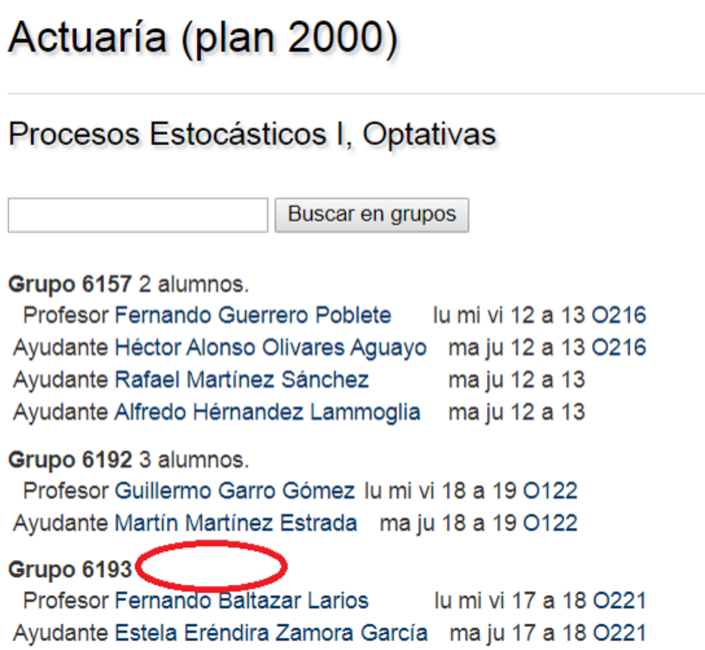


Figura 2.8: *Ejemplo de grupo sin información de alumnos: En este tipo páginas encontramos grupos que no tienen el número de alumnos inscritos.*

- En la figura 2.9 vemos un ejemplo de páginas que tienen grupos sólo con el horario, sin nombre del profesor, salón, ayudante, número de alumnos, lugares disponibles: `http://www.fciencias.unam.mx/docencia/horarios/20091/119/841`

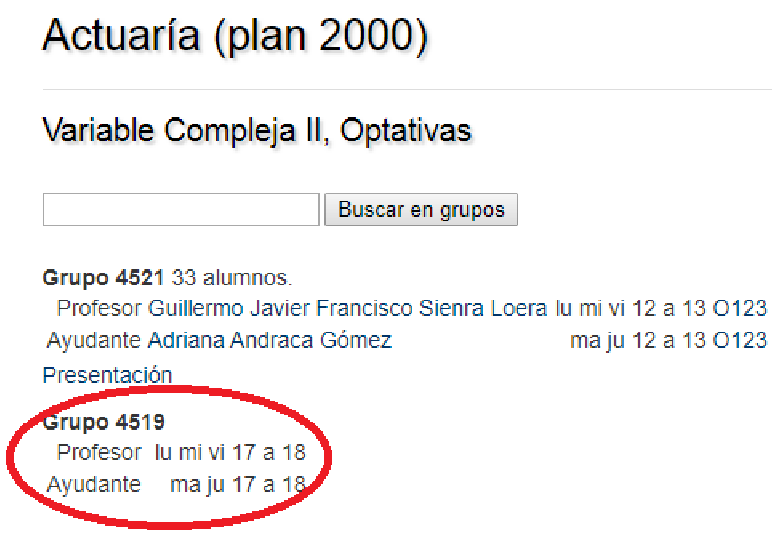


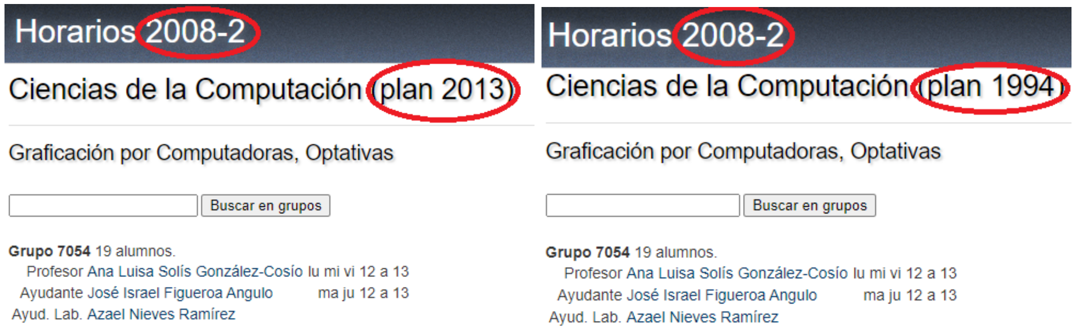
Figura 2.9: *Ejemplo de grupo sólo con horario: En este tipo páginas existen grupos que no tienen información del profesor o salón ni del número de alumnos inscritos, sólo tienen la clave del grupo y el horario.*

2.4.2. Problemas de información repetida

Dentro de los problemas de información repetida, encontramos diferentes casos. Para guardar la información juntamos aquellos grupos que provenían del mismo grupo. A continuación presentamos los casos que encontramos con el problema de tener información repetida.

- En la figura 2.10 mostramos un ejemplo en donde se tiene información de una materia correspondiente a un plan de estudios posterior al semestre en el que se está buscando la información: `http://www.fciencias.unam.mx/docencia/horarios/20082/1556/803` y tener la misma información con el plan de estudios correspondiente: `http://www.fciencias.unam.mx/docencia/horarios/20082/218/803`

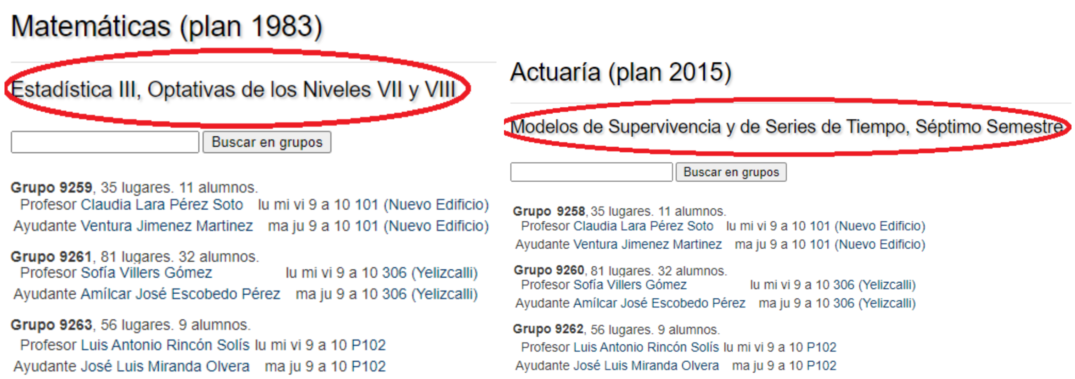
  El número del plan de estudios corresponde al año en que entró en vigencia el plan, por lo que no debería de existir un horario con un plan posterior al año del semestre. En la subfigura (a) de la figura 2.10 podemos ver una materia de la carrera de Ciencias de la Computación del semestre 2008-2, con el plan 2013, lo cual no es lógico. En la subfigura (b) de la misma figura, vemos la infomración de la misma materia y del mismo grupo pero con el plan 1994.



(a) *Plan de estudios posterior* (b) *Plan de estudios correspondiente*

Figura 2.10: *Ejemplo de información repetida (Planes de estudio): No deberían de existir grupos con planes posteriores al año del semestre en el que se busca información.*

- En la figura 2.11 vemos un ejemplo en donde se tiene una misma materia con nombres distintos para las diferentes carreras: `http://www.fciencias.unam.mx/docencia/horarios/20201/217/1712` para Matemáicas, plan 1983 y `http://www.fciencias.unam.mx/docencia/horarios/20201/2017/1739` para Actuaría, plan 2015. Notamos que la información en ambas páginas es la misma, sólo cambian las claves de los grupos.

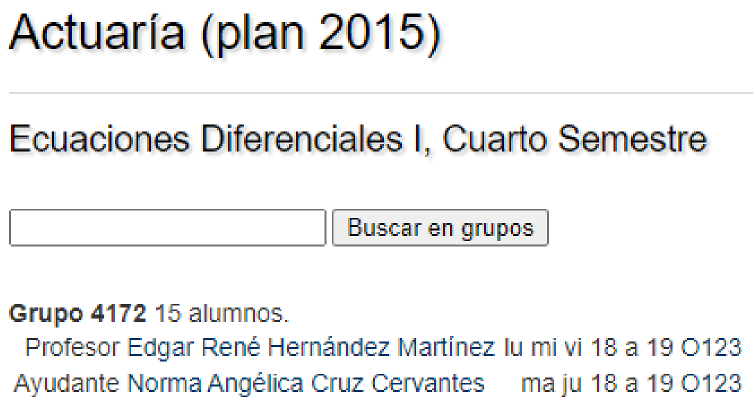


(a) *Matemáticas: 1983* (b) *Actuaría: 2015*

Figura 2.11: *Ejemplo de información repetida: Materia con nombres distintos: En estos casos se tienen materias que tienen nombres diferentes de acuerdo a la carrera o plan de estudios.*

- En la figura 2.12 tenemos un ejemplo de profesores que imparten dos o más clases distintas en el mismo horario y diferente salón: `http://www.fciencias.unam.mx/docencia/horarios/20111/2017/162` para Ecuaciones Diferenciales I y `http://www.fciencias.unam.mx/docencia/horarios/20111/2017/91` para Cálculo Diferencial e Integral I.

  Las materias mencionadas son diferentes, pero las clases comienzan a la misma hora, Ecuaciones de 18-19hrs y Cálculo de 18-20hrs, dado que se tiene la misma ayudante pudiera ser que se intercambien las horas, pero no se puede asignar más de una clase a la misma hora al mismo profesor.



(a) *Ecuaciones Diferenciales I*

(b) *Cálculo Diferencial e Integral I*

Figura 2.12: *Ejemplo de información repetida (mismo profesor, materias distintas): En este caso se tiene más de una clase impartida por el mismo profesor a la misma hora en diferente salón lo cual no debería de ocurrir.*

2.4.3. Otros problemas al extraer información

Al extraer la información surgieron otros problemas. En algunos casos se tuvieron que analizar las materias a mano para poder guardar la infomración de manera adecuada. A continuación se presentan los diferentes casos encontrados:

- Dentro de la obtención de datos del número de alumnos, no se lee la información cuando se tiene *Un alumno*, ya que no se reconoce el texto *Un* como el número 1. En la figura 2.13 vemos un ejemplo de este caso.

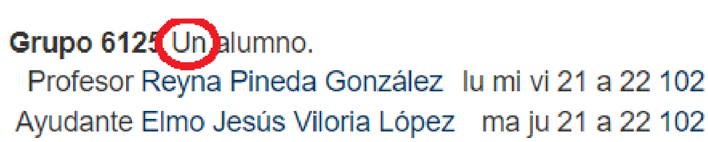


Figura 2.13: *Ejemplo de grupo con un alumno: En este caso se tiene el texto "Un" y no un número "1".*

Para resolver este problema se identificó la variable tipo *string* igual a *Un* para convertir la información y así poder utilizar los datos obtenidos.

- El algoritmo supone que todas las clases duran una hora y no se consideran las medias horas: `http://www.fciencias.unam.mx/docencia/horarios/20172/1556/820`. En la figura 2.14 mostramos un ejemplo en donde se considera que esa materia inicia a las 18hrs.

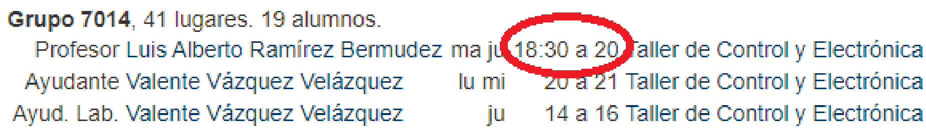


Figura 2.14: *Ejemplo de grupo con medias horas: Se considera que las materias inician en horas enteras y no a las medias horas.*

- Se tienen materias con múltiples horarios:`http://www.fciencias.unam.mx/docencia/horarios/20181/2055/1323`. En estos casos sólo se registran los horarios y salones en los que los profesores imparten su clase, no se toman en cuenta las clases impartidas por los ayudantes.

  En la figura 2.15 tenemos un ejemplo de este caso en donde el profesor imparte su clase los lunes, miércoles y viernes de 13-14hrs en el salón O215, hay una ayudantía los martes y jueves de 13-14hrs en el salón O215 y otra ayudantía los martes de 11-13hrs en el salón 304 (Yelizcalli). Se considera que esta materia inicia a las 13hrs y se imparte en el salón O215.

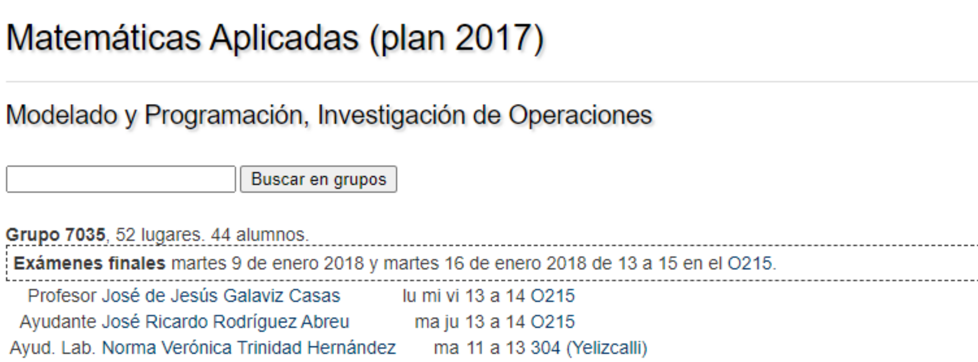


Figura 2.15: *Ejemplo de grupo con horarios múltiples: En estos grupos sólo se toman en cuenta los horarios y salones en los que los profesores imparten clase.*

- Las materias de inglés no se imparten todos los días de la semana, en algunos casos se imparten clases en línea: `http://www.fciencias.unam.mx/docencia/horarios/20202/2017/1135`. Se registran únicamente los horarios de los días en que se imparten las clases presenciales. En la figura 2.16 mostramos un ejemplo de este caso.

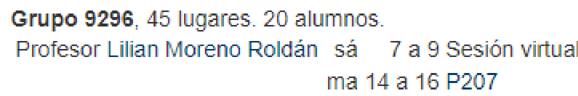


Figura 2.16: *Ejemplo de grupo de inglés: Las clases no se imparten todos los días. Hay sesiones virtuales. Sólo se toma en cuenta el horario de las clases presenciales.*

- Se tienen grupos que no tienen la misma estructura que los tipos de grupos **A**, **B** y **C** definidos en la sección 2.3: `http://www.fciencias.unam.mx/docencia/horarios/20201/2017/872`, debido a ello el código CSS utilizado no sirve para obtener toda la información que se puede obtener del grupo. En la figura 2.17 tenemos un ejemplo de este caso en donde no se lee adecuadamente el número de alumnos inscritos en el grupo.

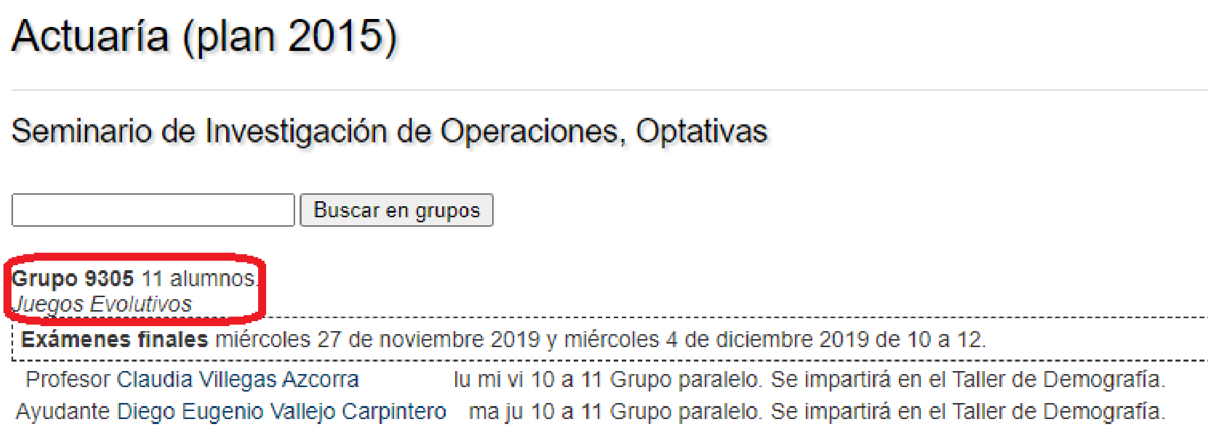


Figura 2.17: *Ejemplo de grupo con estructura diferente: En estos casos no se extrae adecuadamente la infomración de los grupos porque el código CSS utilizado no corresponde a este tipo de grupos.*

2.5. Matrices de datos

Una vez que se realizó el proceso de la limpieza de los datos obtenidos, éstos se guardaron, por semestre, en matrices llamadas *m_grande*, los nombres de sus columnas con su respectiva explicación y posibles valores, se muestran en la siguiente tabla:

Col.	Nombre	Explicación	Posibles valores
1	Materia	Nombre de algún curso impartido en la FC	"Probabilidad I"
2	Profesor	Nombre de la persona que va a impartir alguna materia	"Arrigo Coen Coria"
3	Horario	Hora en la que se imparte alguna materia	"7 a 8", ... , "21 a 22"
4	horario_num	Valores de la columna Horario en variables tipo *numeric*	$7,8,9,\ldots,20,21$
5	Lugares	Espacios disponibles por salón	$\mathbb{N}$

La tabla continúa en la siguiente página

Col.	Nombre	Explicación	Posibles valores
6	Alumnos	Número de estudiantes inscritos por grupo	$\mathbb{N}$
7	Salón	Espacio físico en el que se imparte alguna materia	“O218”, ..., “P105”
8	Grupo	Clave con la que se identifica una asignación	4489, 6114, ...
9	Carrera	Nombre de alguna carrera de la FC	“Actuaría”, “Matemáticas”,...
10	Plan	Año en el que se implemento un nuevo plan de estudios	1983, 1994, ..., 2017
11	Semestre	Semestre al que pertenece la materia (Año y semestre par o impar)	20081, ... , 20192, 20201
12	Cambios	Clave que indica los cambios que se le han hecho al grupo	$\mathbb{N}$
13	Turno	Matutino: 7:00-14:00hrs, Vespertino: 15:00-21:00	M,V
14	Semestre_de_materia	Semestre en el que el plan de estudios dicta que se lleva esa materia	“Primer Semestre”, ... , “Optativas”
15	url	Nombre de la página de los horarios de FC correspondiente al grupo	url’s de FC
16	Act2000	Columna binaria, indica si el grupo pertenece a la carrera de Actuaría, plan 2000	$0, 1$
17	Act2006	Columna binaria, indica si el grupo pertenece a la carrera de Actuaría, plan 2006	$0, 1$
18	Act2015	Columna binaria, indica si el grupo pertenece a la carrera de Actuaría, plan 2015	$0, 1$
19	CdC1994	Columna binaria, indica si el grupo pertenece a la carrera de CdC, plan 1994	$0, 1$
20	CdC2013	Columna binaria, indica si el grupo pertenece a la carrera de CdC, plan 2013	$0, 1$
21	Mat1983	Columna binaria, indica si el grupo pertenece a la carrera de Matemáticas, plan 1983	$0, 1$
22	MatAp2017	Columna binaria, indica si el grupo pertenece a la carrera de MatAp, plan 2017	$0, 1$
23	NomMat_Act2000	Indica el nombre de las materia correspondiente a la carrera de Actuaría plan 2000	Nombres de materias de FC
24	NomMat_Act2006	Indica el nombre de las materia correspondiente a la carrera de Actuaría plan 2006	Nombres de materias de FC
25	NomMat_Act2015	Indica el nombre de las materia correspondiente a la carrera de Actuaría plan 2015	Nombres de materias de FC
26	NomMat_CdC1994	Indica el nombre de las materia correspondiente a la carrera de CdC plan 1994	Nombres de materias de FC
27	NomMat_CdC2013	Indica el nombre de las materia correspondiente a la carrera de CdC plan 2013	Posibles valores
28	NomMat_Mat1983	Indica el nombre de las materia correspondiente a la carrera de Matemáticas plan 1983	Nombres de materias de FC

La tabla continúa en la siguiente página

Col.	Nombre	Explicación	Posibles valores
29	NomMat_MAp2017	Indica el nombre de las materia correspondiente a la carrera de MatAp plan 2017	Nombres de materias de FC
30	URL_Act2000	Indica la URL correspondiente a la carrera de Actuaría plan 2000	url de FC
31	URL_Act2006	Indica la URL correspondiente a la carrera de Actuaría plan 2006	url de FC
32	URL_Act2015	Indica la URL correspondiente a la carrera de Actuaría plan 2015	url de FC
33	URL_CdC1994	Indica la URL correspondiente a la carrera de CdC plan 1994	url de FC
34	URL_CdC2013	Indica la URL correspondiente a la carrera de CdC plan 2013	url de FC
35	URL_Mat1983	Indica la URL correspondiente a la carrera de Matemáticas plan 1983	url de FC
36	URL_MAp2017	Indica la URL correspondiente a la carrera de MatAp plan 2017	url de FC
37	Num_materia	Número de materia de acuerdo al vector *vec_nom_materias*	$\mathbb{N}$

Tabla 2.3: *Descripción de las columnas de la matriz m_grande: En esta tabla se describe el contenido de las columnas de las matrices en las que se guarda la información por semestres.*

La columna *Cambios*, va a guardar todos los cambios que han "sobrevivido" esos grupos. El significado de los números que pueden aparecer en esa columna se explican a continuación:

(1) Grupos revisados a mano.

