Modelos Estadísticos para la Regresión y la Clasificación

Práctico 1 - Introducción

Micaela Long

Instituto de Matemática y Estadística Prof. Rafael Laguardia (IMERL)
Facultad de Ingeniería, Universidad de la República, Uruguay

9 de agosto de 2024

Clases de práctico

Día y horario

- Viernes 8 a 10 am

Estructura de las clases

- Repaso teórico
- Práctico: ejercicios y/o laboratorios
- Espacio de consultas

Lenguajes de programación

- R
  RStudio
- Python
  Jupyter notebooks
  - Anaconda (local)
  - **Google Colaboratory (colab)**

Por cualquier consulta escribir en Foro de EVA o a: mlong@fing.edu.uy.

Práctico 1
Estadística Descriptiva

Material para hacer práctico 1 (disponible en EVA):

- Teórico 7/8
- Notas PyE de Estadística Descriptiva
- Notas PyE de Test χ^2

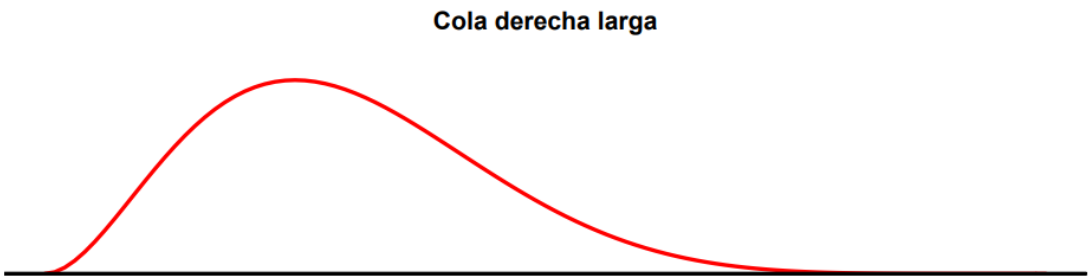
Ejercicio 1: Conclusiones a partir de un resumen numérico.

A continuación se muestran indicadores que caracterizan la distribución de notas de dos clases paralelas de un curso de Inglés. El puntaje máximo es 100.

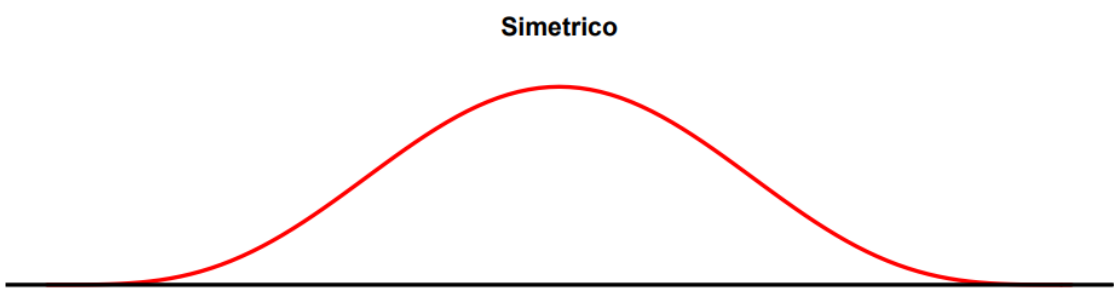
	Clase 1	**Clase 2**
Promedio	78	72
Mediana	65	73
Desvío estándar	16	6

1. Bosquejar el histograma de la distribución de notas de cada clase.
2. ¿En cuál de las dos clases es más probable encontrar un estudiante talentoso?

Ejercicio 1
Parte 1



promedio mayor a la mediana



promedio aproximadamente igual a la mediana

promedio menor a la mediana

Ejercicio 1
Parte 1

Sugerencia para pensar en el histograma de la Clase 1:

