

Parcial 1 - Probabilidad y Estadística Fundamental

Wulfredo Javier Barco Godoy
Ángela Zoraya Cortés Yepes
Juan Manuel García Mejía
Andrés Felipe Patiño Nivia
Jafeth Paz Cortés

Universidad Nacional de Colombia, Bogotá, Colombia

6 de noviembre de 2021

1. Base de datos Alumnos

La base de datos Alumnos contiene las variables de matrícula, carrera, estrato, género, promedio, edad, peso y altura. Con esta, se van a realizar las gráficas y cálculos de medidas de estadística.

2. Gráficas de género, carrera y estrato

Todas las gráficas fueron hechas con ggplot, y los datos procesados con el tidyverse.

2.1. Barras

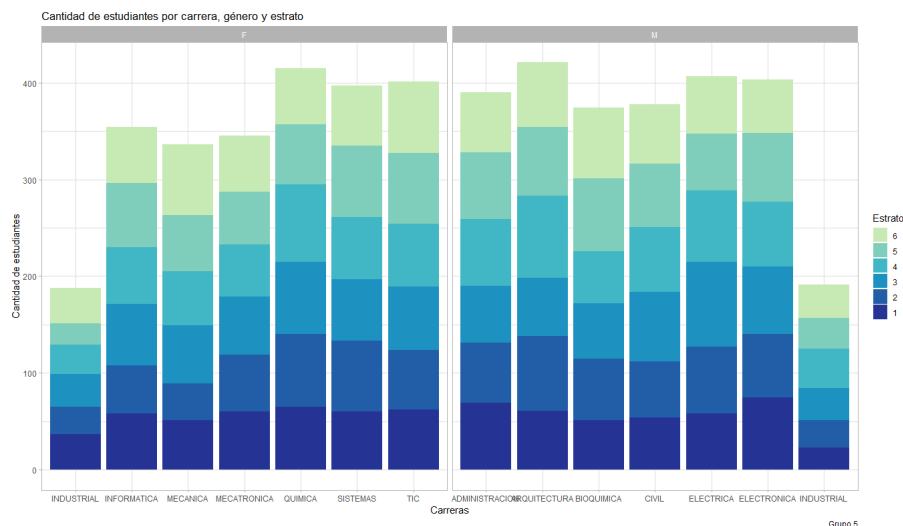
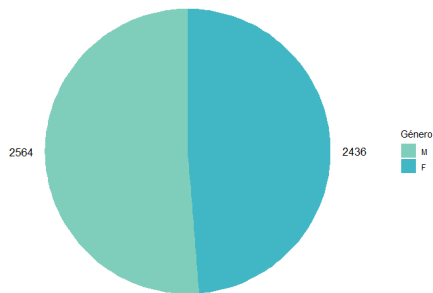


Figura 1: Relación entre género, carrera y estrato

Este gráfico revela, a primera vista, un interesante patrón: las carreras están divididas por género. La mitad de ellas son femeninas y la otra mitad masculinas, a excepción de la Ingeniería Industrial, que la componen en partes casi iguales hombres y mujeres. Este tipo de situaciones no se encuentran en la vida real.

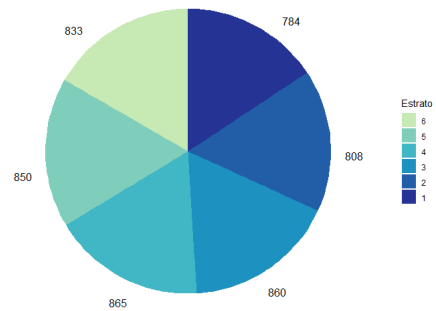
2.2. Tortas

Proporción de género



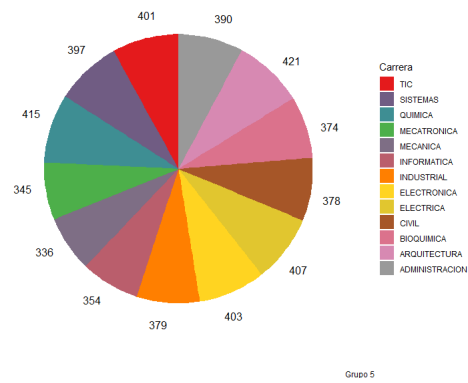
(a) Torta de género

Proporción de estrato



(b) Torta de estratos

Proporción de carrera



(a) Torta de carreras

El conjunto de tortas también muestra un comportamiento peculiar: todos los datos están distribuidos de manera aproximadamente igual. Es razonable en el caso del género, pues es la proporción de la población general (ignorando, claro, la desigualdad de género en la educación); sin embargo, una distribución de este estilo no suele ocurrir en la estratificación y las carreras de una universidad (hay facultades más grandes que otras).