(2) Se anotaron los días en los que se imparte la materia, en la columna *Horario*, por ejemplo cuando había conflicto debido a que el profesor impartía más de una materia a la misma hora, al revisar el caso se encontró que los días en los que se impartía la clase era distinto.

(3) Se eliminaron los grupos repetidos, al juntar la información en un mismo grupo.

(4) Páginas que no tienen información del salón.

Capítulo 3

Análisis estadístico

Debido a que los datos obtenidos no son independientes entre sí, las herramientas elegidas para realizar un análisis estadístico de los datos fueron las series de tiempo. A continuación se describe su definición y aplicación para explicar el motivo de la elección de dichas herramientas estadísticas.

Definimos a una serie de tiempo como una secuencia de observaciones X_t ordenadas cronológicamente, en donde los datos al tiempo presente dependen de las observaciones anteriores, es decir existe una depencia de X_t con $\{X_{t-1}, X_{t-2}, X_{t-3}, \ldots, X_1\}$.

Denotamos a una serie de tiempo como:

$$X_t = m_t + s_t + y_t \tag{3.1}$$

Cada elemento de la serie de tiempo es llamado componente. A continuación se describe cada una de ellas.

- Tendencia (m_t): Se le llama tendencia al cambio, a largo plazo, del promedio de los datos. El cambio puede ser creciente o decreciente.

- Estacionalidad (s_t): Se llama variación estacional a las fluctuaciones periódicas que tiene una serie de tiempo. La longitud de cada periodo es constante y menor o igual a un año, por ejemplo semanal, mensual o semestral.

- Aleatoriedad (y_t): También llamada componente irregular, son series de residuales que pueden o no ser aleatorios.

Chatfield y Xing nos indican, en su libro *The Analysis of Time Series An Introduction with R* [2], que existen 2 tipos de variación estacional:

- Aditiva: Se dice que la estacionalidad es aditiva cuando la longitud de cada periodo es constante año con año.

- Multiplicativa: Se dice que la estacionalidad es multiplicativa cuando la longitud de cada periodo es directamente proporcional a la media de los datos de la serie de tiempo.

Con estos tipos de variaciones se forman 3 modelos de estacionalidad:

1. Aditivo: Se tiene variación estacional aditiva y se utiliza cuando la varianza o la desviación estándar de la serie de tiempo se mantienen constantes a lo largo del tiempo.

$$X_t = m_t + s_t + y_t \tag{3.2}$$

2. Multiplicativo: Se tiene variación estacional multiplicativa. Se utiliza cuando la varianza o la desviación estándar de los datos cambian a través del tiempo. Su variabilididad puede ser mayor o menor conforme pasa el tiempo.

$$X_t = m_t s_t y_t \tag{3.3}$$

3. Mixto: Se utiliza cuando se tiene variación estacional multiplicativa pero la variabilidad de la componente irregular se mantiene constante a lo largo del tiempo.

$$X_t = m_t s_t + y_t \tag{3.4}$$

Los objetivos principales al hacer el análisis de una serie de tiempo son:

- Describir: Leer datos en una tabla es mucho más tardado y en algunas ocasiones más complicado que observar una gráfica de los datos que se tienen. Las gráficas ayudan a ver de una manera más inmediata el comportamiento que tienen los datos y es posible observar si la serie de tiempo tiene alguna tendencia o estacionalidad. También se puede ver la posible falta de información o valores atípicos.

- Predecir: Teniendo una serie de tiempo se desea conocer qué va a pasar en el futuro. Es conveniente tener varios periodos de información para que la predicción sea lo más acertada posible.

Las áreas en las que se pueden aplicar las series de tiempo son por ejemplo en economía, demografía, finanzas, medio ambiente, ingeniería o medicina. Algunos ejemplos más precisos de su aplicación son: precios de acciones diarios, niveles de producción en la agricultura mensuales, medición del sonido por segundos, electrocardiogramas, medición de terremotos, tasa de mortalidad, tasa de natalidad, barriles de petróleo producidos al año, entre otros.

3.1. Análisis estadístico básico

En esta sección haremos un análisis de los datos obtenidos. Las técnicas de suavizamiento de series de tiempo son útiles para mostrar patrones subyacentes en los datos de las series de tiempo. El método que vamos a utilizar para mostrar dichos patrones y para realizar predicciones de los datos es el método Holt-Winters aditivo.

El método se utiliza para describir y predecir valores con series de tiempo que tienen componentes de tendencia y de estacionalidad. Existen pruebas, que hemos hecho a los datos, para comprobar que los datos que obtuvimos cumplen con los supuestos del método. Éstas pruebas se muestran a lo largo de esta sección.

Primero graficamos los datos como serie de tiempo y después mostramos algunas gráficas de barras. Vamos a observar el comportamiento de los datos, veremos si tiene alguna tendencia y variación estacional.

En la figura 3.1 se muestra la gráfica de barras con el número total de alumnos que toman clases por semestre. A simple vista notamos que tiene una tendencia creciente y una estacionalidad semestral. Podemos ver también que el número de alumnos de los semestres impares es siempre mayor al de los semestres pares. Éste fenómeno los vimos en la figura 1.1 al hacer el análisis correspondiente a los datos de la materia *Probabilidad I*.

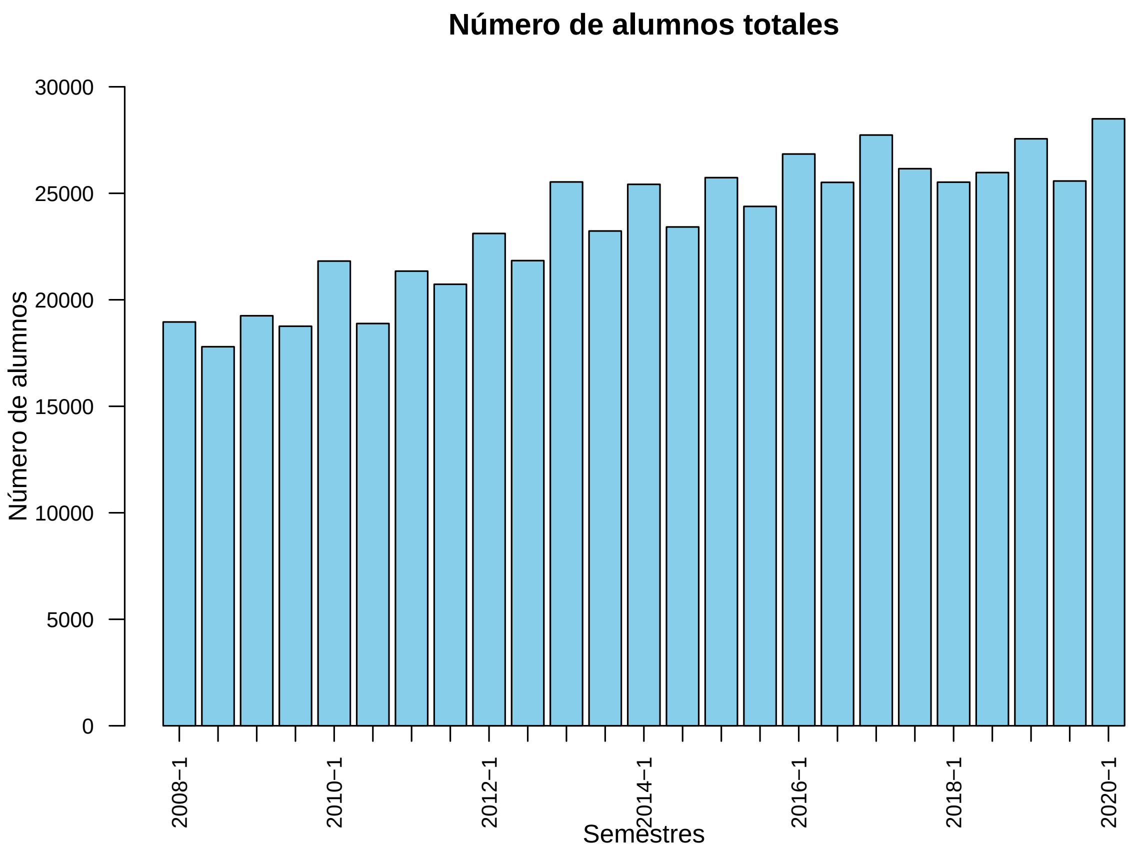


Figura 3.1: *Número total de alumnos por semestre: Vemos la gráfica de barras del número total de alumnos inscritos por cada semestre. Observamos que año con año el número aumenta. Las barras de los semestres impares son mayores que las barras de los semestres pares.*

En la figura 3.2 se muestra la gráfica de la media del número de alumnos que toman clases por semestre de todas las materias. Observamos que los valores tienen una tendencia creciente, esto nos indica que cada semestre, en promedio hay más alumnos tomando clases en la FC.

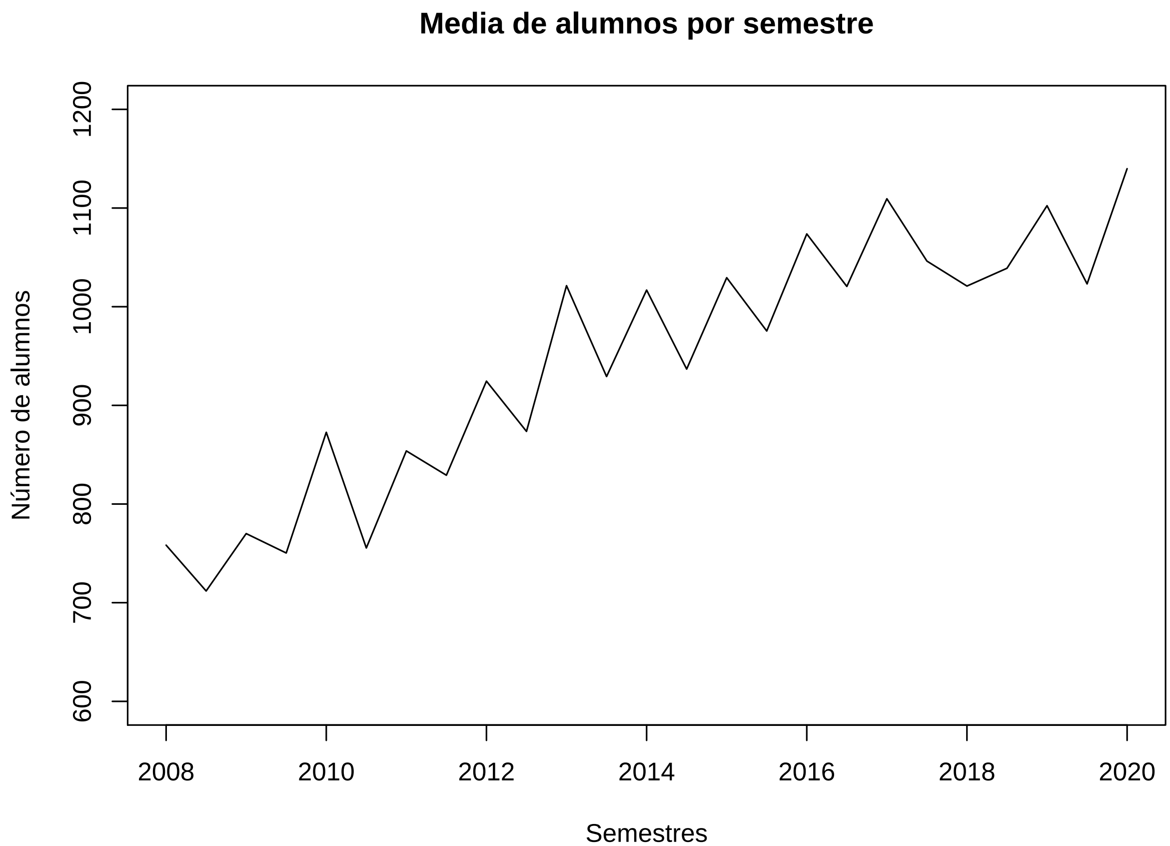


Figura 3.2: *Media de alumnos por semestre: Se observa una tendencia creciente en el promedio de alumnos por semestre. Se tiene la infomraición del semestre 2008-1 al 2020-1.*

En la figura 3.3 se muestra la gráfica de la desviación estándar del número de alumnos por grupo y por semestre de todas las materias. Observamos que los valores se mantienen constantes a lo largo del tiempo. Su rango se encuentra entre 24 y 29.

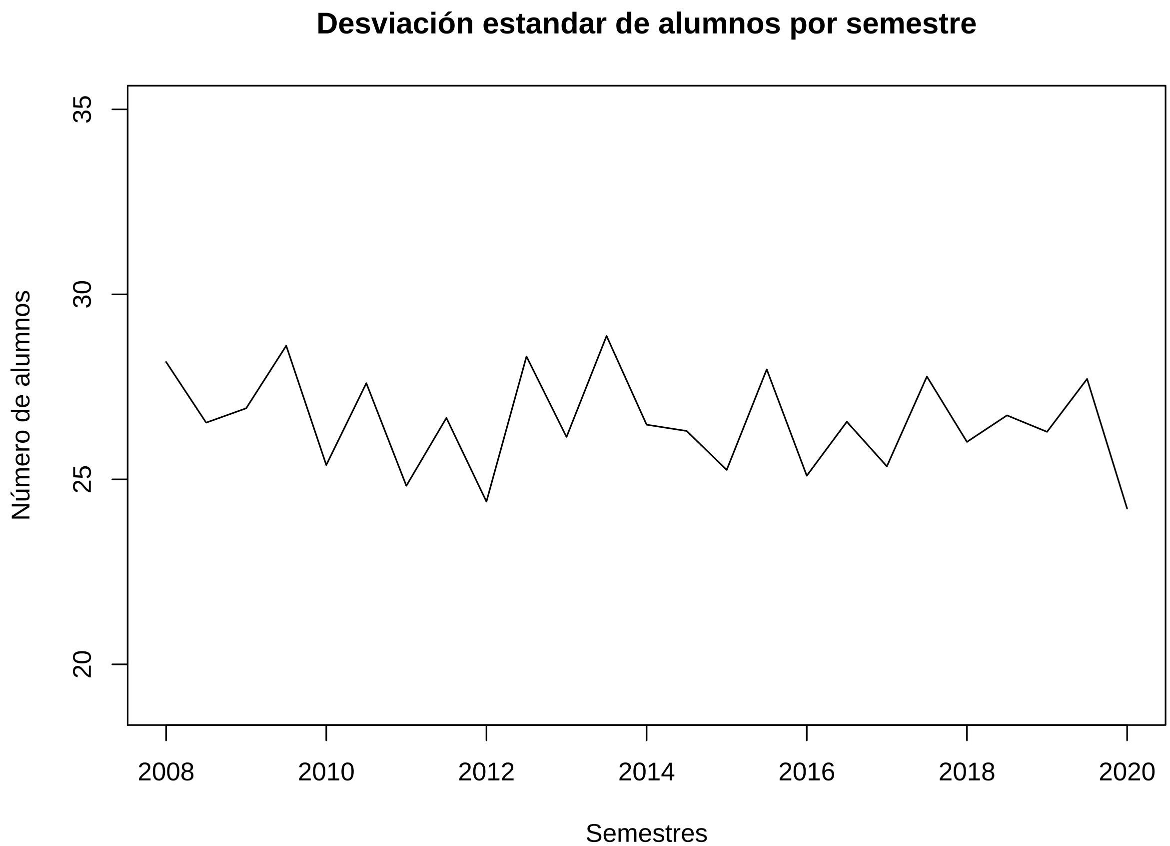


Figura 3.3: *Desviación estándar del número de alumnos por semestre: Vemos que el comportamiento de los datos es constante a lo largo del tiempo.*

En la figura 3.4 se observan 4 diferentes gráficas, en la primera, de arriba hacia abajo, se observan los datos reales del número total de alumnos para cada semestre. En la segunda se muestra la tendecia de los datos, la cual notamos que es creciente. En la tercera vemos la componente estacional que nos indica que los datos tienen una estacionalidad semestral. En la cuarta se ve la componente aleatoria de los datos la cual ya no tiene estacionalidad ni tendencia.

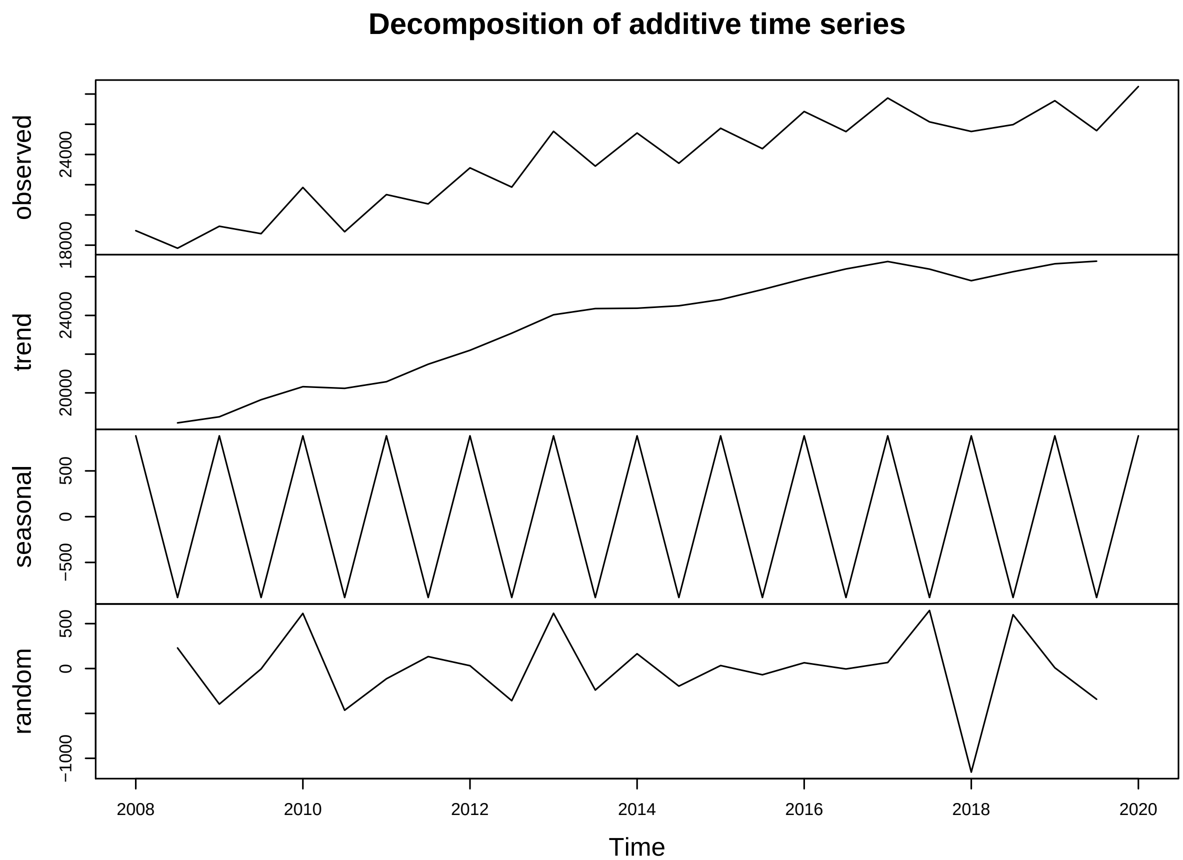


Figura 3.4: *Descomposición por el método aditivo de Holt-Winters: Los datos considerados en esta descomposición es el número total de alumnos por semestre.*

Viendo la figura 3.4 notamos que los datos tienen estacionalidad semestral y una tendencia creciente, por lo tanto confirmamos que se puede utilizar el método Holt-Winters ya que se cumplen los supuestos que se requieren. Para verificar que el modelo de estacionalidad adecuado es el aditivo, vemos la figura 3.3 y notamos que la desviación estándar permanece constante a lo largo del tiempo.