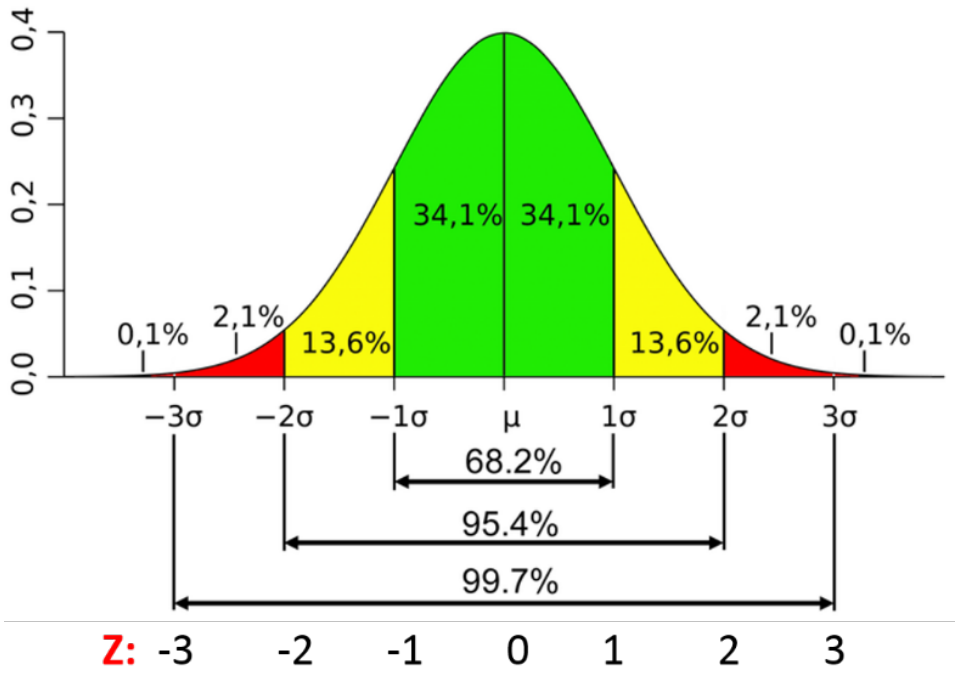
1. Escribir en Excel, Python o R una lista de números. Empezar incluyendo el 65.
2. Agregar la misma cantidad de números a la izquierda (menores) y a la derecha (mayores) de 65.
3. Calcular promedio y desviación estándar.
4. Modificar la lista hasta que el promedio sea aproximadamente 78 y la desviación estándar 16.
5. Graficar histograma!

Ver un ejemplo en Laboratorio Práctico 1:

`https://colab.research.google.com/drive/1CXJLWXACUzWc_TXvf_gpvvB4VdqUnesd`

Ejercicio 1
Parte 2

Cuál es la probabilidad de encontrar una nota mayor a 90 en la Clase 2?



- En la clase 2 tenemos $3\sigma = 18$.
- Por tanto, aproximadamente el $99,7\,\%$ de los individuos tendrán notas en el intervalo $[\bar{x} - 3\sigma, \bar{x} + 3\sigma] = [54, 90]$.
- Pensar qué sucede con la clase 1.

Ejercicio 2: Presión arterial diastólica.

Las siguientes mediciones corresponden a la presión arterial diastólica para una muestra de 15 adolescentes:

60 52 75 54 85 45 76 64 58 71 65 60 55 63 70

1. Construir un diagrama de tallos y hojas para estas mediciones, indicando la profundidad de cada medición.
2. Calcular el resumen de cinco números.
3. Hacer un diagrama de caja (boxplot).
4. Calcular la presión arterial diastólica promedio y el desvío estándar de la muestra.

Ejercicio 2
Parte 1

60 52 75 54 85 45 76 64 58 71 65 60 55 63 70

Diagrama de tallos y hojas.

- Estrategia para ordenar los datos y estudiar su distribución.
- Los tallos se ordenan verticalmente, mientras que las hojas se marcan horizontalmente en el valor del tallo correspondiente.

4	5
5	2 4 5 8
6	0 0 3 4 5
7	0 1 5 6
8	5

4	*
5	* * * *
6	* * * * *
7	* * * *
8	*

- Profundidad $p(\cdot)$ de una medición nos dice qué tan lejos está de la medición más chica.
- Ej.: $p(55) = 4$.

Ejercicio 2
Partes 2, 3 y 4

- Calcular el resumen de cinco números.
- Hacer un diagrama de caja (boxplot).
- Calcular la presión arterial diastólica promedio y el desvío estándar de la muestra.