3. Carrera & género

	Carrera	Genero	Cantidad
1	ADMINISTRACION	M	390
2	ARQUITECTURA	M	421
3	BIOQUIMICA	M	374
4	CIVIL	M	378
5	ELECTRICA	M	407
6	ELECTRONICA	M	403
7	INDUSTRIAL	F	188
8	INDUSTRIAL	M	191
9	INFORMATICA	F	354
10	MECANICA	F	336
11	MECATRONICA	F	345
12	QUIMICA	F	415
13	SISTEMAS	F	397
14	TIC	F	401

Esta tabla muestra más precisamente lo concluido en la gráfica 1: las carreras son exclusivamente femeninas o masculinas, a excepción de la Ingeniería Industrial. Esto también explica la paridad de género de la gráfica 2a.

4. Carrera & Estrato

	Carrera	Est 1	Est 2	Est 3	Est 4	Est 5	Est 6
1	ADMINISTRACION	0.18	0.16	0.15	0.18	0.18	0.16
2	ARQUITECTURA	0.14	0.18	0.14	0.20	0.17	0.16
3	BIOQUIMICA	0.14	0.17	0.15	0.14	0.20	0.20
4	CIVIL	0.14	0.15	0.19	0.18	0.17	0.16
5	ELECTRICA	0.14	0.17	0.22	0.18	0.14	0.15
6	ELECTRONICA	0.19	0.16	0.17	0.17	0.18	0.14
7	INDUSTRIAL	0.16	0.15	0.18	0.19	0.14	0.19
8	INFORMATICA	0.16	0.14	0.18	0.17	0.19	0.16
9	MECANICA	0.15	0.11	0.18	0.17	0.17	0.22
10	MECATRONICA	0.17	0.17	0.17	0.16	0.16	0.17
11	QUIMICA	0.16	0.18	0.18	0.19	0.15	0.14
12	SISTEMAS	0.15	0.18	0.16	0.16	0.19	0.16
13	TIC	0.15	0.15	0.16	0.16	0.18	0.18

Cuadro 1: Tabla de frecuencia relativa

La tabla 1 muestra datos exactos, pero es más fácil ver estos números ilustrados:

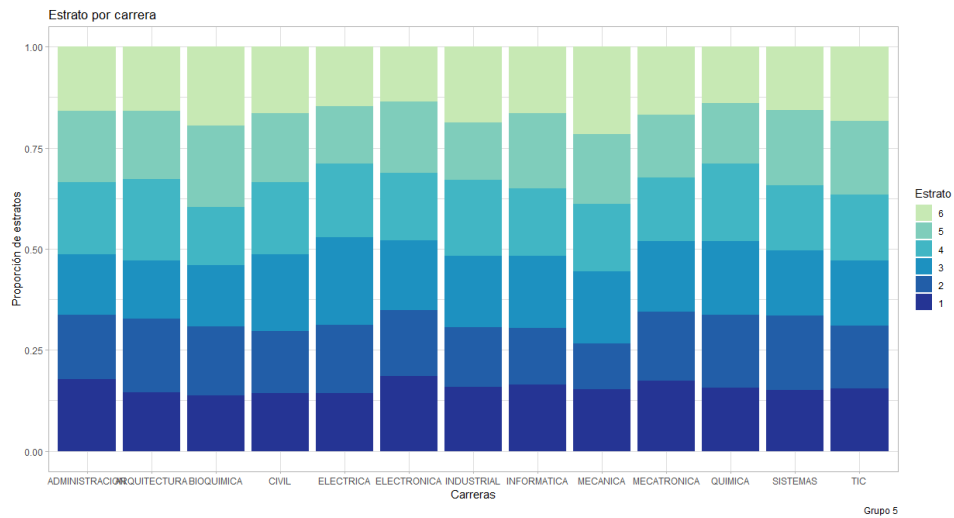


Figura 4: Proporción de estratos por carrera

Esta gráfica de frecuencia relativa de estratos muestra una distribución aproximadamente igual por carrera. No parece haber correlación entre ellas.