Con estas observaciones concluímos que el método Holt-Winters aditivo es el método adecuado para poder hacer predicciones con nuestros datos.

3.2. Análisis estadístico por grupo de datos

En la figura 3.5 vemos la gráfica del número de alumnos separado por semestres pares e impares. Se observa un comportamiento similar al de la figura 1.1, de la sección 1.6. Vemos con mayor claridad lo que ocurre en la figura 3.1, los datos efectivamente tienen una tendencia creciente. Notamos que el número de alumnos de los semestres impares es mayor al número total de alumnos de los semestres pares en todos los semestres, salvo en el semestre 2018-1 en donde el número de alumnos es menor a los de los semestres adyacentes, los cuales son el 2017-2 y el 2018-2.

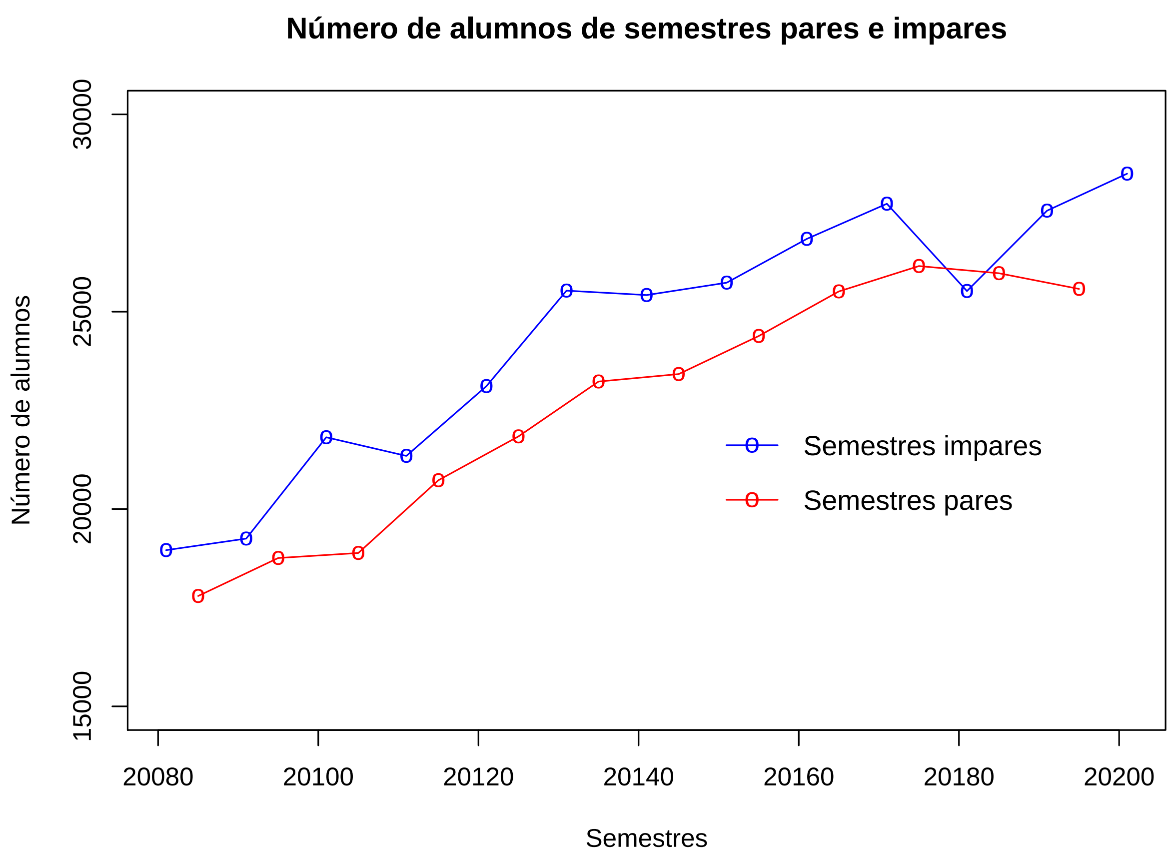


Figura 3.5: *Número de alumnos de semestres pares e impares: Observamos una tendencia creciente y en general el número de alumnos se semestres impares (línea azul) es mayor al número de alumnos de semestres pares (línea roja).*

En la figura 3.6 observamos el histograma doble del número total de alumnos de semestres pares e impares con sus respectivas densidades ajustadas. Notamos que hay una ligera diferencia entre el número de alumnos de los semestres pares con respecto al número de alumnos de los semestres impares. Existe una mayor cantidad de grupos en los semestres pares con un menor número de alumnos, que en los semestres impares. Hay una mayor cantidad de grupos en los semestres impares contra los semestres pares, que tienen entre 35 y 100 alumnos. Tanto para los semestres pares como para los impares, el comportamiento de las distribuciones ajustadas es muy parecido.

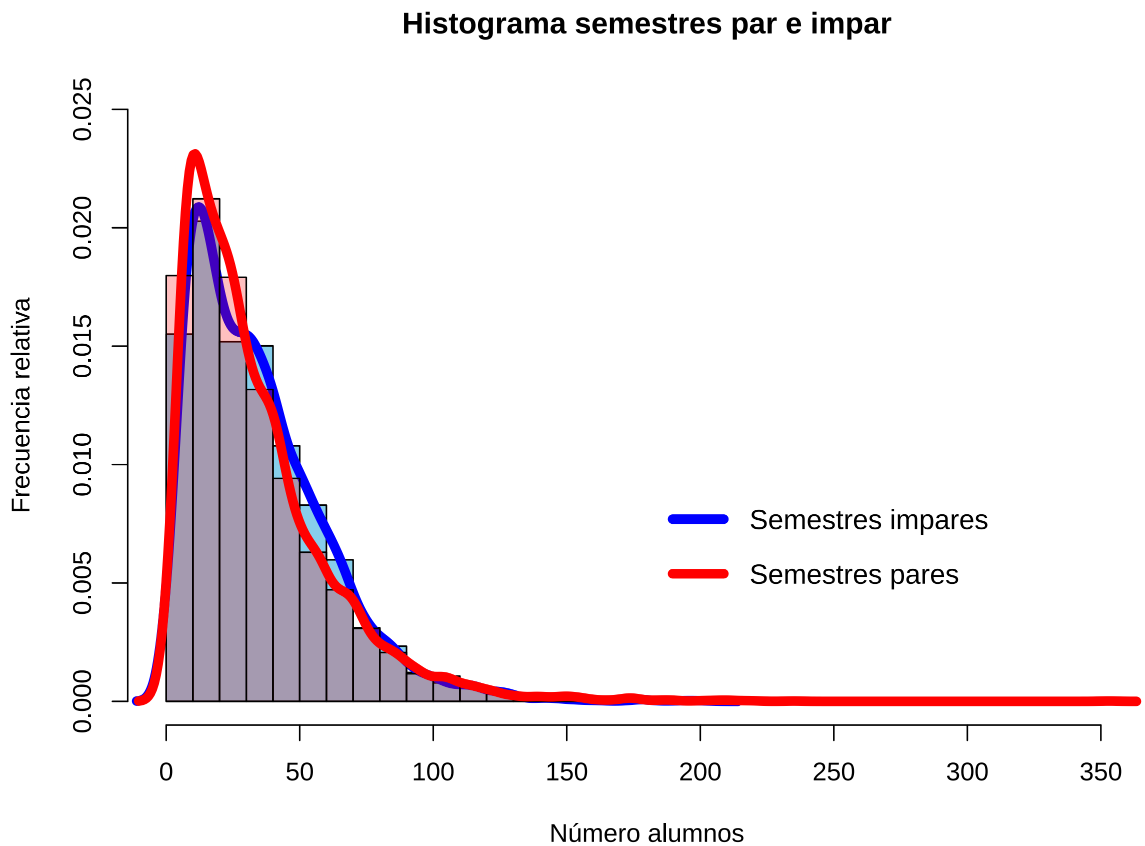


Figura 3.6: *Histograma del número de alumnos de semestres pares e impares: Se observa que las distribuciones ajustadas son muy parecidas, sin importar si los datos son de semestres pares o impares.*

En la figura 3.7 mostramos la gráfica del número de alumnos por turno: matutino y vespertino. Se puede observar que en todo momento el número de alumnos del turno matutino es mayor al número de alumnos del turno vespertino. Este comportamiento lo vimos en la figura 1.3 al hacer el análisis para la materia *Probabilidad I* en la sección 1.6.

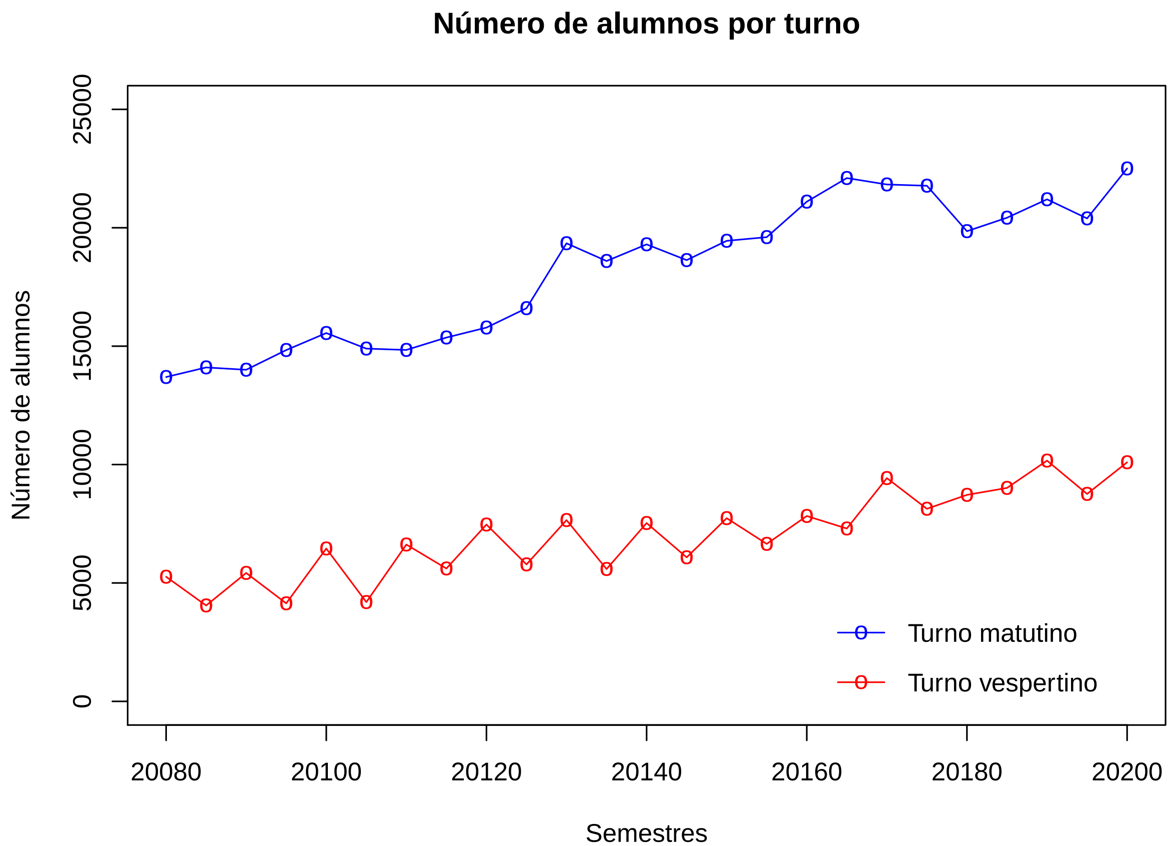


Figura 3.7: *Número de alumnos por turno de todos los semestres: Observamos que la línea azul (turno matutino) está en todo momento por encima de la línea roja (turno vespertino).*

Los datos que se graficaron en el histograma de la figura 3.8 son los alumnos totales por hora de cada semestre. En dicha figura se muestra el histograma doble de los datos divididos en los turnos matutino y vespertino. Notamos que las densidades ajustadas de cada turno son completamente diferentes. Al ver la gráfica podemos concluir que hay más alumnos en el turno matutino que en el vespertino.

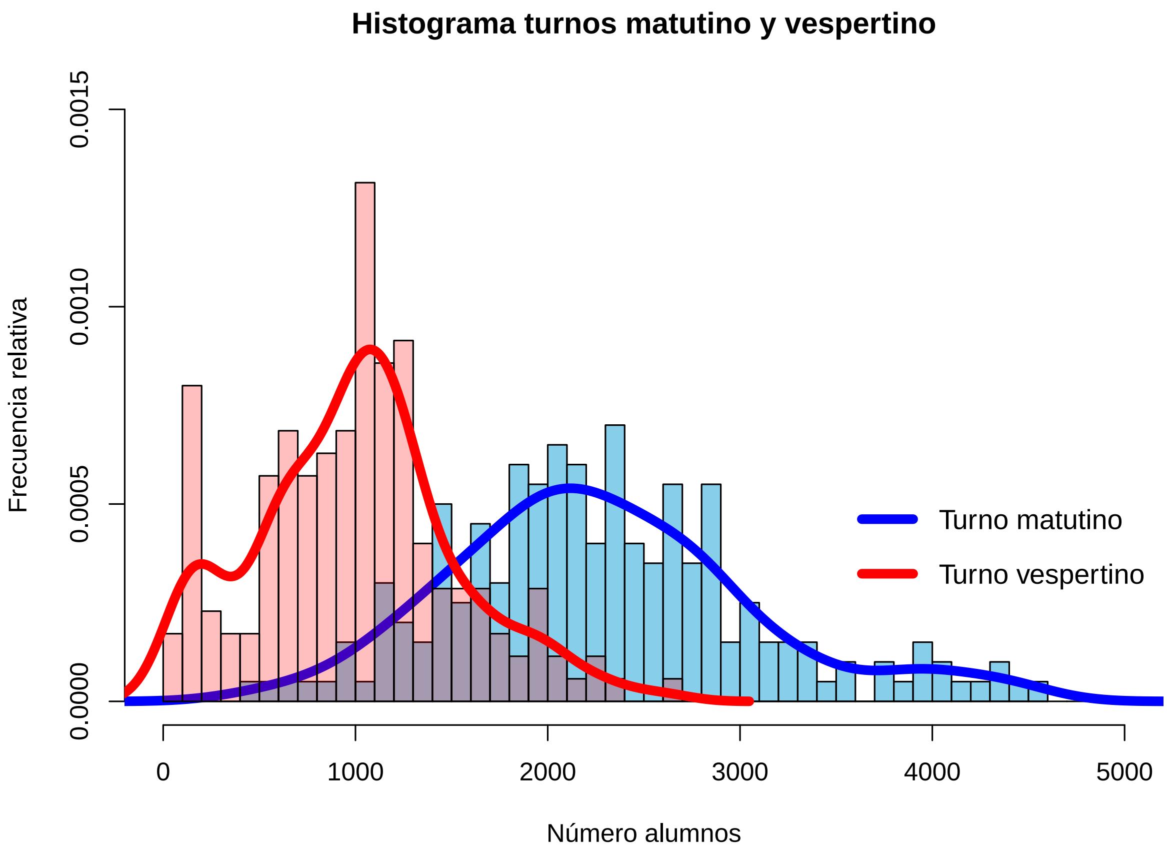


Figura 3.8: *Histograma del número de alumnos de los turnos matutino y vespertino: Podemos concluir qeu hay más alumnos en el turno matutino que en el vespertino. Sus densidades ajustadas son diferentes.*

3.3. Análisis estadístico por carrera

En la figura 3.9 vemos un histograma para cada carrera con el número total de alumnos por grupos. Vemos que los datos tienen un comportamiento similar. Esto se puede explicar por el hecho de que todas las carreras tienen básicamente el mismo tronco común. Es decir, comparten muchas de las materias impartidas en los primeros 4 semestres. Cabe mencionar que el número máximo de alumnos por grupo para la CdC es 211 y para el resto de las carreras es 353.

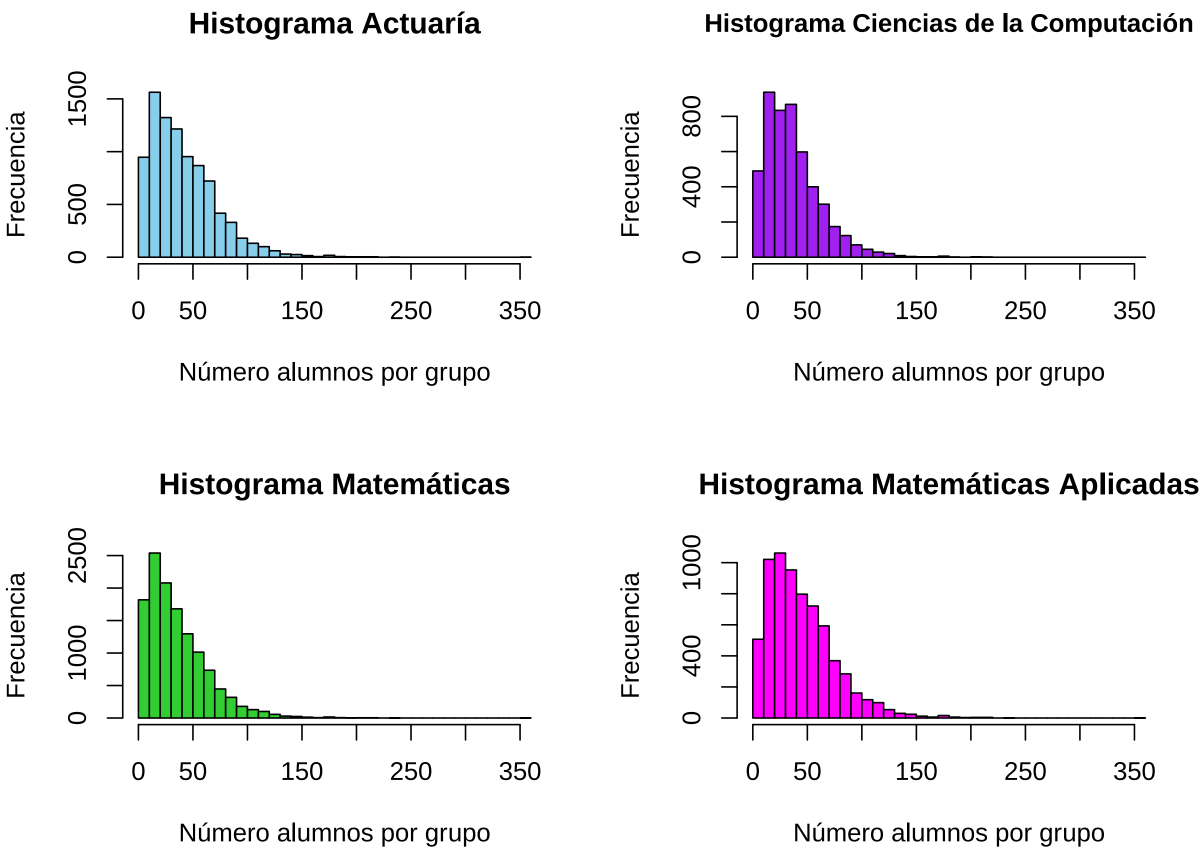


Figura 3.9: *Histograma del número de alumnos por carrera: Se observa un comportamiento similar en todas las carreras.*

En la figura 3.10 vemos una gráfica con las densidades ajustadas a los datos del número de alumnos por grupos para cada carrera. Con esta gráfica podemos confirmar lo que vimos en la figura 3.9, que el comportamiento es similar para todas las carreras.