Lo podemos hacer a mano, pero veamos como hacerlo en Python:

`https://colab.research.google.com/drive/1CXJLWXACUzWc_TXvf_gpvvB4VdqUnesd`

Ejercicio 3: Resumen numérico.
Resumen de los cinco números y diagrama de tallos y hojas.

Ejercicio 4: Alturas.
Promedio.
Sugerencia: usar la definición de media muestral y plantear el problema en una ecuación.

Ejercicio 5: Estudio farmacéutico.
Diagrama de tallos y hojas extendido.
Histogramas.

Ejercicio 6: Tablas de contingencia I

En un estudio se investiga la relación entre hacer ejercicio frecuentemente y fumar. Para una muestra de 200 individuos, los resultados son los siguientes:

Ejercicio frecuente	**Fumador**	**No fumador**	**Total**
Sí	37	53	90
No	63	47	110
Total	100	100	200

Realizar el test χ^2 de independencia. Indicar en qué intervalo se encuentra el p-valor.

Test χ^2 de independencia

Objetivo:

Determinar si existe una relación significativa entre dos variables cualitativas.

Procedimiento:

Comparar las cantidades observadas en una tabla de contingencia, con las cantidades esperadas si las variables fueran independientes.

Test χ^2 de independencia

1. **Formulación de Hipótesis.**
   - Hipótesis nula (H_0): Las dos variables son independientes.
   - Hipótesis alternativa (H_1): Las dos variables no son independientes.
2. **Construcción de la tabla de contingencia.**
3. **Cálculo de las cantidades esperadas.**

   Se calculan bajo la suposición de que las variables son independientes:

$$E_{ij} = \frac{(\text{Total de la fila i}) x (\text{Total de la columna j})}{(\text{Total de muestras})}$$

4. **Cálculo del estadístico χ^2**

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

5. **Determinación del p-valor**

   Se obtiene comparando el estadístico χ^2 con la distribución χ^2 con $(r-1)x(c-1)$ grados de libertad, donde r es el número de filas y c es el número de columnas en la tabla de contingencia.
6. **Conclusión**
   - Se elige nivel de significancia α (en general 0,05).
   - Si el p-valor es menor que α, se rechaza la hipótesis nula (variables no independientes).
   - Si el p-valor es mayor que α, no se rechaza la hipótesis nula (no hay evidencia suficiente).

Ejercicio 6

Paso 1: Formulación de hipótesis.

- Hipótesis nula (H_0): Hacer ejercicio frecuentemente y fumar son independientes.
- Hipótesis alternativa (H_1): Hacer ejercicio frecuentemente y fumar no son independientes.

Paso 2: Construcción de la tabla de contingencia.

Ejercicio frecuente	**Fumador**	**No fumador**	**Total**
Sí	37	53	90
No	63	47	110
Total	100	100	200

Paso 3: Cálculo de las cantidades esperadas.

$$E_{ij} = \frac{\text{(Total de la fila i)}x\text{(Total de la columna j)}}{\text{(Total de muestras)}}$$

$$\begin{aligned} E_{11} &= 200 \cdot \mathbb{P}(\{\text{Fumador}\} \cap \{\text{Ejercicio frecuente}\}) \\ &= 200 \cdot \mathbb{P}(\text{Fumador}) \cdot \mathbb{P}(\text{Ejercicio frecuente}) \\ &= 200 \cdot \frac{90}{200} \cdot \frac{100}{200} \\ &= 45 \end{aligned}$$

$$E_{21} = \frac{100x110}{200} = 55$$

$$E_{12} = \frac{90x100}{200} = 45$$

$$E_{22} = \frac{100x110}{200} = 55$$

Paso 4: Cálculo del estadístico χ^2

Ejercicio frecuente	Fumador	No fumador	Total
Sí	37	53	90
No	63	47	110
Total	100	100	200

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(37-45)^2}{45} + \frac{(53-45)^2}{45} + \frac{(63-55)^2}{55} + \frac{(47-55)^2}{55}$$

$$\approx 5{,}172$$

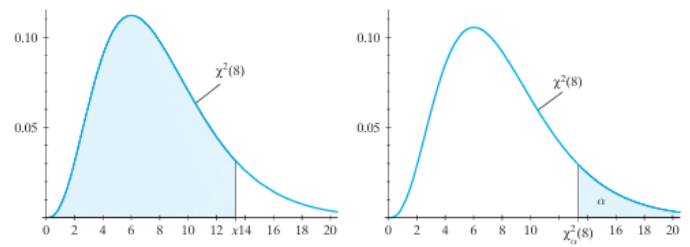
Paso 5: Determinación del p-valor.