5. Gráficos de Edad

5.1. Barras

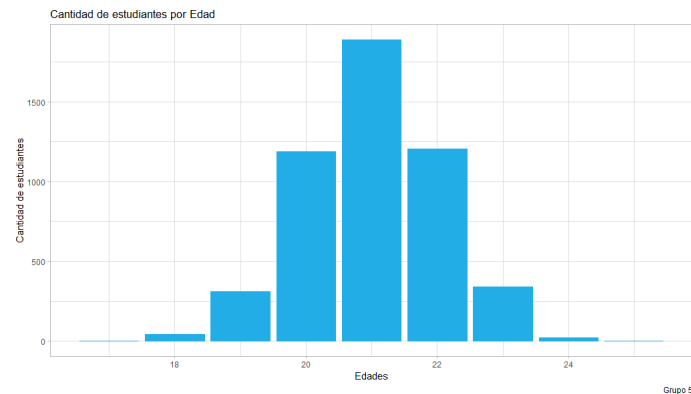


Figura 5: Edad por carrera

Es evidente la poca prevalencia de las edades 17, 18, 24 y 25. Adicionalmente, muestra mejor la ligera mayoría de personas de 21 años; mientras que los pares 20-22 y 19-23 tienen frecuencias similares. Recuerda a una distribución normal.

5.2. Torta

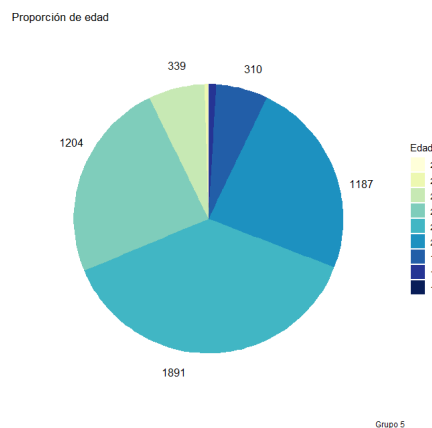


Figura 6: Proporción de edades

La torta no aporta información adicional al gráfico 5.

6. Gráficos de Variables Continuas

Se tomaron como variables continuas la altura, promedio y peso.