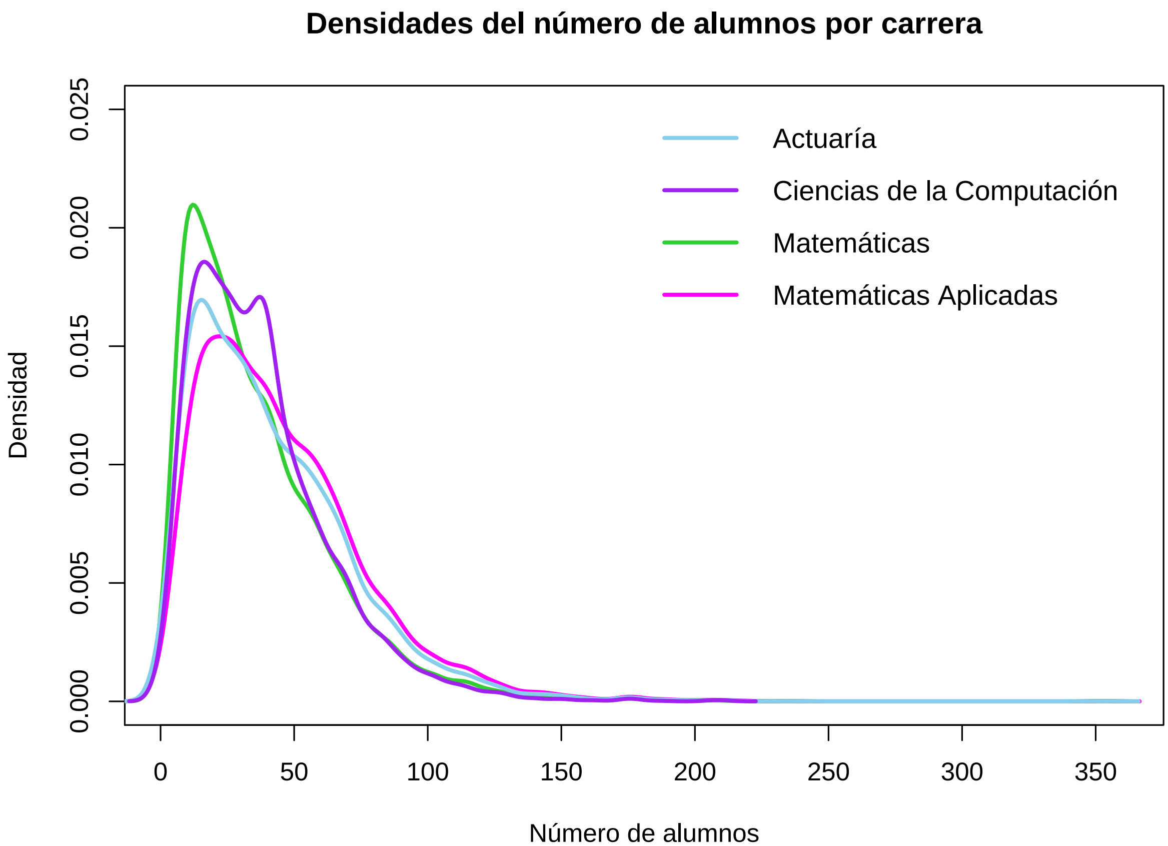


Figura 3.10: *Densidades del número de alumnos por carrera: Vemos que las densidades ajustadas, para cada carrera, son muy similares.*

3.4. Distribución del tamaño de los grupos

En la figura 3.11 se muestra el histograma del número de alumnos por grupo de todos los semestres, desde el 2008-1 hasta el 2020-1. Observamos el mismo comportamiento que en las figuras 3.6, 3.9 y 3.10. La mayor frecuencia se encuentra en los grupos que tienen entre 10 y 20 alumnos.

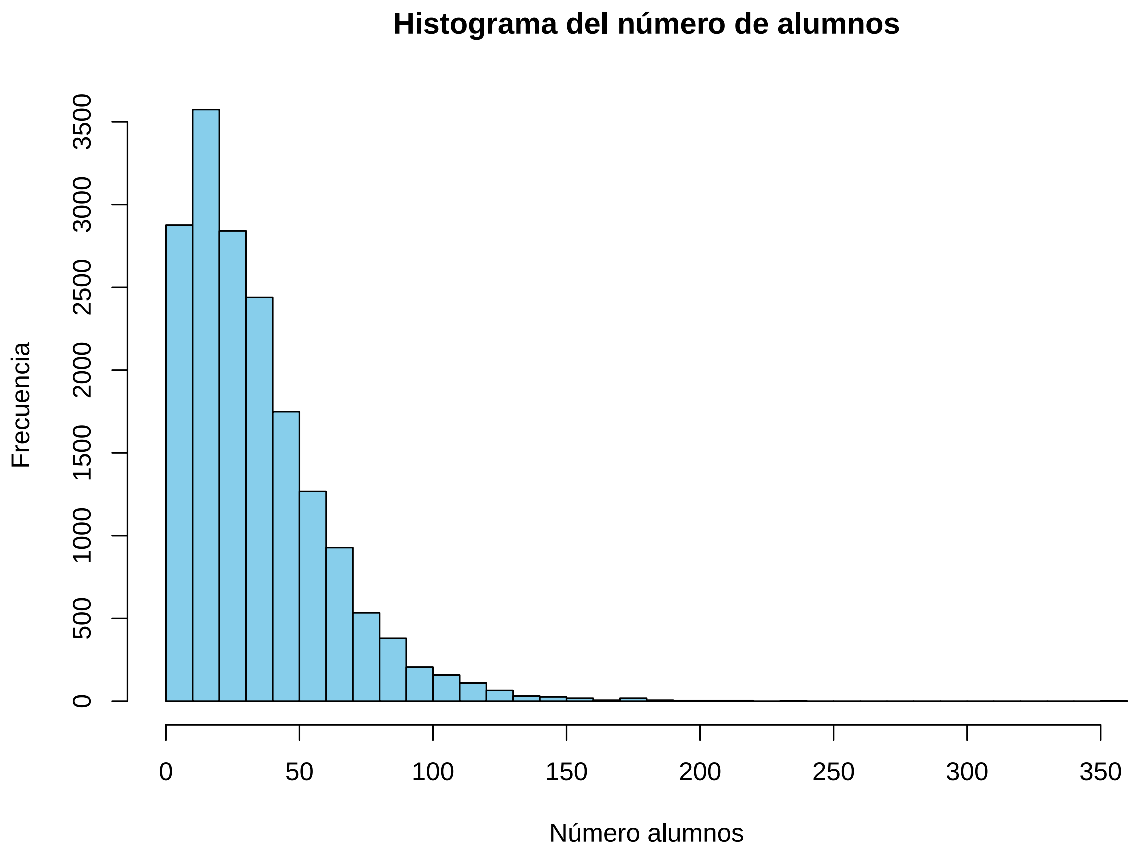


Figura 3.11: *Histograma del número de alumnos por grupo de todos los semestres: La información es de los semestres del 2008-1 al 2020-1. Vemos una mayor concentración en los grupos que tienen entre 10 y 40 alumnos.*

En la figura 3.12 vemos diferentes líneas con las densidades ajustadas a los valores del número de alumnos por grupo de cada semestre. Cada línea corresponde a un semestre. Se tomaron los datos de 25 semestres, del 2008-1 al 2020-1. Notamos que el comportamiento es prácticamente el mismo en todos los semestres, salvo en dos cuyo pico máximo es mucho menor que los demás.

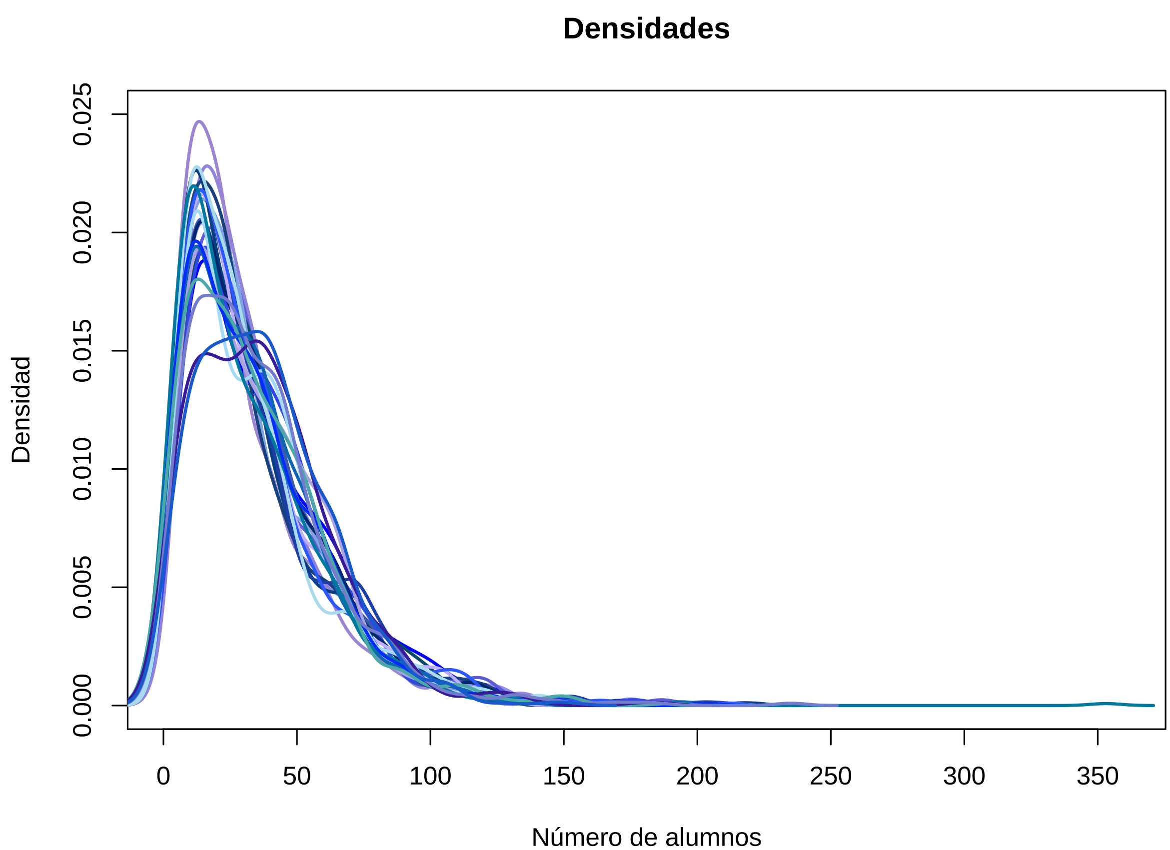


Figura 3.12: *Densidades del número de alumnos por grupo de cada semestre: Cada línea corresponde a la densidad ajustada de los datos de un semestre entre el 2008-1 y el 2020-1*

Viendo las gráficas de las figuras 3.11 y 3.12, podríamos concluir que la distribución que mejor se ajusta al tamaño de los grupos es la distribución Poisson por la forma en la que están distribuidos los datos. Para probar esta hipótesis utilizamos la función `ks.test(X,Y)`, de *R*, para hacer la prueba de Kolmogorov-Smirnov.

La prueba de Kolmogorov-Smirnov, dice que se rechaza H_0 cuando $D_n > D_n^{1-\alpha}$. Donde $D_n^{1-\alpha}$ nos indica el valor en donde inicia la región de rechazo para un nivel de significancia de α y n es el número de datos de la muestra. Tomamos como hipótesis nula $H_0 : X$ y Y tienen la misma distribución.

Definimos a X como el vector con el número de alumnos por cada grupo del semestre 2008-1 al 2020-1. Definimos a Y como un vector de números aleatorios de una distribución $Poisson(\lambda)$. Por el resultado B.1 sabemos que el estimador máximo verosímil de λ para la distribución $Poisson(\lambda)$ es la media de los datos. Con este estimador ($\hat{\lambda} = 34.18746$), obtuvimos los números aleatorios de Y. Tenemos que $Y \sim Poisson(34.18)$.

Por [4] sabemos que:

$$D_n^{1-\alpha} = \sqrt{\frac{ln\left(\frac{1}{\alpha}\right)}{2n}} - 1.6693n^{-1} - 0.20562n^{-\frac{3}{2}} \tag{3.5}$$

En nuestro caso los valores de las variables son: $n = 17,246$ y $\alpha = 0.01$. Sustituyendo en la ecuación 3.5 tenemos que $D_{17246}^{0.99} = 0.01145794823$. Con la función `ks.test(X,Y)`, de *R*, obtenemos el valor de $D_{17246} = 0.39974$

Como $D_{17246} = 0.39974 > 0.01145 = D_{17246}^{0.99} \Rightarrow$ se rechaza H_0, por lo tanto los datos no siguen una distribución Poisson con $\lambda = 34.18$.

Hicimos otra prueba suponiendo que los datos tienen una distribución $Normal(\mu,\sigma)$. Para simular los datos de Y utilizamos los estimadores máximo verosímiles de μ y σ. Estos estimadores los obtuvimos con la función `fitdistr(X, densfun="normal")`, en *R*. Los valores de los estimadores son $\hat{\mu} = 34.1874638$ y $\hat{\sigma} = 26.5768345$. El resultado de la función de la prueba de Kolmogorov-Smirnov es $D_{17246} = 0.10513$.

Como $D_{17246} = 0.10513 > 0.01145 = D_{17246}^{0.99} \Rightarrow$ se rechaza H_0, por lo tanto los datos no siguen una distribución $Normal(34.18, 26.57)$.

En la figura 3.13 vemos el histograma con las frecuencias relativas de los datos del número de alumnos por grupo para cada semestre. La línea azul es la densidad ajustada generada por *R*, la línea morada es la densidad de n números aleatorios con distribución $Poisson(34.18)$ y la línea roja es la densidad de n números aleatorios con distribución $Normal(34.18, 26.57)$.

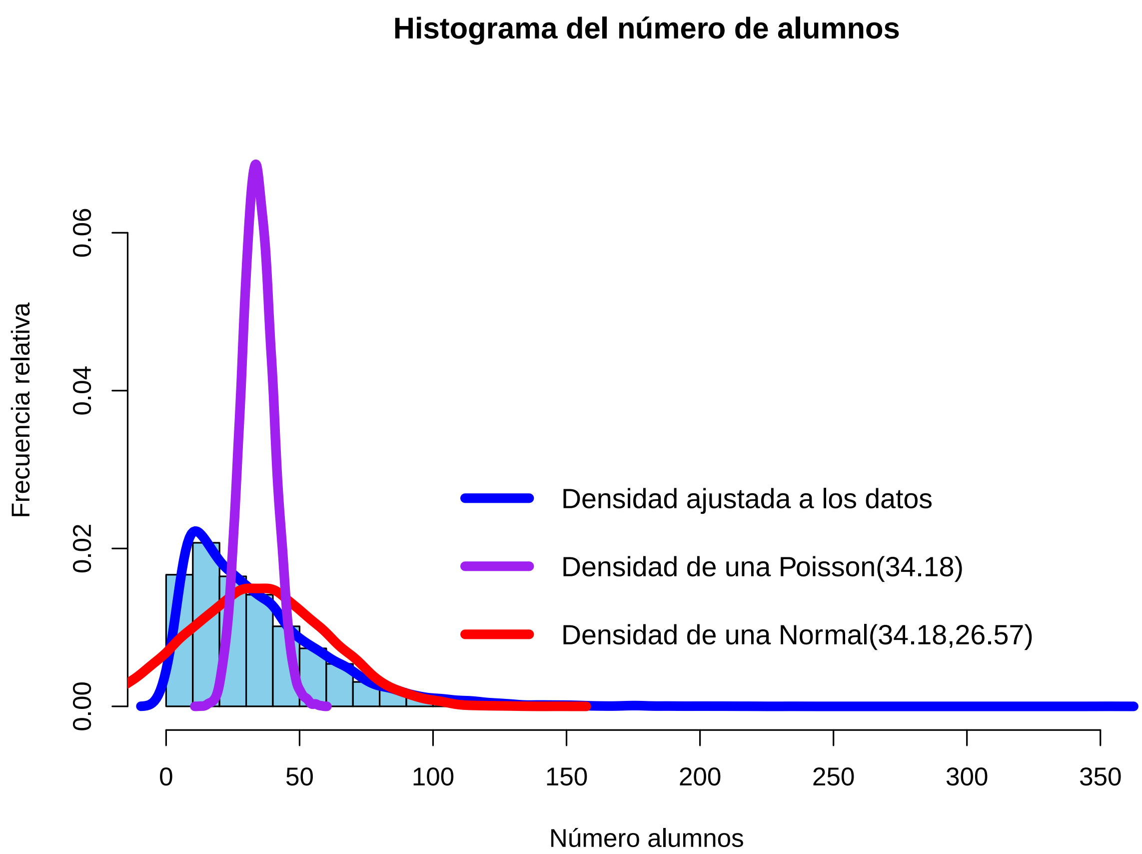


Figura 3.13: *Histograma con densidades ajustadas: Se muestran 3 densidades ajustadas. La línea azul es la ajustada por R. La morada corresponde a una Poisson(34.18). La roja corresponde a una Normal(34.18,26.57). Ninguna de las distribuciones propuestas se ajustan de manera adecuada a los datos.*

Hicimos más pruebas con otras distribuciones y en todos los casos rechazamos la hipótesis nula. Con estos resultados concluímos que el ajuste que se pudiera hacer a los datos tendría que ser en dos partes. Un ejemplo de una posible partición de los datos es que se puede ajustar una distribución para los datos que están entre 0 y 100, y otra distribución para los datos mayores a 100. Esto debido a que a pesar de ser pocos los datos mayores a 100, si impactan en la distribución total. El análisis con esta propuesta no lo realizamos para este trabajo.

3.5. Comportamientos por hora

En esta sección veremos algunas gráficas cuyo eje x corresponde a las horas en las que se imparten las clases. Se empieza por la clase de 7-8hrs y se termina con la clase de 21-22hrs. Primero mostraremos el comportamiento del promedio de grupos por hora y después el comportamiento del promedio del número de alumnos por hora.

En la figura 3.14 vemos la gráfica de barras con el número promedio de grupos por hora. Tomamos la información de 25 semestres. Observamos una disminución considerable de los grupos a las 15hrs. Podemos concluir que es debido a que a esa hora, usualmente la gente sale a comer. A las 21hrs se tiene el menor número de grupos, esto se puede explicar por el hecho de que es la última clase impartida en la FC.

Hay un descenso leve a las 9hrs donde se pudiera suponer que la gente sale a desayunar. En mi caso particular, al estudiar la carrera de Actuaría, tenía clases desde las 7a.m. por lo que las 9a.m. era una buena hora para hacer una pequeña pausa en el horario y tener un descanso para desayunar. Esta hora coincide con el cambio en el que se dejan de impartir las materias exclusivas para los actuarios como Teoría del Seguro, MASP o MASD. Es decir, a partir de las 9 de la mañana se imparten materias de todas las carreras.

A las 10 de la mañana se tiene el número máximo de grupos. Con esta información se podría medir la capacidad que debería de tener la facultad en cuanto al número de salones necesarios para cubrir la demanda de grupos. Si se está preparado para cubrir la demanda del pico más alto de todas las horas, entonces los demás casos están cubiertos por tener menor número de grupos.

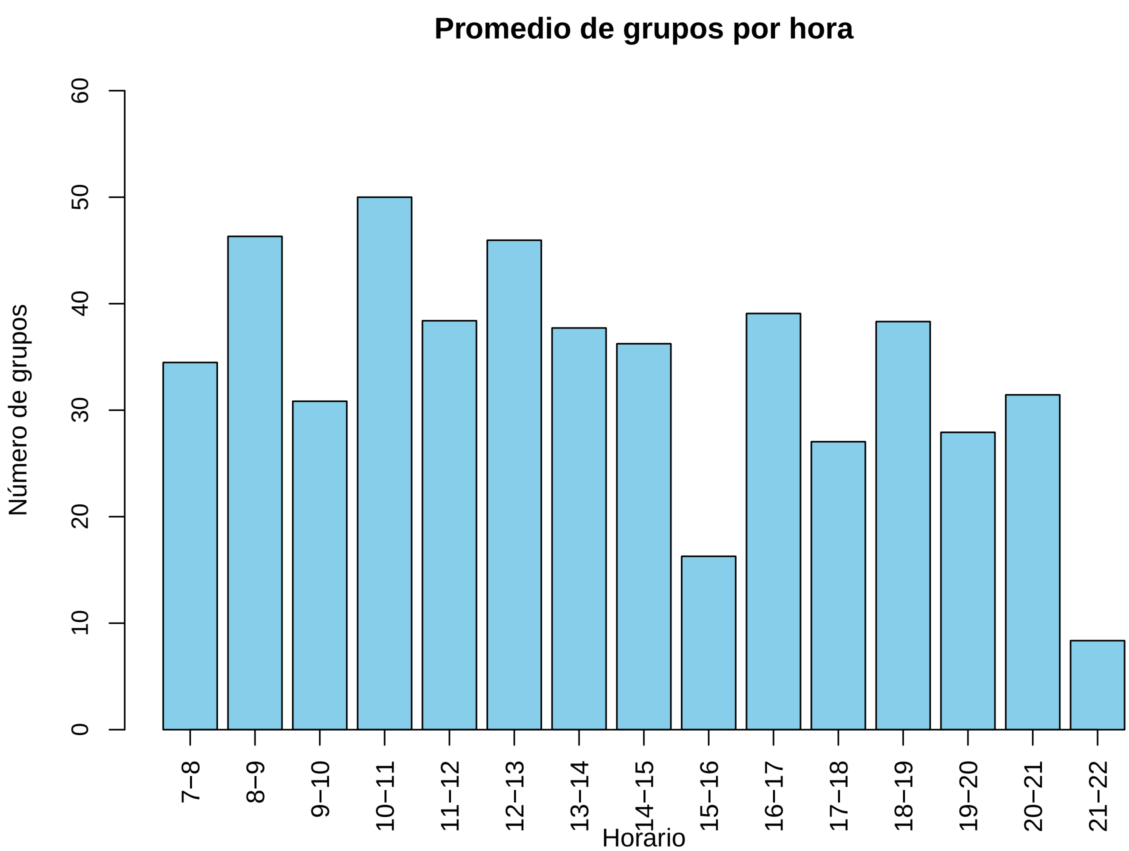


Figura 3.14: *Número promedio de grupos por hora: Se observa una disminución considerable a las 15hrs y a las 21hrs. El valor más alto se encuentra a las 10hrs.*

En la figura 3.15 se muestra la gráfica de barras con el promedio del número de alumnos por hora. Notamos que el comportamiento de ésta gráfica es muy similar al de la gráfica mostrada en la figura 3.14. El pico más alto de los datos también se tiene a las 10 de la mañana y el menor número de alumnos se encuentra a las 21hrs. También hay una disminución considerable a las 15hrs.