Grados de libertad (tenemos dos filas y dos columnas): $(2-1)(2-1) = 1$.

Vamos a la tabla de χ^2 ...

Ejercicio 6

Table IV The Chi-Square Distribution



$$P(X \le x) = \int_0^x \frac{1}{\Gamma(r/2)2^{r/2}} w^{r/2-1} e^{-w/2} dw$$

	$P(X \le x)$								
	0.010	0.025	0.050	0.100	0.900	0.950	0.975		0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$		$\chi^2_{0.01}(r)$
1 grado de libertad -----> 1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	5.172	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378		9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348		11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14		13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83		15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45		16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01		18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54		20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02		21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48		23.21
11	3.053	3.816	4.575	5.578	17.28	19.68	21.92		24.72
12	3.571	4.404	5.226	6.304	18.55	21.03	23.34		26.22
13	4.107	5.009	5.892	7.042	19.81	22.36	24.74		27.69
14	4.660	5.629	6.571	7.790	21.06	23.68	26.12		29.14
15	5.229	6.262	7.261	8.547	22.31	25.00	27.49		30.58
16	5.812	6.908	7.962	9.312	23.54	26.30	28.84		32.00
17	6.408	7.564	8.672	10.08	24.77	27.59	30.19		33.41
18	7.015	8.231	9.390	10.86	25.99	28.87	31.53		34.80
19	7.633	8.907	10.12	11.65	27.20	30.14	32.85		36.19
20	8.260	9.591	10.85	12.44	28.41	31.41	34.17		37.57
21	8.897	10.28	11.59	13.24	29.62	32.67	35.48		38.93
22	9.542	10.98	12.34	14.04	30.81	33.92	36.78		40.29
23	10.20	11.69	13.09	14.85	32.01	35.17	38.08		41.64
24	10.86	12.40	13.85	15.66	33.20	36.42	39.36		42.98
25	11.52	13.12	14.61	16.47	34.38	37.65	40.65		44.31
26	12.20	13.84	15.38	17.29	35.56	38.88	41.92		45.64
27	12.88	14.57	16.15	18.11	36.74	40.11	43.19		46.96
28	13.56	15.31	16.93	18.94	37.92	41.34	44.46		48.28
29	14.26	16.05	17.71	19.77	39.09	42.56	45.72		49.59
30	14.95	16.79	18.49	20.60	40.26	43.77	46.98		50.89
40	22.16	24.43	26.51	29.05	51.80	55.76	59.34		63.69
50	29.71	32.36	34.76	37.69	63.17	67.50	71.42		76.15
60	37.48	40.48	43.19	46.46	74.40	79.08	83.30		88.38
70	45.44	48.76	51.74	55.33	85.53	90.53	95.02		100.4
80	53.34	57.15	60.39	64.28	96.58	101.9	106.6		112.3

This table is abridged and adapted from Table III in *Biometrika Tables for Statisticians*, edited by E.S.Pearson and H.O.Hartley.

$$\mathbb{P}(\chi^2 > 5{,}172) \in [0{,}01, 0{,}025]$$

Y por tanto

$$\mathbb{P}(\chi^2 > 5{,}172) < 0{,}05 = \alpha$$

¿Intuición?

A mayor χ^2, mayor es la diferencia entre las cantidades observadas y esperadas.

⬇

Menor es el área bajo la curva (p-valor), es decir, la probabilidad de observar un estadístico χ^2 tan extremo o más extremo que el calculado, bajo la hipótesis nula.

⬇

Mayor es la probabilidad de que la hipótesis alternativa sea cierta.

Ejercicio 6

Paso 6: Conclusión

Dado que el p-valor es menor que el nivel de significancia α, rechazamos la hipótesis nula.

Por tanto, hay evidencia significativa para afirmar que existe una relación entre hacer ejercicio frecuentemente y fumar.

Ejercicio 7: Tablas de contingencia II

Aplicar lo anterior en este ejercicio!