6.1. Histogramas

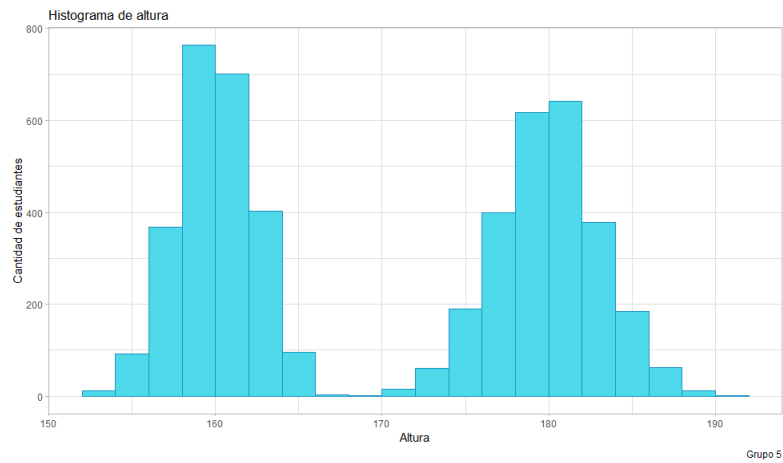


Figura 7: Histograma de altura

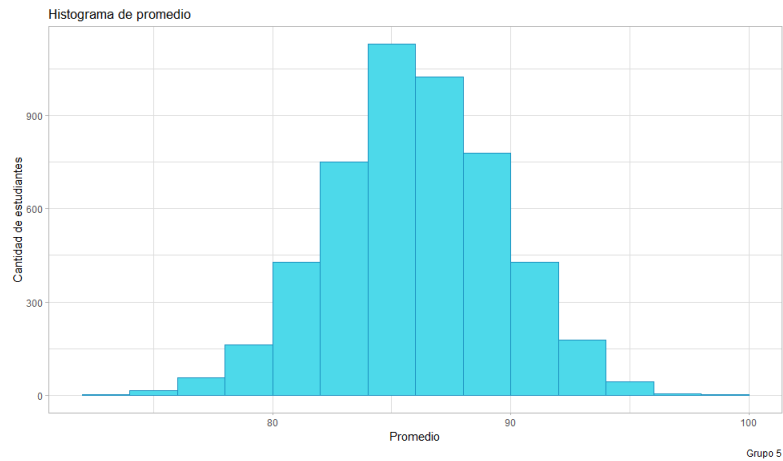


Figura 8: Histograma de promedio

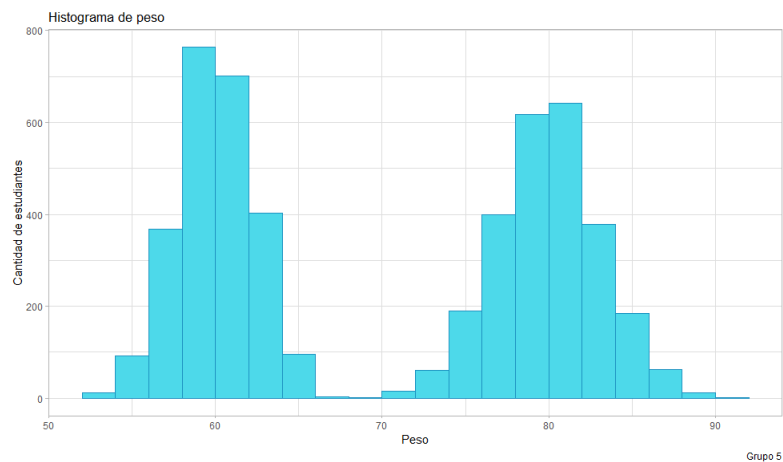


Figura 9: Histograma de peso

El histograma de promedio no aporta mucha información, aunque tal vez sugiera una ligera asimetría hacia la izquierda. Sin embargo, las variables fisiológicas (altura y peso) sí muestran patrones interesantes: hay dos picos completamente separados entre sí, con distribuciones generales casi idénticas.