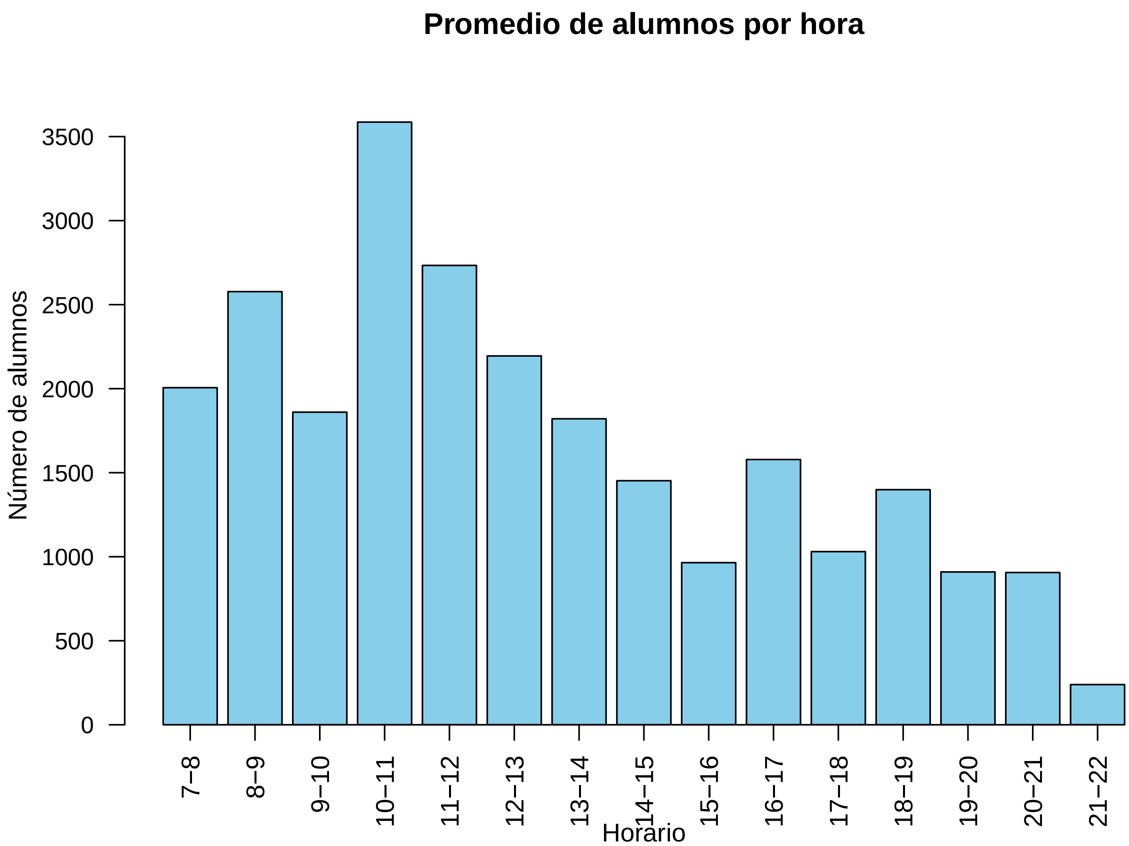


Figura 3.15: *Número promedio de alumnos por hora: Notamos una disminución de los valores a las 9hrs, a las 15hrs y a las 21hrs. El valor más alto lo encontramos a las 10hrs.*

Viendo las gráficas 3.14 y 3.15 podemos concluir que existe una correlación entre el número promedio de grupos por hora y el número promedio de alumnos por hora. Por ejemplo, si no hay alumnos que tomen clases a las 15hrs entonces no tiene caso que se abran grupos a esa hora. Análogamente para las 21hrs. Por el contrario entre más alumnos haya por hora, se deben abrir más grupos a esas horas, como es el caso de las 10hrs.

Capítulo 4

Simulación

La simulación es un proceso que nos permite estudiar el comportamiento de un sistema complejo que es muy difícil de examinar de manera analítica. Nos ayuda a determinar de manera empírica las probabilidades de ciertos eventos. La simulación nos permite experimentar con diversos supuestos que podrían ser muy costosos de realizar. Las áreas en las que se utiliza la simulación como herramienta diaria son por ejemplo en biología, estadística, medicina, química, matemáticas, investigación de operaciones, física, ingeniería o en las ciencias sociales.

Algunos ejemplos de su aplicación van desde simular el lanzamiento de una moneda justa, hasta la simulación de colisiones de átomos en un acelerador de partículas. Se utiliza para todo tipo de propósitos, por ejemplo para poder realizar predicciones en base a datos históricos o enseñar a los pilotos a volar un avión sin poner en riesgo a la población al volar el avión real.

Actualmente se combinan diferentes metodologías de simulación con el software disponible, el análisis de sensibilidad y la optimización estocástica para poder obtener un mejor resultado al momento de simular sistemas que se hacen cada vez más complejos como las redes neuronales.

A lo largo de este capítulo explicaremos el proceso que seguimos para obtener la asignación de los horarios para cada materia con su respectivo profesor.

4.1. Funciones hechas en R

El diagrama de flujo de la función **gen_asignacion** se puede ver en la figura 4.1.

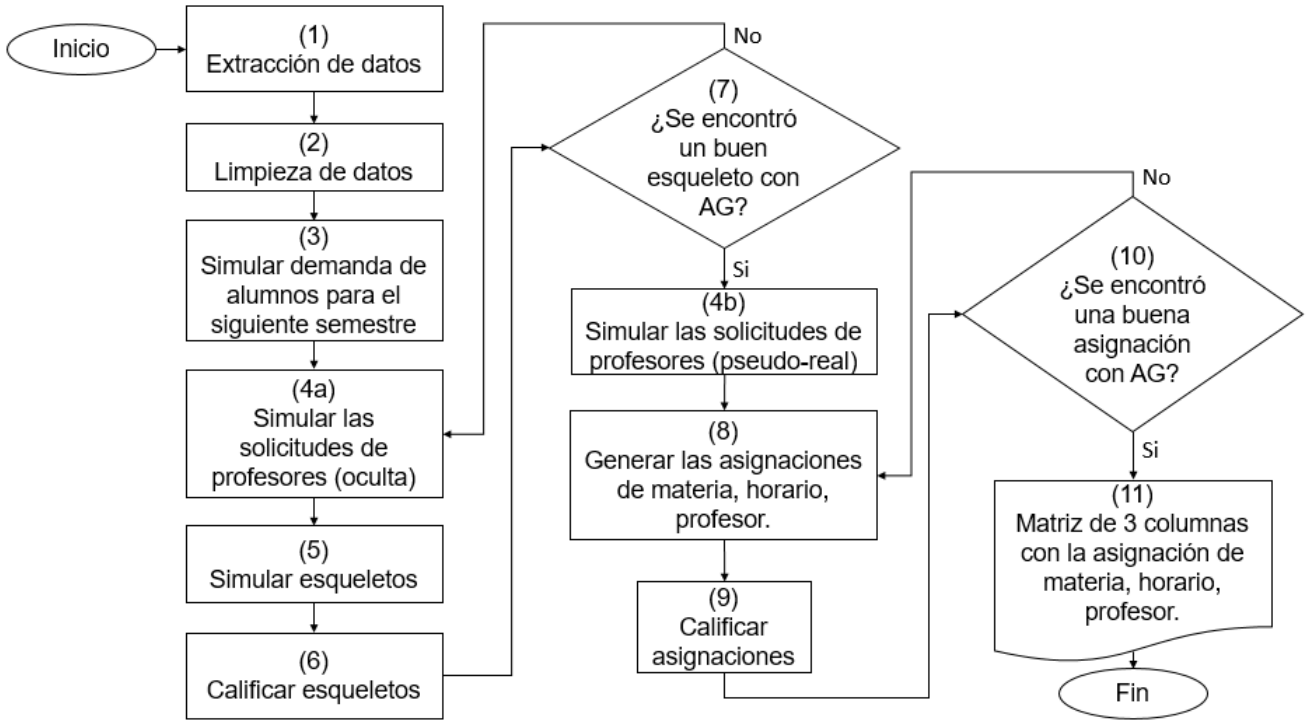


Figura 4.1: *Diagrama de flujo de la función* ***gen_asignacion***

- *posibles_url*
- *gen_m_grande*
- *gen_m_grande_total*
- *gen_esqueleto:* Función que genera esqueletos la cual carga la función *simula_grupos* y ésta a su vez carga la función *estima_grupos*.
- *gen_solicitudes*
- *gen_asignacion*
- *gen_simula_alumnos*
- *gen_simula_tamano_grupo*

4.2. Obtención de los parámetros q_1 y q_2

En esta sección vamos a explicar cómo obtuvimos los valores de q_1 y q_2. Son parámetros que se introducen en la función `hw()` de *R*. Representan los cuantiles utilizados al calcular los intervalos de confianza. Por ejemplo si $q_1 = 80$ entonces se calcula el intervalo al 80 % de confianza. Si se introducen a la función los dos parámetros entonces se calculan dos intervalos, uno al q_1 % de confianza y el otro al q_2 % de confianza.

Primero seleccionamos los parámetros generales necesarios para las simulaciones:

1. Fijamos la semilla con `set.seed(8654)`.

2. Elegimos 3 semestres para simular la demanda del número de alumnos. Los seleccionamos de los semestres que ya teníamos guardados con información real. Hicimos una comparación entre nuestros datos simulados y los reales de cada semestre. Los semestres que elegimos fueron: 2019-1, 2019-2 y 2020-1.

3. Fijamos $k = 5$ (número de semestres que se tienen como ventana de información).

4. Fijamos $num\_sim = 10$ (número de simulaciones de la demanda de alumnos para el semestre a simular).

Después fijamos 5 materias que consideramos representativas para hacer las pruebas iniciales: *Cálculo Diferencial e Integral I*, *Demografía I*, *Modelos no Paramétricos y de Regresión*, *Administración de Riesgos Financieros* y *Seminario de Investigación de Operaciones*.

Tomamos 12 posibles combinaciones de valores para q_1 y q_2, las cuales podemos ver en la tabla 4.1. La letra L indica que se tomó la cota inferior de q_1 y la letra U indica que se tomó la cota superior de q_2. Con estas cotas formamos intervalos de los cuales obtuvimos las simulaciones para los 3 diferentes semestres previamente elegidos.

$q_1 \backslash q_2$	80	85	90	99
80	-	L80,U85	L80,U90	L80,U99
85	L85,U80	-	L85,U90	L85,U99
90	L90,U80	L90,U85	-	L90,U99
99	L99,U80	L99,U85	L99,U90	-

Tabla 4.1: *Posibles valores para q_1 y q_2*

Una vez hecha la simulación obtuvimos una tabla con 7 columnas: materia, intervalo, mín, media, máx, sd y seg. Donde en el renglón i se tienen los datos de la matriz de diferencias relativas de la i-ésima materia para cada intervalo de q_1 y q_2. En la figura 4.2 vemos los primeros 10 renglones de la tabla obtenida.

Materia	Intervalo	Min	Media	Max	sd	Seg
Cálculo Diferencial e Integral I	L80,U85	-2.622222	-0.21911543	0.8627586	0.7287619	7.50
Demografía I	L80,U85	-1.985714	-0.09395821	0.8378378	0.4695739	7.53
Modelos no Paramétricos y de Regresión	L80,U85	-6.922222	-0.45848861	1.0000000	1.5742750	6.27
Administración de Riesgos Financieros	L80,U85	-1.816667	-0.03119518	0.6312500	0.3145900	3.08
Seminario de Investigación de Operaciones	L80,U85	-2.300000	-0.05121275	0.9384615	0.4202550	5.50
Cálculo Diferencial e Integral I	L80,U90	-2.588889	-0.25311699	0.7220690	0.7108189	8.34
Demografía I	L80,U90	-3.228571	-0.20226571	0.7270270	0.6654177	7.16
Modelos no Paramétricos y de Regresión	L80,U90	-6.744444	-0.48396359	1.0000000	1.6007042	5.94
Administración de Riesgos Financieros	L80,U90	-2.316667	-0.04418860	0.6375000	0.3924680	3.28
Seminario de Investigación de Operaciones	L80,U90	-2.233333	-0.05595981	0.9461538	0.4210018	5.49

Figura 4.2: *Matriz con medidas de dispersión*

Decidimos elegir q_1 y q_2 en base a la desviación estándar. Con la tabla 4.2 obtuvimos una matriz de dos columnas que contine en su primer columna el intervalo y en la segunda el promedio de la desviación estándar para cada intervalo de las 5 materias. Los datos de dicha tabla los podemos ver en la figura 4.3.

Intervalo	Promedio_sd
L85,U80	0.6877112
L90,U80	0.6893502
L80,U85	0.7014912
L90,U85	0.7218125
L80,U90	0.7580821
L85,U90	0.7705116
L99,U90	0.8014339
L90,U99	0.9032661
L99,U80	0.9045421
L99,U85	0.9422762
L85,U99	0.9579213
L80,U99	0.9615854

Figura 4.3: *Promedio de la desviación estándar: 5 materias, 12 pruebas*

Los datos en la tabla 4.3 están ordenados de menor a mayor con respecto al promedio de la desviación estándar. Para la segunda prueba se eligieron los primeros 6 intervalos de dicha tabla. Se eligieron 10 materias: *Álgebra Lineal I*, *Álgebra Superior II*, *Algoritmos Genéticos*, *Análisis Matemático IV*, *Análisis Numérico*, *Teoría de la Medida I*, *Cálculo Diferencial e Integral IV*, *Graficas y Juegos*, *Inglés I* y *Matemáticas Actuariales para Seguro de Daños*. La tabla con el promedio de la desviación estandar de sus datos se puede ver en la figura 4.4

Intervalo	Promedio_sd
L85,U90	0.4694684
L85,U80	0.4805732
L80,U90	0.4893892
L90,U85	0.4992955
L80,U85	0.5030579
L90,U80	0.5167806

Figura 4.4: *Promedio de la desviación estándar: 10 materias, 6 pruebas*

Para la tercera prueba elegimos, de la tabla 4.4 los intervalos que tuvieran un promedio en la desviación estándar menor a 0.5. Se eligieron otras 10 materias: *Estadística III*, *Teoría del*

Seguro, *Programación Entera*, *Investigación de Operaciones*, *Geometría Moderna I*, *Geometría Analítica II*, *Lógica Matemática I*, *Cálculo Diferencial e Integral III*, *Estadística I* y *Bases de Datos*. La tabla con el promedio de la desviación estandar de sus datos se puede ver en la figura 4.5.

Intervalo	Promedio_sd
L90,U85	0.4133900
L80,U90	0.4292204
L85,U80	0.4292348
L85,U90	0.4410803

Figura 4.5: *Promedio de la desviación estándar: 10 materias, 4 pruebas*

Podemos ver que los valores de la tabla 4.5 son muy parecidos. Se hizo una prueba con los mismos intervalos pero con 5 materias que se dan en todos los semestres y además tienen muchos alumnos. La prueba se hizo para ver si habpía alguna diferencia en los datos y se pudiera elegir un sólo intervalo. Las materias que se eligieron para esta prueba fueron: *Geometría Analítica I*, *Cálculo Diferencial e Integral II*, *Finanzas I*, *Probabilidad II* y *Procesos Estocásticos I*. La tabla con el promedio de la desviación estandar de sus datos se puede ver en la figura 4.6.

Intervalo	Promedio_sd
L85,U80	0.5829679
L90,U85	0.6027183
L80,U90	0.6127408
L85,U90	0.6260881

Figura 4.6: *Promedio de la desviación estándar: 5 materias, 4 pruebas*

Analizando la información de las matrices de las figuras 4.5 y 4.6, decidimos elegir los valores de $q_1 = 85$ y $q_2 = 80$. Por lo que el intervalo que buscamos estará formado por la cota inferior del intervalo de confianza al 85 % y por la cota superior del intervalo de confianza al 80 %. Para visualizar de una mejor manera cómo se encuentra el intervalo formado, podemos ver la figura 4.7. De dicho intervalo vamos a obtener los valores para simular la demanda de alumnos.

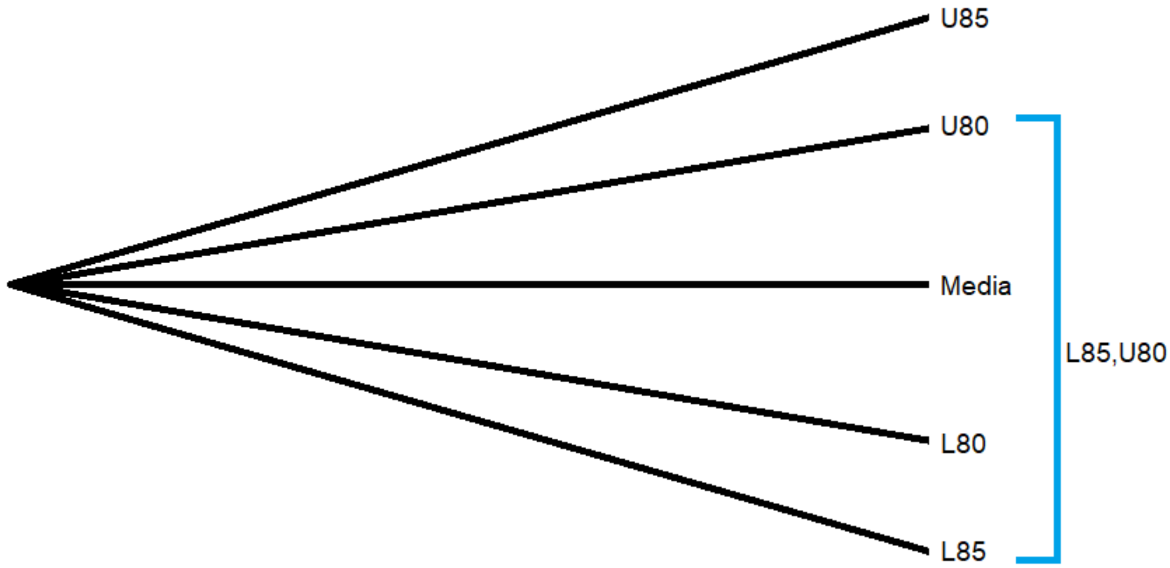


Figura 4.7: *Diagrama de los intervalos de confianza*

Finalmente con los valores de $q_1 = 85$ y $q_2 = 80$ se hizo una prueba aleatoria (eliminando la semilla). Las materias que elegimos para dicha prueba son: *Estadística III*, *Teoría del Seguro*, *Cálculo Diferencial e Integral I*, *Investigación de Operaciones*, *Geometría Moderna I*, *Geometría Analítica II*, *Lógica Matemática I*, *Cálculo Diferencial e Integral III*, *Estadística I*, *Bases de Datos*, *Matemáticas Financieras*, *Cálculo Diferencial e Integral II*, *Probabilidad I*, *Probabilidad II* y *Procesos Estocásticos I*. Los resultados de la prueba aleatoria los podemos ver en la figura 4.8. El promedio de la desviación estándar de todas las materias es 0.48.