6.2. Caja

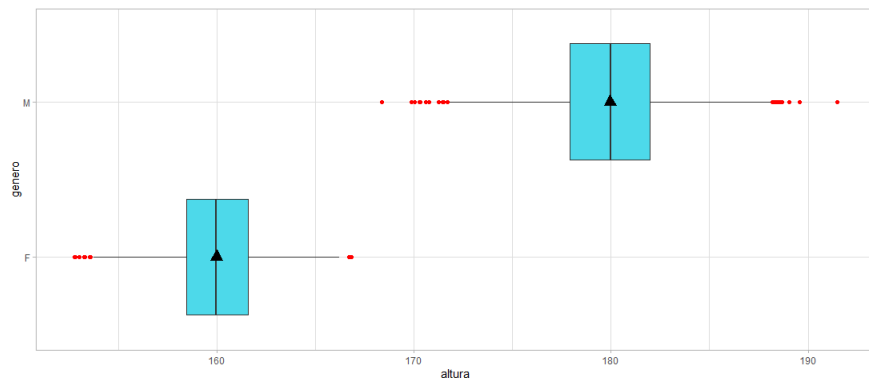


Figura 10: Cajas de altura

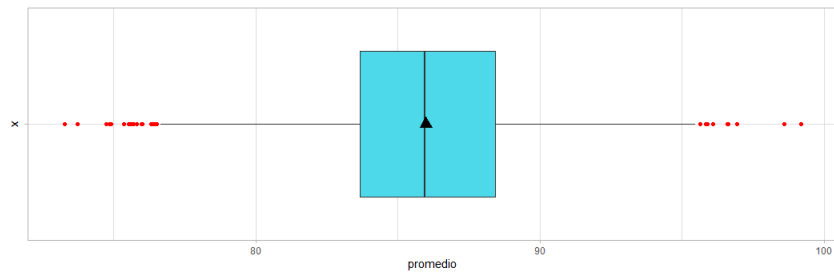


Figura 11: Caja de promedio

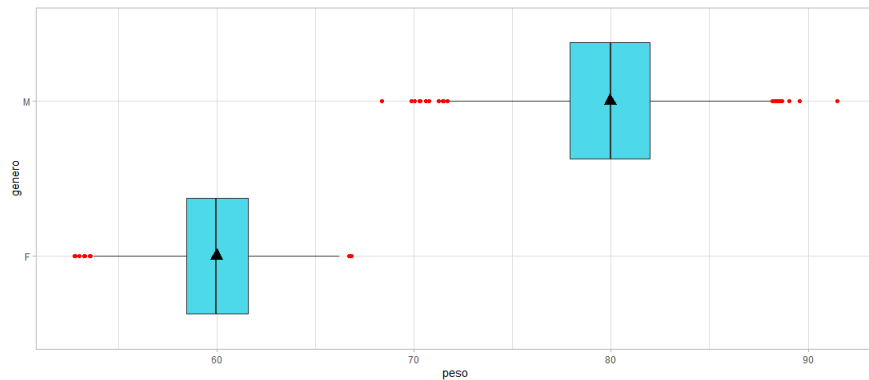


Figura 12: Cajas de peso

La caja de promedio no aporta información nueva. Sin embargo, se aclara el patrón de las variables fisiológicas: cada uno de los conjuntos correspondía perfectamente a la diferenciación de género, sin datos aislados cruzados. Esto tampoco suele suceder en la vida real: hombres y mujeres están mucho más cerca en términos de peso y altura.

7. Medidas de tendencia central

Todas las medidas se calcularon con funciones ya diseñadas.

Vale notar que las medidas fisiológicas son casi idénticas: la diferencia es de 100 en cada caso. Además, hay dos modas exactas en ambos géneros, con frecuencias iguales. Todo esto sugiere, de nuevo, que la base de datos es artificial.

	Variables	Promedio	Mediana	Moda
1	promedio	85.99	85.95	83.73
2	edad	21.00	21.00	21
3	peso	70.24	73.69	M: 60.41 H: 81.22
4	altura	170.24	173.69	M: 160.41 H: 181.22

8. Medidas de variabilidad

Todas las medidas se calcularon con funciones ya diseñadas.

	Variables	Rango	Desviación	Varianza	CVariacion
1	promedio	25.92	3.56	12.65	0.04
2	edad	8.00	1.06	1.13	0.05
3	peso	38.69	10.36	107.27	0.15
4	altura	38.69	10.36	107.27	0.06

Se evidencia, de nuevo, otro patrón curioso: el rango, la desviación y la varianza son iguales en las variables fisiológicas. Esto definitivamente no ocurre en la vida real, lo que refuerza la tesis de que los datos fueron generados artificialmente

9. Medidas de posición relativa

Se dividió el conjunto de datos en 5 partes.

	Variables	.25	.75	.85	.92
1	promedio	83.68	88.42	89.71	90.97
2	edad	20.00	22.00	22.00	22.00
3	peso	60.05	80.10	81.65	83.14
4	altura	160.05	180.10	181.65	183.14

- **Promedio:** Comportamiento creciente.
- **Edad:** Sólo crece al inicio.
- **Peso:** Comportamiento creciente.
- **Altura:** Comportamiento creciente.

10. Asimetría y apuntamiento

Sabiendo, a través de las gráficas 7, 9, 10 y 12, que los datos fisiológicos tienen dos patrones diferentes (algo así como dos histogramas en uno), hay que dividir las medidas en género:

	Variables	Asimetría	Apuntamiento
1	promedio	-0.04	-0.04
2	edad	-0.07	-6e-5

	Variables	Asim M	Asim F	Apun M	Apun F
3	peso	-0.04	-0.06	0.04	-0.19
4	altura	-0.04	-0.06	0.04	-0.19

10.1. Análisis de cada caso:

- **Promedio:** Con una asimetría y apuntamiento negativos, es una gráfica ligeramente platicúrtica a la izquierda.
 - **Edad:** Con una asimetría negativa, la gráfica tiende ligeramente a la izquierda. El apuntamiento, aunque negativo, es muy pequeño: se aproxima más a ser mesocúrtica.
 - **Peso:**
 - *Mujeres:* Ligeramente platicúrtica a la izquierda.
 - *Hombres:* Mesocúrtica (con pequeña tendencia a leptocurtismo) a la derecha.
 - **Altura:** Exactamente igual al peso.
-