Materia	Intervalo	Min	Media	Max	sd	Seg
Estadística III	L85,U80	-2.4053333	-0.062628178	1.0000000	0.7625277	6.68
Teoría del Seguro	L85,U80	-0.7475410	-0.007189294	0.9685714	0.2066836	4.46
Cálculo Diferencial e Integral I	L85,U80	-1.8055556	-0.150509219	0.8544828	0.5743739	7.67
Investigación de Operaciones	L85,U80	-1.5109589	-0.113587042	0.5523364	0.4197767	8.46
Geometría Moderna I	L85,U80	-1.0729730	0.062571864	1.0000000	0.3758780	5.20
Geometría Analítica II	L85,U80	-1.6896552	-0.075029650	1.0000000	0.6112092	7.42
Lógica Matemática I	L85,U80	-1.4857143	0.009737679	1.0000000	0.4441402	6.46
Cálculo Diferencial e Integral III	L85,U80	-1.6142857	-0.118954103	0.8689189	0.5222578	7.26
Estadística I	L85,U80	-1.8022222	-0.045271946	0.9751515	0.5525782	6.57
Bases de Datos	L85,U80	-0.5000000	0.030660747	0.8080000	0.1858422	4.25
Matemáticas Financieras	L85,U80	-0.8403974	0.030471884	0.8584270	0.2544040	4.69
Cálculo Diferencial e Integral II	L85,U80	-3.2192308	-0.053958329	1.0000000	0.6681179	6.77
Probabilidad I	L85,U80	-1.6750000	-0.047856672	0.9188034	0.4289253	5.91
Probabilidad II	L85,U80	-1.9750000	-0.124282718	1.0000000	0.7012144	7.28
Procesos Estocásticos I	L85,U80	-1.8096154	-0.109287138	0.7419643	0.5144173	4.86

Figura 4.8: *Matriz con medidas de dispersión de prueba aleatoria*

4.3. Simulación de la demanda de alumnos

La demanda del número de alumnos para el siguiente semestre la hicimos por materia y por hora. Para poder hacer la simulación lo primero que hicimos fue acomodar la información que teníamos por semestres y por hora. El procedimiento que seguimos es el siguiente:

1. Definir el semestre del cual se quiere obtener la simulación.

2. Definir el número de semestres que se quieren como ventana de información.

3. Tomar una submatriz de *m_grande_total* con la información de una materia para los semestres en la ventana de información.

4. Para cada semestre dentro de la ventana de información se suma el número de alumnos en cada hora.

5. Se obtiene una matriz de $15 \times k$ (k es el número de semestres en la ventana) como la que se puede ver en la figura 4.9.

	20181	20182	20191	20192	20201
7-8	0	0	0	0	0
8-9	0	0	0	0	0
9-10	0	0	71	0	52
10-11	198	0	75	0	144
11-12	0	44	0	9	0
12-13	0	75	0	97	0
13-14	0	0	0	0	0
14-15	0	0	0	0	0
15-16	0	0	0	0	0
16-17	0	0	0	0	0
17-18	0	52	0	40	47
18-19	0	0	0	0	88
19-20	78	0	63	0	0
20-21	0	53	79	69	0
21-22	0	0	0	0	0

Figura 4.9: *Ejemplo de matriz con alumnos corregidos*

Con el procedimiento descrito pudimos generar vectores por hora y aplicar la función `hw()` en *R* para obtener la demanda de alumnos esperados para el siguiente semestre. En la figura 4.10 vemos el vector con la demanda de alumnos simulados para el semestre 2020-2 de la materia *Estadística III*.

Notamos que el valor de la demanda de alumnos es cero cuando en todos los semestres de alguna hora no hay datos. En el ejemplo, es el caso de las 7hrs, 8hrs, 13hrs, 14hrs, 15hrs, 16hrs y 21hrs.

Observando los datos de las 10hrs. vemos que en los semestres pares no hay alumnos, por lo que en la simulación se obtiene únicamente un alumno. Si vemos los datos de las 17hrs vemos que de los 5 semestres en la ventana se tienen alumnos en los semestres pares y en un semestre impar, el número de alumnos simulados para esa hora son 31 alumnos.

Con estos ejemplos podemos ver de manera tangible que el modelo si respeta la estacionalidad semestral que tienen los datos.

	20202
7-8	0
8-9	0
9-10	61
10-11	1
11-12	8
12-13	122
13-14	0
14-15	0
15-16	0
16-17	0
17-18	31
18-19	29
19-20	6
20-21	132
21-22	0

Figura 4.10: *Ejemplo de vector con demanda simulada para el 2020-2*

Obtuvimos vectores con la demanda simulada para cada una de las materias y formamos una matriz de 15×335. En la figura 4.11 podemos ver un ejemplo de cómo se ve la matriz formada.

Analicemos 2 pares de grupos, primero veamos la segunda y la quinta columna, que corresponden a las materias de *Álgebra Superior II* y *Geometría Analítica I*, respectivamente. Ambas son materias obligatorias para Actuaría, Matemáticas y Matemáticas Aplicadas. La primera corresponde a semestres pares y la segunda a semestres pares. Notamos que para *Geometría Analítica I*, se tienen alumnos prácticamente en cada hora, pero el número no es muy grande, a diferencia de los alumnos simulados para *Álgebra Superior II*, en donde hay varias horas con cero alumnos simulados pero hay dos grandes cantidades, una a las 9hrs con 832 alumnos y la otra a las 18hrs con 224 alumnos. Con esta comparación podemos ejemplificar la diferencia entre una materia que corresponde a semestres pares con una de semestres impares.

Ahora analicemos las columnas 4 y 8, correspondientes a las materias de *Seminario de Topología A* y *Probabilidad II*. La primera es una materia optativa para Matemáticas y la segunda es una materia obligatoria para Actuaría, correspondiente a semestres pares y optativa para Ciencias de la Computación, Matemáticas y Matemáticas Aplicadas. El número total de alumnos simulados para *Seminario de Topología A* es menor a 20, en cambio para *Probabilidad II* se tiene una gran cantidad de alumnos a las 8hrs, 9hrs y 10hrs. Considerando los valores que se tienen en el turno vespertino para *Probabilidad II*, notamos que a las 19hrs también hay una gran cantidad de alumnos. Con esta comparación podemos ejemplificar la diferencia entre una materia obligatoria y una optativa, así como la diferencia entre el turno matutino y vespertino.

	Topología I	Álgebra Superior II	Teoría de Gráficas	Seminario de Topología A	Geometría Analítica I	Graficas y Juegos	Geometría Moderna I	Probabilidad II	Análisis Matemático II	Análisis Matemático III	Series de Tiempo
7-8	0	16	0	0	37	0	0	0	0	0	1
8-9	2	59	0	0	39	46	53	264	0	0	0
9-10	0	832	0	0	38	4	55	160	44	0	0
10-11	15	0	0	1	16	7	0	187	1	0	0
11-12	56	0	0	2	0	0	0	5	13	0	0
12-13	0	0	0	1	12	90	6	0	31	9	2
13-14	2	0	0	5	30	113	19	0	3	12	0
14-15	32	8	0	0	107	14	86	0	0	0	0
15-16	20	7	0	0	43	0	1	0	0	0	0
16-17	0	0	0	0	0	0	36	0	41	0	0
17-18	1	0	0	10	27	0	0	8	32	8	3
18-19	44	224	0	0	187	0	0	9	18	0	1
19-20	0	0	0	0	45	8	0	85	0	8	0
20-21	0	9	0	0	10	0	16	0	25	0	2
21-22	0	6	0	0	16	0	0	0	0	0	0

Figura 4.11: *Ejemplo de matriz con demanda simulada para el 2020-2*

4.4. Simulación de tamaño de grupos

En esta sección vamos a explicar cómo hicimos la simulación del tamaño de grupos. Vamos a definir al tamaño de un grupo como el número de alumnos que va a tener cada grupo.

Hicimos una función en *R* que realiza los siguientes pasos:

1. Definir *m_grande_2015* la cual es una submatriz de *m_grande_total* con los datos de los semestres del 2015-1 al 2020-1.
2. Obtener, de *m_grande_2015*, la información del número de alumnos que ha tenido un profesor.
3. Tomar el mínimo (a) y el máximo (b) de esos datos.
4. Generar un número aleatorio con distribución uniforme en ese intervalo con la función `runif(1,min = a, max = b)` en *R*.
5. Redondear el número aleatorio con la función `ceiling()` en *R*.
6. Regresa el número redondeado.

Con este procedimiento simulamos el tamaño de los grupos con respecto a los profesores. En la vida real cuando un alumno decide inscribirse a una materia a cierta hora, la decisión que toma para elegir el grupo al que se quiere inscribir es el profesor con el que le gustaría tomar esa materia a esa hora. Decidimos realizar de esta manera la simulación porque queremos que el número de alumnos de cada grupo dependa de los profesores y no de la distribución general que tiene el tamaño de los grupos (ver 3.4).

4.5. Obtención de información para solicitudes

Antes de iniciar las simulaciones de elección de materia y de horario obtuvimos un vector y una matriz con la información de las materias y de los profesores, respectivamente.

En el caso de las materias, el vector *vec_nom_materias_total* lo obtuvimos a partir de la matriz *m_grande_total* del semestre 2008-1 al 2020-1. No tiene nombres repetidos. Tiene 335 materias.

En el caso de los profesores, la matriz *mat_nom_prof_total* tiene 2 columnas. En la primer columna se tienen los nombres de todos los profesores que han impartido clase desde el semestre 2015-1 hasta el 2020-1. Dichos nombres los obtuvimos de la matriz *m_grande_total* de los semestres correspondientes.

En la segunda columna de la matriz, se tiene un 1 si el profesor es de tiempo completo y un 0 si no. Para llenarla ingresamos a la página `http://www.matematicas.unam.mx/index.php/nosotros/profesores-de-tiempo-completo` del Departamento de Matemáticas. Con la aplicación *SelectorGadget* seleccionamos el vector con el nombre de los profesores de tiempo completo. En la figura 4.12 podemos ver el código CSS que utilizamos para obtener los datos en R. También observamos que se seleccionaron 94 profesores.

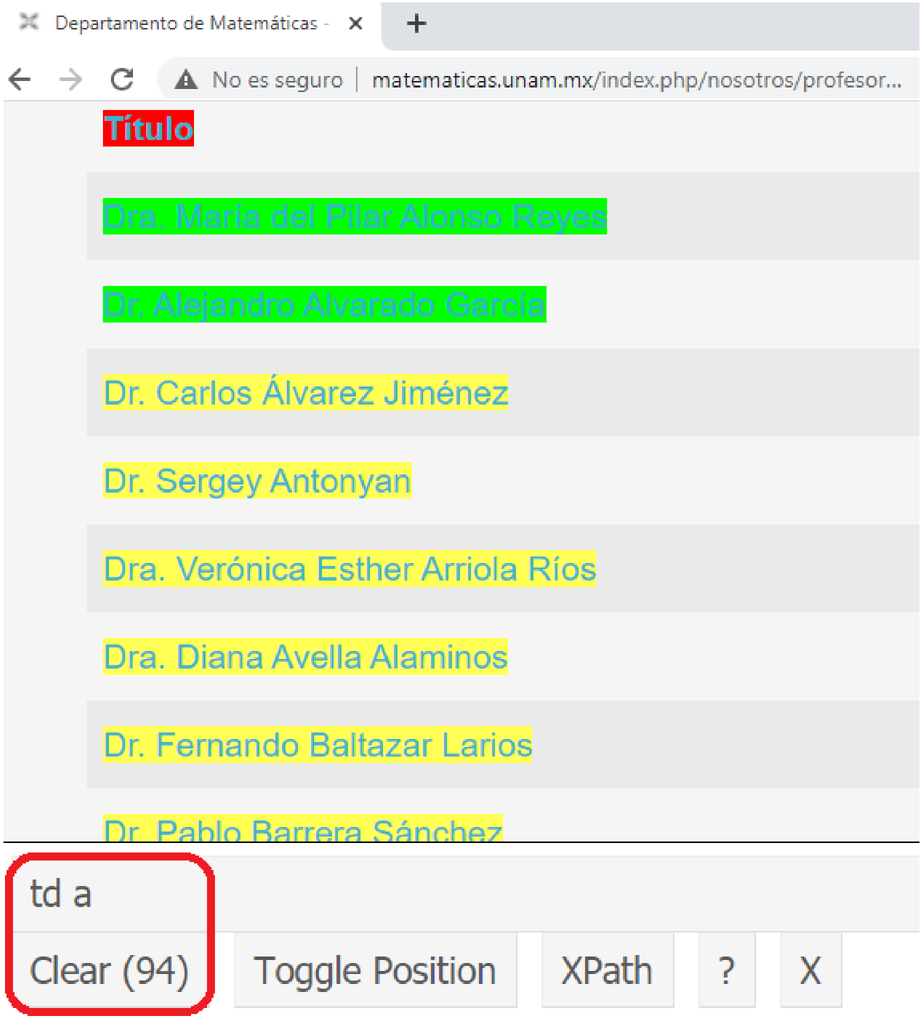


Figura 4.12: *Profesores de tiempo completo: SelectorGadget*

Al extraer la información en R obtuvimos un vector con 94 entradas. En la figura 4.13 podemos ver los primeros 20 valores del vector. Notamos que cada entrada del vector inicia con los caracteres $\backslash n\backslash t\backslash t\backslash t\backslash t\backslash t\backslash t$. Estos caracteres, en la presentación final de la página de internet, indican un salto de línea y las tabulaciones o espacios que se tienen de izquierda a derecha.

```
 [1] "\n\t\t\t\t\t\t\t Dra. María del Pilar Alonso Reyes"
 [2] "\n\t\t\t\t\t\t\tDr. Alejandro Alvarado García"
 [3] "\n\t\t\t\t\t\t\t Dr. Carlos Álvarez Jiménez "
 [4] "\n\t\t\t\t\t\t\t Dr.  Sergey Antonyan"
 [5] "\n\t\t\t\t\t\t\tDra. Verónica Esther Arriola Ríos"
 [6] "\n\t\t\t\t\t\t\tDra. Diana Avella Alaminos"
 [7] "\n\t\t\t\t\t\t\tDr. Fernando Baltazar Larios"
 [8] "\n\t\t\t\t\t\t\tDr. Pablo Barrera Sánchez"
 [9] "\n\t\t\t\t\t\t\tDr. Fernando Brambila Paz"
[10] "\n\t\t\t\t\t\t\tM. en C. Alejandro Bravo Mojica"
[11] "\n\t\t\t\t\t\t\t Dra. Gabriela Campero Arena  "
[12] "\n\t\t\t\t\t\t\t Dr. Humberto Andrés Carrillo Calvet "
[13] "\n\t\t\t\t\t\t\t Dr.  Fidel Casarrubias Segura"
[14] "\n\t\t\t\t\t\t\tMat. Margarita Elvira Chávez Cano"
[15] "\n\t\t\t\t\t\t\tM. en C. Elena de Oteyza de Oteyza"
[16] "\n\t\t\t\t\t\t\tDra. Guillermina Eslava Gómez "
[17] "\n\t\t\t\t\t\t\tDra. María de Lourdes Esteva Peralta "
[18] "\n\t\t\t\t\t\t\tDr. Manuel Jesús Falconi Magaña  "
[19] "\n\t\t\t\t\t\t\tDra. Ma. Asunción Begoña Fernández Fernández"
[20] "\n\t\t\t\t\t\t\tAct.  Javier Fernández García"
```

Figura 4.13: *Vector de profesores de tiempo completo*

Limpiamos los datos para obtener un vector que sólo tuviera los nombres de los profesores. Eliminamos el título de cada uno porque en los horarios publicados en las páginas de la FC sus nombres no tienen título. También eliminamos los espacios finales que había en algunos nombres.

De esta manera obtuvimos el vector con el nombre de los profesores de tiempo completo del Departamento de Matemáticas. Dicho vector lo comparamos con la primer columna de la matriz *mat_nom_prof_total*, cuando los nombres coincidieron, pusimos un 1 en el renglón correspondiente.

Al limpiar los datos encontramos 11 nombres que analizamos a mano porque no aparecía el 1 en su respectivo renglón. Encontramos que no aparecía la información necesaria en la matriz *mat_nom_prof_total* por diferencias en los nombres. Encontramos diferencias por acentos, por mayúsculas y por nombre incompleto. En la tabla 4.2 vemos los nombres que aparecen en las páginas de FC comparados con los que aparecen en la página del Departamento de Matemáticas.

Nombre en páginas de FC	**Nombre en página del Depto. de Matemáticas**
Alejandro Ricardo Garciadiego Dantan	Alejandro Ricardo Garciadiego Dantán
Edith Corina Sáenz Valadez	Edith Corina Sáenz Valadéz
Emilio Esteban Lluis Puebla	Emilio Lluis Puebla
Guillermo Javier Francisco Sienra Loera	Guillermo Sienra Loera
María Asunción Begoña Fernández Fernández	Ma. Asunción Begoña Fernández Fernández
María Concepción Ana Luisa Solís González-Cosío	Ana Luisa Solís González Cosío
María Isabel Puga Espinosa	Isabel Puga Espinosa
María Lourdes Velasco Arreguí	María de Lourdes Velasco Arregui
Mucuy-Kak del Carmen Guevara Aguirre	Mucuy-kak del Carmen Guevara Aguirre
Oscar Alfredo Palmas Velasco	Óscar Alfredo Palmas Velasco
Úrsula Xiomara Iturrarán Viveros	Úrsula Iturrarán Viveros

Tabla 4.2: *Diferencias en nombres de profesores de tiempo completo*

Finalmente en la matriz *mat_nom_prof_total* se tiene la información de 1387 profesores de los cuales 94 son profesores de tiempo completo.

Algunas notas a considerar de esta matriz son:

- Hay profesores que se repiten por diferencia de acentos. Ej. *César Alejandro Arellano Ruíz, Luis Eduardo García Hernández*
- Hay profesores que se repiten por tener a lado el nombre de los ayudantes. Ej. *Fermín Alberto Viniegra Heberlein, Edgar Vázquez Luis*
- Puede haber profesores que ya no impartan clases en la FC.

4.6. Simulación de solicitudes de profesores

En esta sección vamos a explicar cómo hicimos la simulación de la solicitud de los profesores. En la vida real los profesores pueden elegir libremente las materias que quieren impartir y seleccionan las horas a las que desean impartir sus clases. Dado que no contamos con esa información decidimos simular la elección de materias y horarios en base a la información que tenemos de semestres anteriores.

Como vimos en el diagrama 4.1 simulamos dos veces las solicitudes de los profesores, en el proceso de asignación. A la primera vez que simulamos las solicitudes la llamaremos *Solicitud oculta* y a la segunda la llamaremos *Solicitud pseudo-real*. La explicación de su uso lo vemos a continuación.

- Solicitud oculta: La llamamos oculta porque nos ayuda para la generación de los esqueletos. No influye directamente en la asignación final.
- Solicitud pseudo-real: Es la simulación de las posibles elecciones que los profesores harían en la vida real. Nos ayuda directamente a realizar la asignación final.

El procedimiento para ambos casos es el mismo. Al finalizar obtuvimos una matriz, llamada *mat_1_solicitud*. La matriz tiene 5 columnas (Profesor, TC, Materia, Num_Materia, Horario) y 6 renglones. Tiene la información de la solicitud de un profesor. Se eligen 2 materias y hasta 3 diferentes horarios. Los pasos que realizamos para obtener la matriz *mat_1_solicitud*, con la solicitud de un profesor, son los siguientes:

1. Llenar la columna *Profesor* de *mat_1_solicitud* con el nombre del profesor del cual queremos realizar la solicitud.
2. Llenar la columna *TC* de *mat_1_solicitud*. Esta columna tiene unos si el profesor es de tiempo completo y ceros si es un profesor de asignatura.
3. Obtener, de *m_grande_2015*, la información de las materias que ha impartido el profesor elegido. Guardar la información en el vector `materias_profesor`. Se tienen 3 casos con respecto al número de materias impartidas:
    a) El número de materias es 2: Llenar los primeros 3 renglones, de la columna *Materia*, con la información de la materia 1 y los últimos 3 renglones con la información de la segunda materia.

b) El número de materias es mayor o igual a 3: Se toma una muestra de dos materias, con la función `sample(materias_profesor, size = 2)` en *R*. Se llena la columna *Materia* como el caso anterior.

c) El número de materias es 1: Llenar la columna *Materia* con la materia impartida por el profesor elegido.

4. Llenar la columna *Num_Materia* de *mat_1_solicitud* con los números de materia correspondientes a las materias elegidas en el paso anterior.

5. Obtener, de *m_grande_2015*, la información de las horas en las que ha dado clases el profesor elegido. Guardar la información en el vector `horas_profesor`. Se tienen 4 casos con respecto al número de horas:

   a) El número de horas es 3: Llenar los renglones 1 y 4, de la columna *Horario*, con la información de la hora 1; los renglones 2 y 5 con la información de la hora 2 y los renglones 3 y 6 con la información de la hora 3.

   b) El número de horas es mayor o igual a 4: Se toma una muestra de 3 horas, con la función `sample(horas_profesor, size = 3)` en *R*. Se llena la columna *Horario* como el caso anterior.

   c) El número de horas es 2: Llenar los renglones 1,2,4 y 5, de la columna *Horario*, con la información de la hora 1 y los renglones 3 y 6 con la información de la hora 2.

   d) El número de horas es 1: Llenar la columna *Horario* con la hora en la que ha dado clases el profesor elegido.

En la figura 4.14 vemos un ejemplo de la matriz *mat_1_solicitud*.

	Profesor	TC	Materia	Num_Materia	Horario
1	Margarita Elvira Chávez Cano	1	Estadística I	42	9
2	Margarita Elvira Chávez Cano	1	Estadística I	42	11
3	Margarita Elvira Chávez Cano	1	Estadística I	42	10
4	Margarita Elvira Chávez Cano	1	Estadística III	139	9
5	Margarita Elvira Chávez Cano	1	Estadística III	139	11
6	Margarita Elvira Chávez Cano	1	Estadística III	139	10

Figura 4.14: *Ejemplo de matriz de solicitudes de un profesor*

El proceso se repite para cada uno de los profesores en la matriz *mat_nom_prof_total* obtenida en la sección 4.5. En la matriz formada quitamos los renglones repetidos para poder simular adecuadamente los esqueletos.

4.7. Simulación de esqueletos

En esta sección vamos a explicar cómo generamos la matriz *mat_esqueleto*, la cual tiene 15 renglones y 335 columnas. En la entrada (i, j) se tiene el numero de grupos simulados para la *i-ésima* hora y la materia j. La matriz *mat_esqueleto* depende de la demanda de alumnos y de las solicitudes de los profesores.

El proceso que seguimos para obtener la matriz *mat_esqueleto* es el siguiente:

1. Obtener la matriz *mat_demanda_alumnos* con la demanda simulada del número de alumnos para el siguiente semestre (ver sección 4.3).
2. Obtener la matriz *mat_solicitudes* con las solicitudes simuladas de los profesores (ver sección 4.6).
3. Elegir un profesor de tiempo completo al azar.
4. Elegir al azar un horario y una materia que haya solicitado el profesor elegido en el paso anterior. Con estos datos sabemos las coordenadas (i, j) para las matrices *mat_demanda_alumnos* y *mat_esqueleto*.
5. Verificar que a esa materia en esa hora aún le sobran alumnos. En la entrada (i, j) de *mat_demanda_alumnos*.
6. Simular el número de alumnos para ese grupo (ver sección 4.4).
7. Restar el número de alumnos simulados en el paso anterior, de la materia y hora elegidas. En la entrada (i, j) de *mat_demanda_alumnos*.
8. Ese profesor ya no puede impartir clases a esa hora. Retirar renglones correspondientes de *mat_solicitudes*.
9. Repetir los pasos de 3 a 8.
10. Una vez que se terminen los profesores de tiempo completo, hacer los pasos de 3 a 9 con los profesores de asignatura.

Algunas notas a considerar del procedimiento son:

- Se le da prioridad de asignación a los profesores de tiempo completo.
- Los profesores sólo pueden tener asignadas a lo más 2 materias.
- Las condiciones de paro del proceso son:
    a) Ya se cubrió toda la demanda
    b) Ya no hay más profesores
    c) Llegar a una cota predefinida para que el ciclo no se haga infinito o tarde mucho en cumplir las condiciones anteriores.

En la figura 4.15 vemos un ejemplo de la matriz *mat_esqueleto* para el semestre 2020-1. Observemos las últimas 4 columnas que corresponden a las materias de Cálculo I, II, III y IV. Notamos que el número de grupos simulados para las materias de semestres pares (Cálculo II y IV) es mayor que para las materias de semestres impares. Esto se debe al comportamiento observado en 3.2 y el semestre 2020-2 es par.

	Estadística I	Investigación de Operaciones	Análisis de Redes	Cálculo de las Variaciones	Cálculo Diferencial e Integral I	Cálculo Diferencial e Integral II	Cálculo Diferencial e Integral III	Cálculo Diferencial e Integral IV
7-8	0	0	1	0	1	3	1	4
8-9	0	1	1	0	0	0	0	0
9-10	1	2	0	0	0	0	0	0
10-11	1	0	1	0	0	3	0	2
11-12	4	1	1	1	2	10	1	8
12-13	1	1	0	0	0	0	0	1
13-14	0	0	0	1	0	0	2	0
14-15	0	0	0	0	5	1	3	0
15-16	0	0	0	0	0	0	0	0
16-17	2	0	1	0	3	7	6	1
17-18	1	0	1	0	1	1	1	0
18-19	2	0	0	0	0	3	0	3
19-20	0	1	0	0	0	1	0	0
20-21	0	2	0	0	1	1	1	3
21-22	0	0	0	0	0	0	0	0

Figura 4.15: *Ejemplo de esqueleto para el semestre 2020-2*

4.7.1. Calificación de esqueletos de horario

$$
\begin{aligned}
L_{materia} &= -1 \quad \text{por cada materia no impartida} \\
x &= \text{promedio} \\
y &= \text{cupo} \\
L_{dif\_p\_c}(x,y) &= \begin{cases} \dfrac{a}{190}(x-y) & \text{si } x < y \\ -\dfrac{b}{190}(x-y) & \text{si } x \geqslant y \end{cases} \\
a &= 0.5 \\
b &= 0.8 \\
L^1_{categoria}(mat, solicitud) &= -c(categoria-1) \\
L^2_{categoria}(mat, solicitud) &= -c_1(categoria)
\end{aligned}
$$

Al momento de simular las solicitudes de materias para los profesores suponemos que la que está en primer lugar es la materia que más quiere dar, en segundo lugar, la segunda que quiere dar.

Primero asignar grupos a los profesores de tiempo completo. Después asignar grupos faltantes a los profesores de asignatura.

$L_{materia}$ es la penalización por no tener en el esqueleto una materia que necesitamos.

$L_{dif\_p\_c}$ es la penalización en la asignación de salones. Se tiene un grupo de tamaño x y un salón con capacidad y. Se penaliza con $\frac{a}{190}$ veces la diferencia entre x y y cuando el tamaño del grupo es menor a la capacidad del salón y se penaliza con $-\frac{b}{190}$ veces la diferencia entre x y y cuando el tamaño del grupo es mayor a la capacidad del salón.

El esqueleto depende de la demanda de alumnos y de las solicitudes de los profesores.

Primero se asignan materias a los profesores de tiempo completo y después a los de asignatura.

Los profesores de asignatura pueden quedarse sin materias asignadas.

Penalizaciones:

1. Si algún profesor pidió alguna materia y no se la dieron.
2. Si hay alumnos que necesitan una clase a alguna hora y no existe profesor que la imparta.
3. Con $\alpha \times num\_alumnos\_faltantes$ por cada alumno que te faltó en cada hora-materia que tenías que dar. $\alpha > 0$
4. Con $\beta \times num\_alumnos\_sobrantes$ por cada alumno que te pasaste en cada hora-materia que tenías que dar. $\beta > 0$

Queremos el esqueleto con el menor valor en $\alpha + \beta$

Con esto se obtiene un nuevo esqueleto.

4.8. Generación de asignaciones

Matriz de 3 columnas (Materia-Horario-Profesor), la cual tiene la información de las asignaciones. A cada renglón de la matriz de esqueletos se agrega un profesor. Se genera con el esqueleto obtenido del proceso del AG y de las solicitudes de los profesores.

4.8.1. Calificación de asignaciones de grupo

Capítulo 5

Teoría del Algoritmo Genético aplicado a los horarios

*** ESCRIBIR ACERCA DE LA TEORÍA DE AG ***

El algoritmo genético actualmente se utiliza para resolver problemas de optimización tanto discretos como continuos. Se basa en el mecanismo de la selección natural de Darwin, el cual nos indica que el individuo más apto sobrevive, por lo que entre mejores sean los padres, mejor es la descendencia.

Definimos a un cromosoma como una posible solución al problema. En nuestro caso representamos a un cromosoma por medio de una matriz con *j_materias* renglones y con 3 columnas las cuales representan la asignación de profesor, día y salón, respectivamente, por lo que el renglón j indica que la materia j es impartida por el profesor i, el día t, en el salón k.

El valor de adaptabilidad *fit(x)*, de cada cromosoma, se asigna al evaluar su utilidad en la función objetivo, entre mejor sea el cromosoma, más alto será su valor de adaptabilidad. Los mejores cromosomas de la población actual pasan directamente a la siguiente generación. Se dice que la población evoluciona por medio de tres operadores hasta una condición de paro, los operadores son: *selección, entrecruzamiento (crossover) y mutación*.

Los pasos del algoritmo se muestran a continuación:

1. Se inicia con un grupo de cromosomas generados aleatoriamente, a los cuales se les calcula su valor de adaptabilidad

2. La probabilidad de que el cromosoma k sea elegido para el entrecruzamiento *(crossover)*, es:

$$p_k = \frac{fit(x)}{\sum_{h=1}^{pop} fit(h)} \quad \text{donde } pop \quad \text{es el tamaño de la población de cromosomas}$$

3. En el entrecruzamiento se mezclan dos padres para generar nuevas soluciones. Se genera un número aleatorio entre cero y uno, r, si $r < 0.6$ la primer columna de M_{ij} y la primer columna de M_{ti} del padre 1 se copian en la nueva solución, las demás columnas se llenan con las columnas del padre 2. Si la nueva solución no es factible, en la matriz M_{ij}, si alguna materia tiene asignada dos profesores, se selecciona uno de ellos de manera aleatoria y el otro se elimina de esa asignación; en caso de que alguna materia no tenga profesor asignado, se le asigna uno aleatoriamente.

4. Se actualiza la matriz M_{ti}.

5. Se aplica el operador *mutación*, se selecciona un profesor de manera aleatoria y se cambia el día en el que más tiene clase por el día que menos clases imparte. Ésto se aplica para cada profesor de manera aleatoria, sin repetición.

6. Una vez generadas las nuevas soluciones se elige la mejor entre todas ellas.

5.1. Ciclo de la evolución natural

Figura 5.1: *Algoritmo Genético*

5.1.1. Selección

5.1.2. Cruce

5.1.3. Mutación

5.1.4. Reemplazamiento

5.2. Algoritmo Genético aplicado a la generación de esqueletos de horario

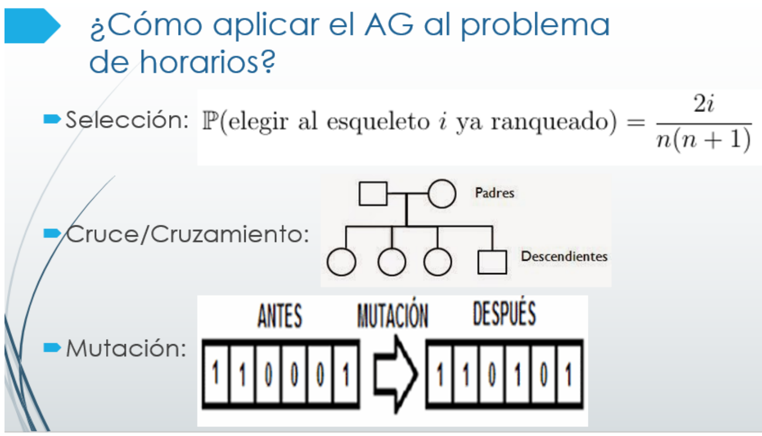


Figura 5.2: *Algoritmo Genético aplicado*

5.3. Algoritmo Genético aplicado a la generación de asignaciones de grupos

Capítulo 6

Resultados del Algoritmo Genético

Capítulo 7

Comportamiento de la selección

Capítulo 8

Conclusiones

La división que se hizo de los datos es estadísticamente adecuada.

Se encontró que el AG es una buena opción para solucionar este problema de maximización.

Este trabajo apoya las necesidades de los alumnos de la Facultad.

Con el fin de encontrar más posibles aplicaciones del programa realizado en este trabajo, se buscaron diferentes páginas de horarios en distintas facultades de la UNAM y de otras universidades. No se pudieron encontrar páginas con todas las características que tienen las páginas de la FC. Algunas de las páginas que se encontraron son las siguientes:

- *Facultad de Filosofía y Letras UNAM:* Se encontró una estructura en las páginas web con las cuales se puede acceder a la información por carrera, pero no se puede acceder a la información de semestres anteriores y dentro de éstas páginas no se pueden encontrar el número de alumnos inscritos por cada materia, por lo que no sería posible una simulación del número de alumnos.

  `https://servicios-galileo.filos.unam.mx/horarios/ordinarios/1354`

  `https://servicios-galileo.filos.unam.mx/horarios/ordinarios/1355`

  `https://servicios-galileo.filos.unam.mx/horarios/ordinarios/1359`

- *Facultad de Ingeniería UNAM:* En la siguiente página web se puede seleccionar una materia y buscar la información de ella del semestre en curso, no se puede acceder a información de semestres anteriores y no se tiene alguna estructura para buscar de manera automática los datos.

  `https://www.ssa.ingenieria.unam.mx/horarios.html`

  Una vez que se ingresa a la materia, se puede encontrar información del salón, horario, cupo y vacantes, se podría obtener el número de alumnos inscritos al restar el cupo del número de vacantes, pero al no tener información de semestres anteriores en todo momento, la recopilación de información tardaría años.

- *FES Acatlán (Actuaría):* En la siguiente página web se pueden descargar los horarios del semestre en curso en un archivo de Excel el cual no contiene información del número de alumnos inscritos en el grupo, tampoco se puede obtener información de semestres anteriores. No se puede utilizar la aplicación *SelectorGadget* para obtener la información.

`http://www.actuaria.acatlan.unam.mx/`

- *FES Iztacala (Psicología):* Se encontró una estructura en las páginas web con las cuales se puede acceder a la información en archivos pdf de algunos semestres, dependiendo si el semestre en curso es par o impar. No se puede utilizar la aplicación *SelectorGadget* para obtener la información.

  `https://psicologia.iztacala.unam.mx/avisos2020/horarios21_1/21-1_3-TERCER%20SEMESTRE.pdf`

  `https://psicologia.iztacala.unam.mx/avisos2020/horarios21_1/21-1_5-QUINTO%20SEMESTREv1108.pdf`

- *Centro de Nanociencias y Nanotecnología (Nanotecnología):* Al igual que en el caso anterior la información de las páginas que se muestran a continuación están en archivos pdf por lo que no se puede utilizar la aplicación *SelectorGadget* para obtener la información.

  `https://nanolic.cnyn.unam.mx/sitio/wp-content/uploads/2020/09/H-1A-2021-1.pdf`

  `https://nanolic.cnyn.unam.mx/sitio/wp-content/uploads/2020/09/H-1B-2021-1.pdf`

  `https://nanolic.cnyn.unam.mx/sitio/wp-content/uploads/2020/09/H-3A-2021-1.pdf`

- *Facultad de Química UNAM:* En la siguiente página web se pueden seleccionar todas las materias impartidas en la facultad o por carrera.

  `http://escolares.quimica.unam.mx/Horarios/hor_def_e2.php4`

  Una vez que se eligió alguna opción, se muestra un listado, en la siguiente url, con las posibles materias que se pueden elegir.

  `http://escolares.quimica.unam.mx/Horarios/hor_def_pre_e2.php4`

  Finalmente se accede a la información con la siguiente página web.

  `http://escolares.quimica.unam.mx/Horarios/hor_tot_e2.php4`

  No importa las opciones que se elijan, siempre se obtienen esas mismas urls por lo que no hay alguna estructura para poder buscar la información automáticamente.

- *ESIME Zacatenco IPN (Ingeniería en Control y Automatización):* Se encontró una estructura en las páginas web pero no se puede encontrar el número de alumnos inscritos por materia por lo que no es posible realizar una simulación del número de alumnos.

  `http://horarios.esimez.ipn.mx/horarios/VHorGpoAl.aspx?Gpo=1AM8&PaId=57`

  `http://horarios.esimez.ipn.mx/horarios/VHorGpoAl.aspx?Gpo=1AV1&PaId=57`

- *ITAM:* En este caso se debe de seleccionar una materia y luego se despliega la información, sin importar la selección de la materia, las url son las mismas por lo que no se tiene una estructura en las páginas web.

  `http://escolar1.rhon.itam.mx/licenciaturas/horarios/seleccion_03.asp`

  `http://escolar1.rhon.itam.mx/licenciaturas/horarios/pormateria_03.asp`

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

09:14:54 p. m. del 25-septiembre-2020

Los grupos programados para el semestre OTO?O 2016 LICENCIATURA de la materia CALCULO DE PROBABILIDADES.,I son:

DEPTO.	CLAVE	GRUPO	TEORÍA O LABORATORIO	NOMBRE	PROF.	CRÉDITOS	HORARIO	DÍAS	SALÓN	CAMPUS	COMENTARIOS
EST	14101	001	T	CALCULO DE PROBABILIDADES.,I	VICTOR MANUEL ARMANDO AGUIRRE TORRES	6	10:00-11:30	LU MI	RH314	RIO HONDO	
EST	14101	002	T	CALCULO DE PROBABILIDADES.,I	ANA MEDA GUARDIOLA	6	10:30-12:00	MA JU	RH313	RIO HONDO	
EST	14101	003	T	CALCULO DE PROBABILIDADES.,I	ERICK MIER MORENO	6	08:30-10:00	LU VI	RHPB2	RIO HONDO	
EST	14101	004	T	CALCULO DE PROBABILIDADES.,I	LUIS ENRIQUE NIETO BARAJAS	6	13:00-14:30	LU MI	RH314	RIO HONDO	
EST	14101	005	T	CALCULO DE PROBABILIDADES.,I	MIGUEL ANGEL MENDEZ ANTONIO	6	08:30-10:00	LU MI	RH313	RIO HONDO	

Figura 8.1: *ITAM Probabilidad I*

- *Universidad La Salle:* Se encontró que las páginas tienen una cierta estructura y también se tiene la información del número de alumnos inscritos por materia pero los archivos son pdf por lo que no se puede utilizar la aplicación *SelectorGadget* para obtener la información.

  `https://cienciasquimicas.lasalle.mx/wp-content/uploads/2020/08/QFB-291.pdf`

  `https://cienciasquimicas.lasalle.mx/wp-content/uploads/2020/08/QFB-391.pdf`

  `https://cienciasquimicas.lasalle.mx/wp-content/uploads/2020/08/QFB-991.pdf`

- *Universidad Panamericana:* No se encontraron horarios de materias, sólo de exámenes y de entrenamientos.

  `https://www.up.edu.mx/sites/default/files/fechas_de_examenes_humanidades_1202.pdf`

  `https://www.up.edu.mx/en/media/22960`

Sólo se encontró la misma estructura en las otras carreras de la FC, por lo que se puede ajustar el programa realizado en este trabajo para ellas. Algunas consideraciones que se deberían de tomar en cuenta son por ejemplo que las materias impartidas en los laboratorios duran más de una hora, no todas las materias se imparten todos los días, existen varias materias que no duran horas enteras. A continuación se presentan algunos ejemplos:

- *Biología:*

  `http://www.fciencias.unam.mx/docencia/horarios/20172/181/1601`

- *Ciencias de la Tierra:*

  http://www.fciencias.unam.mx/docencia/horarios/20182/1439/1318

- *Física:*

  http://www.fciencias.unam.mx/docencia/horarios/20191/1081/830

- *Física Biomédica:*

  http://www.fciencias.unam.mx/docencia/horarios/20192/2016/1735

- *Manejo Sustentable de Zonas Costeras:*

  http://www.fciencias.unam.mx/docencia/horarios/20181/1262/386

Apéndice A

Observaciones / Notas

1. La matriz `mat_posibles_url` se define con un tamaño fijo antes de correr el algoritmo para que no se demore por tener un objeto que va cambiando de tamaño, por lo que al final de haberle aplicado la función se le deben de quitar los renglones que no tienen información.

2. La función *casos_alumnos* convierte los *NA* de la columna *Alumnos* de *m_grande* en ceros pero al generar *m_grande_total* y pasarla por la función *limpia_m_grande* se eliminan los *NA* y se cambian por ceros por lo tanto no es necesaria la función *casos_alumnos*, basta pasar la columna correspondiente a *Alumnos* de *m_grande_total*.

3. Cuando se hacen comparaciones se toman los valores reales y se les restan los valores simulados $(Reales - \mathbb{E}[Simulados])$

4. Con las gráficas *heatmap* se revisa si el modelo es adecuado o si se debe modificar algo. Se espera que las gráficas sean de color claro ya que nos interesa que el número de grupos y alumnos simulados se parezca al real.

5. Se tienen dos tipos de matrices las cuales llamaremos *m_objetivo* y *m_definición*; las matrices *m_objetivo* son las que tienen la información que se utiliza para la asignación; las matrices *m_definición* nos sirven para dos cosas:

    a) Respaldo de la descripción de cada columna

    b) Para guardar los índices en los que se encuentran las columnas

6. Las matrices tipo *m_definición* son:

    a) mat_def_columnas_MG

    b) mat_def_grupos_reales

    c) mat_def_grupos_simulados

7. Las matrices tipo *m_objetivo* son:

    a) m_grande

    b) m_grande_total

c) …

8. La función *checa_ind_materia* se encarga de obtener lo índices de las columnas de las matrices tipo *m_definición* para poder sacar información de *m_grande* o de *m_grande_total*.

9. Para las simulaciones se utiliza la información anterior a la del semestre que se quiere simular para no tener información real dentro de los datos para la simulación.

10. En caso de querer elegir la capacidad del salón se va a elegir la mayor de sus capacidades (comparando las capacidades que se han tenido a lo largo de varios semestres).

11. Las matrices *m_grande* y de *m_grande_total* tienen información real.

12. En los ciclos que recorren renglones y columnas de matrices, siempre es más rápido hacer (de afuera hacia adentro) primero las columnas y luego los renglones.

    Si se tiene una matriz con entradas (i, j) entonces:

```
for(j){
 for(i){
     m[i,j]
   }
 }
```

Código A.1: *Ejemplo de ciclo for*

13. El vector *vec_nom_materias_total* tiene los nombres de las materias, sin repeticiones, que se utiliza para las simulaciones.

14. El vector *vec_excepciones* tiene las posibles ecxepciones en las que las funciones que extraen información pueden caer, de esta manera se pueden generar nuevas funciones para corregir esos casos.

15. La siguiente imagen es el resultado de la función *imprime_info_idiomas* la cual muestra la información de los idiomas. Dicha función arroja un vector con los semestres que requieren modificación.

```
La matriz m_grande del semestre  20081  no tiene clases de  Inglés
La matriz m_grande del semestre  20082  no tiene clases de  Inglés
La matriz m_grande del semestre  20091  no tiene clases de  Inglés
La matriz m_grande del semestre  20092  no tiene clases de  Inglés
La matriz m_grande del semestre  20101  no tiene clases de  Inglés
La matriz m_grande del semestre  20102  no tiene clases de  Inglés
La matriz m_grande del semestre  20111  no tiene clases de  Inglés
La matriz m_grande del semestre  20112  no tiene clases de  Inglés
La matriz m_grande del semestre  20121  no tiene clases repetidas de  Inglés
La matriz m_grande del semestre  20122  no tiene clases repetidas de  Inglés
La matriz m_grande del semestre  20131  no tiene clases repetidas de  Inglés
La matriz m_grande del semestre  20132  no tiene clases repetidas de  Inglés
La matriz m_grande del semestre  20141  no tiene clases repetidas de  Inglés
La matriz m_grande del semestre  20142  no tiene clases repetidas de  Inglés
La matriz m_grande del semestre  20151  no tiene clases repetidas de  Inglés
En el semestre  20152  se tienen  2  clases repetidas de  Inglés
En el semestre  20161  se tienen  3  clases repetidas de  Inglés
En el semestre  20162  se tienen  4  clases repetidas de  Inglés
En el semestre  20171  se tienen  1  clases repetidas de  Inglés
En el semestre  20172  se tienen  2  clases repetidas de  Inglés
En el semestre  20181  se tienen  4  clases repetidas de  Inglés
En el semestre  20182  se tienen  3  clases repetidas de  Inglés
La matriz m_grande del semestre  20191  no tiene clases repetidas de  Inglés
En el semestre  20192  se tienen  1  clases repetidas de  Inglés
En el semestre  20201  se tienen  1  clases repetidas de  Inglés
```

Figura A.1: *Resumen de clases de ingles antes de modificación*

Con esta información se decidió observar caso por caso los renglones que requieren modificación para la matriz *m_grande*

16. Debido a la situación en la que estamos viviendo actualmente, ahora más que nunca es necesario tener un programa para la asignación de horarios que permita la realización de las asignaciones sin tener la necesidad de hacer reuniones en persona, ya que al proseguir con las medidas de distanciamiento social, las reuniones antiguamente hechas en persona se tendrían que hacer por medio de alguna plataforma digital las cuales no necesariamente son las más óptimas ya que dependen de la señal de todos los participantes para que haya una comunicación de manera fluída. Debido a ésto, el programa es una buena solución.

17. Al hacer las simulaciones del número de alumnos el redondeo es hacia arriba, usando la función *ceiling*.

18. El vector *vec_nom_materias_total*, que contiene el nombre de las materias se definió en la lista *param* para poder tomarlo en las diferentes funciones.

19. Para resolver un problema, pensar en los pasos en los que se puede dividir dicho problema, usualmente se requieren entre 3 y 8 pasos o casos para obtener un producto final. Para cada paso hacer una función.

    Se tienen dos posibles estructuras:

    a) La función del paso n manda a llamar a la del paso $n-1$.

       Ej.

       simula_grupos{*simula_gpos_1_sem* {*simula_gpos_1_materia* {*simula_tam_gpo*

    b) Se tiene una función principal que manda a llamar a las funciones de cada paso:

       Ej.

```r
gen_asignacion_completa <- function(sem_ini,sem_fin){
  # Se carga y se limpia la lista de urls (para no tener paginas sin informacion,...)
   list_url <- Actualiza_list_url(list_url)

  # Se obtiene "m_grande" y se genera un archivo para cada semestre
  for(k in 1:length(semestres)){
    sem_info <- semestres[k]
    directorio_info[k] <- gen_m_grande(sem_info,list_url)
  }

  # Se genera el esqueleto del semestre que se quiere obtener
  mat_esqueleto <- gen_esqueleto(directorio_info,param)

  # Se genera la matriz de solicitudes de todos los profesores
  mat_solicitudes <- gen_solicitudes(param)

  # Se genera la matriz de asignaciones de todos los profesores
```

```
    mat_asignaciones <- gen_asignacion(mat_esqueleto,mat_solicitudes,param)

    return(mat_asignaciones)
  }
}

```

Código A.2: *Ejemplo de estructura de funciones*

20. Pudiera ser que haya un apéndice con “Observaciones” utlizando las notas escritas.

21. Todo lo que se escriba debe tener un propósito, sino quitarlo.

22. La información que se puede encontrar actualmente (debido a la pandemia) en las páginas web de los horarios de la FC no es la misma que la mostrada a lo largo del trabajo ya que ahora no se tiene información del salón, o del número de alumnos inscritos por materia, ni los lugares disponibles por grupo.

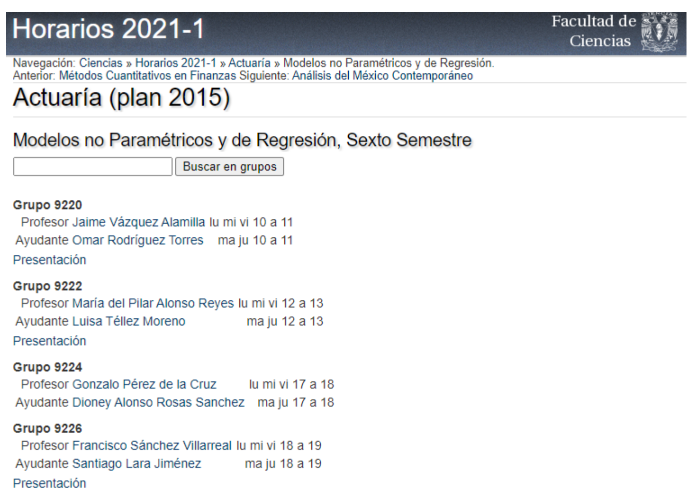


Figura A.2: *Ejemplo de horarios de semestre 2021-1*

23. Notas de T26

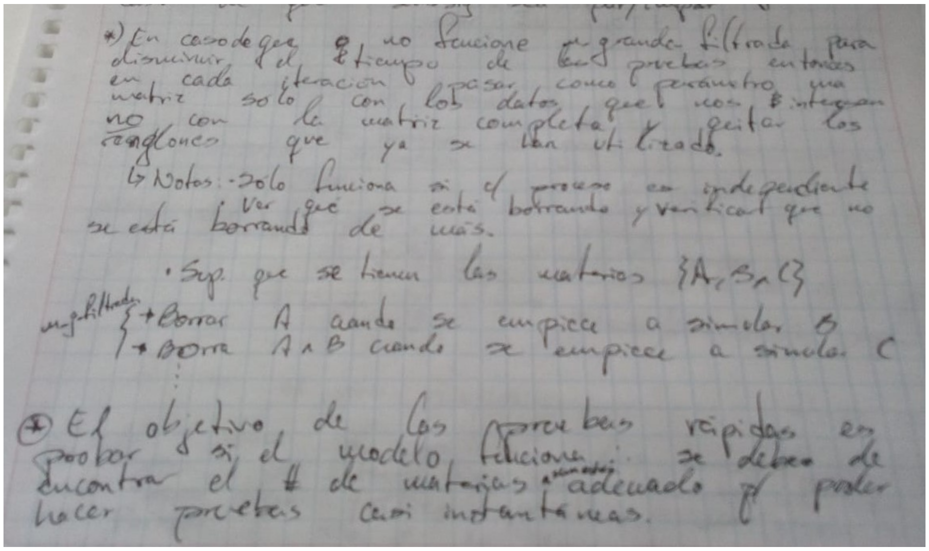

*) En caso de que no funcione el grande filtrado para disminuir el tiempo de las pruebas entonces en cada iteración pasar como parámetro una matriz solo con los datos que nos integran no con la matriz completa y quitar los renglones que ya se han utilizado.
↳ Notas: - Solo funciona si el proceso es independiente. Ver que se está borrando y verificar que no se está borrando de más.
• Sup. que se tienen las matrices {A, B, C}
ej. filtrado { + Borrar A cuando se empiece a simular B
+ Borra A ∧ B cuando se empiece a simular C
⊛ El objetivo de las pruebas rápidas es probar si el modelo funciona. Se deben de encontrar el # de matrices adecuado p/ poder hacer pruebas casi instantáneas.

Figura A.3: *Notas de T26*

24. En caso de tener subsecciones: entre 3 y 4
25. La estructura de cada párrafo debe ser de tipo *reloj de arena*. Ir de lo general a lo particular y volver a lo general con una conclusión.
26. Sea $D = \dfrac{r-s}{s}$, donde r son datos reales, s datos simulados y D la diferencia relativa, se busca que $D \in \left[-\frac{1}{2}, \frac{1}{2}\right]$.
27. Ejemplo del uso del comando *Roxygen* para comentar las funciones en *R*.

```
#' Add together two numbers
#'
#' @param x A number
#' @param y A number
#' @return The sum of \code{x} and \code{y}
#' @examples
#' add(1, 1)
#' add(10, 1)
add <- function(x, y) {
  x + y
}
```

Figura A.4: *Ejemplo de Roxygen*

28. Escribir en el archivo de LaTeX pequeños comentarios de la idea que se quiere transmitir en cada párrafo (de 2 a 3 palabras claves). Ésto sirve para referencias futuras y para ordenar los párrafos con mayor facilidad.
29. Escribir párrafos de 2 a 3 enunciados completos, no dejar enunciados solos a menos que contengan información muy importante.

30. En caso de tener más de 10 referencias bibliográficas utilizar *Mendeley* para genera un archivo *.bib* y ponerlo en la tesis para tener la bibliografía.

31. Cuidar el tamaño de letra en las gráficas que se pongan

32. No poner abreviaturas en los títulos.

33. Ser muy directa al escribir, pero explicar mucho más (platicar más). No hacer enunciados tan largos. Escribir una idea por enunciado. No sólo escribir en párrafos, utilizar listas, tablas, ...

34. La imagen 3.4 tiene título en inglés, se tienen 2 opciones: dejarlo así o buscar cómo cambiarlo.

35. Recordar la diferencia entre:

    - Número de alumnos inscritos

    - Número de alumnos reales

    - Número de alumnos que toman clase por cada horario (no se toman en cuenta los alumnos que empalman clases)

36. Para la elección de q_1 y q_2 se debe darle prioridad a la varianza no al mín y al máx porque se pueden tener casos en los que el mín y el máx estén muy cercanos a cero (gráfica superior) pero su varianza es grande. Queremos que la varianza se encuentre alrededor del cero (gráfica inferior).

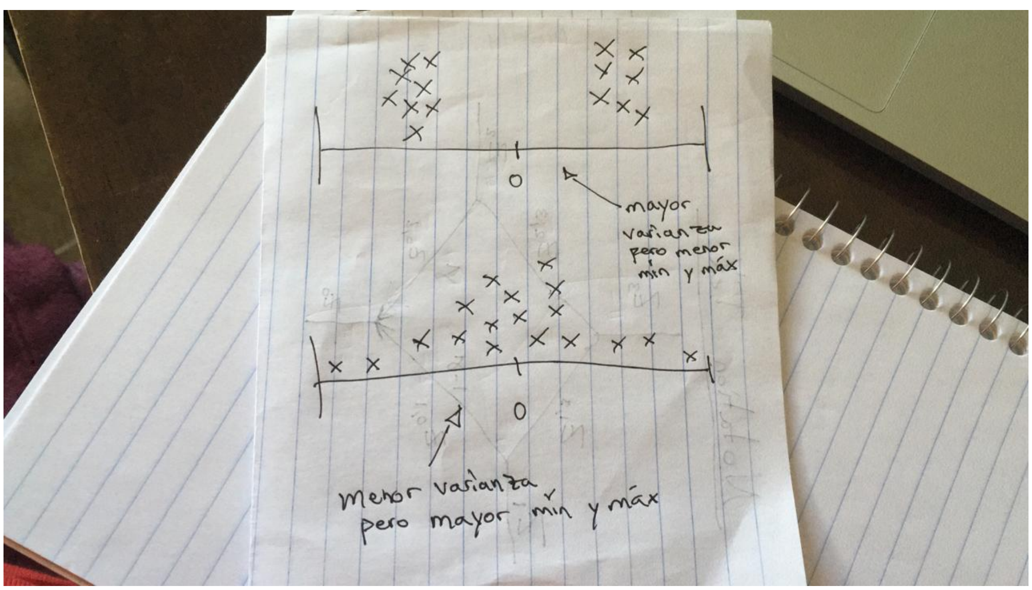


Figura A.5: *Ejemplo de varianza*

37. Preferir sacrificar el B/N en las imágenes impresas para tener una mejor versión digital a color.

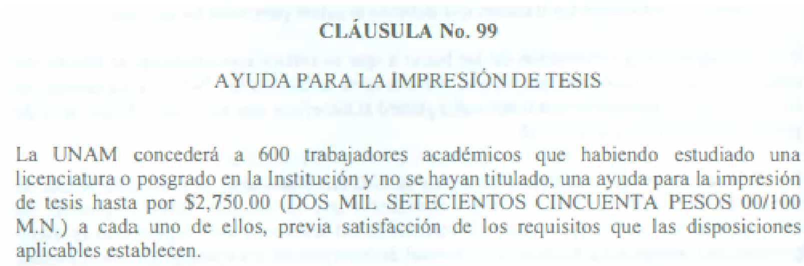
38. Guardar figuras hechas en R con el comando: *dev.print(pdf, “Figures/Fig_Examples_of_GB_distributions* *height=5)*

39. Arrigo dijo que posiblemente alguien se va a quejar de no tomar en cuenta la preferencia de los profesores al realizar las solicitudes.

40. Un histograma nos muestra la representación de la distribución empírica de un conjunto de datos. Cada barra en el histograma representa la frecuencia de un intervalo sobre el rango de las observaciones que se tienen.

41. Cláusula 99 CCTPA: Ayuda para la impresión de la tesis.

    `https://www.personal.unam.mx/Docs/Contratos/AAPAUNAM20132015.pdf`



CLÁUSULA No. 99

AYUDA PARA LA IMPRESIÓN DE TESIS

La UNAM concederá a 600 trabajadores académicos que habiendo estudiado una licenciatura o posgrado en la Institución y no se hayan titulado, una ayuda para la impresión de tesis hasta por $2,750.00 (DOS MIL SETECIENTOS CINCUENTA PESOS 00/100 M.N.) a cada uno de ellos, previa satisfacción de los requisitos que las disposiciones aplicables establecen.

Figura A.6: *Cláusula 99 CCTPA: Ayuda para la impresión de la tesis*

42. Equivalencias de nombres para estadística:
    a) Estadística I - Inferencia Estadística
    b) Estadística II - Modelos no Paramétricos y de Regresión
    c) Estadística III - Modelos de Supervivencia y de Series de Tiempo

43. La frecuencia relativa en los histogramas no refleja directamente el porcentaje. Se debe multiplicar el valor del eje Y por el ancho del intervalo por 100 para obtener cifras en porcentaje. El área total de las barras sumará 1 (10).

44. No confundir las carpetas de *Figuras* del GitHub con la del pdf.

45. Ya no son necesarias las pruebas de bondad de ajuste porque los tamaños de grupo se van a simular con respecto a los profesores. Ver $T_{32}xx$)

46. Los archivos *README* sirven para explicar las cosas a los demás.

47. Si los grupos pequeños dan muchos problemas podemos considerar quitarlos.

48.

49.

50.

Apéndice B

Resultados útiles

Definición B.1. ***Estimador máximo verosímil de*** λ

Sean $X_1, X_2, \ldots, X_n$ *una muestra aleatoria de una población con función de densidad de probabilidad Poisson*(λ)*. Su función de densidad es:*

$$f(x) = \mathrm{e}^{-\lambda} \frac{\lambda^x}{x!} \tag{B.1}$$

$$
\begin{aligned}
\mathscr{L}(X_1, X_2, \ldots, X_n; \lambda) &= \prod_{i=1}^{n} \left(\mathrm{e}^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \right) \\
&= \mathrm{e}^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \\
\textit{Sacamos ln} & \\
ln\mathscr{L}(X_1, X_2, \ldots, X_n; \lambda) &= -n\lambda + \sum_{i=1}^{n} x_i ln\lambda - ln \prod_{i=1}^{n} x_i! \\
\textit{Derivamos con respecto a } \lambda & \\
\frac{\partial}{\partial \lambda} ln\mathscr{L}(\underline{X}; \lambda) &= -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} \\
\textit{Igualamos a cero} & \\
-n + \frac{\sum_{i=1}^{n} x_i}{\lambda} &= 0 \\
\Rightarrow & \\
\textit{Despejamos } \lambda & \\
\hat{\lambda} &= \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} \\
\textit{Derivamos otra vez} & \\
\frac{\partial^2}{\partial \lambda} ln\mathscr{L}(\underline{X}; \lambda) &= -\frac{\sum_{i=1}^{n} x_i}{\lambda^2} < 0 \\
\therefore \hat{\lambda} &= \bar{x} \quad \textit{es el estimador máximo verosímil}
\end{aligned}
$$

Apéndice C

Abreviaturas

ABREVIATURA	SIGNIFICADO
CdC	Ciencias de la Computación
ESIME	Escuela Superior de Ingeniería Mecánica y Eléctrica
FC	Facultad de Ciencias de la UNAM
FES	Facultad de Estudios Superiores
ITAM	Instituto Tecnológico Autónomo de México
MatAp	Matemáticas Aplicadas
TC	Tiempo Completo
UNAM	Universidad Nacional Autónoma de México
URL	Uniform Resource Locator
a	b

Tabla C.1: *Abreviaturas*

Bibliografía

[1] Casella G., (2006), *Statistical Inference*, Thomson Press

[2] Chatfield C. y Xing H., (2019), *The Analysis of Time Series An Introduction with R*, Chapman & Hall/CRC

[3] Gibbons J. D. y Chakraborti S., (2011), *Nonparametric Statistical Inference*, Chapman & Hall/CRC

[4] Miller L. H., (1956), *Journal of the American Statistical Association*, Vol. 51, No. 273, pp. 111-121.

[5] Montgomery D., Jennings C. y Kulahci M., (2015), *Introduction to Time Series Analysis and Forecasting*, Wiley

[6] Rincón L., (2007), *Curso intermedio de probabilidad*, UNAM

[7] Rubinstein R. y Kroese D., (2016), *Simulation and the Monte Carlo Method*, Wiley

[8] Shumway R. y Stoffer D., (2017), *Time Series Analysis and Its Applications: With R Examples*, Springer

[9] Yazdani M., Naeri B. y Zeinali E., (2017), *Algorithms for university course scheduling problems*, Tehnički vjesnik 24, Suppl. 2, 241-247

[10] `https://estadisticafciencias.files.wordpress.com/2019/08/introduccic3b3n-a-la-estadc3adstica-versic3b3n-final.pdf`