

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

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.

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.

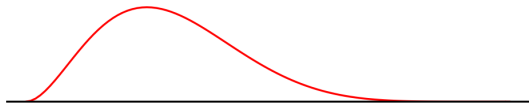
	Clase 1	Clase 2
Promedio	78	72
Mediana	65	73
Desvío estándar	16	6

- 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

Cola derecha larga



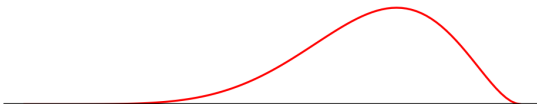
promedio mayor a la mediana

Simetrico



promedio aproximadamente igual a la mediana

Cola izquierda larga



promedio menor a la mediana

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!

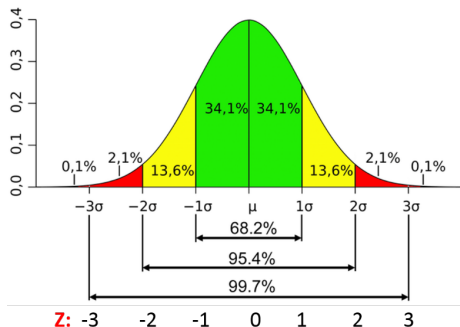
Ver un ejemplo en Laboratorio Práctico 1:

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.

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.

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:

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

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 $(r - 1) \times (c - 1)$ 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.

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

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

$$E_{21} = \frac{100 \times 110}{200} = 55$$

$$E_{12} = \frac{90 \times 100}{200} = 45$$

$$E_{22} = \frac{100 \times 110}{200} = 55$$

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

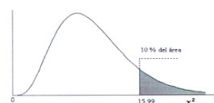
$$\begin{aligned}\chi^2 &= \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \frac{(37 - 45)^2}{45} + \frac{(53 - 45)^2}{45} + \frac{(63 - 55)^2}{55} + \frac{(47 - 55)^2}{55} \\ &\approx 5,172\end{aligned}$$

Paso 5: Determinación del p-valor.

Grados de libertad: $(2 - 1)(2 - 1) = 1$.

Vamos a la tabla de χ^2 ...

Test χ^2



Ejemplo:
Para $\phi = 10$ grados de libertad
 $P(\chi^2 > 15.99) = 0.10$

Grados de libertad

α	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005	α
1	3.841	3.841	3.841	3.841	3.841	3.841	3.841	3.841	3.841	3.841	3.841	3.841	3.841	1
2	5.991	5.991	5.991	5.991	5.991	5.991	5.991	5.991	5.991	5.991	5.991	5.991	5.991	2
3	7.879	7.879	7.879	7.879	7.879	7.879	7.879	7.879	7.879	7.879	7.879	7.879	7.879	3
4	9.488	9.488	9.488	9.488	9.488	9.488	9.488	9.488	9.488	9.488	9.488	9.488	9.488	4
5	11.070	11.070	11.070	11.070	11.070	11.070	11.070	11.070	11.070	11.070	11.070	11.070	11.070	5
6	12.592	12.592	12.592	12.592	12.592	12.592	12.592	12.592	12.592	12.592	12.592	12.592	12.592	6
7	14.067	14.067	14.067	14.067	14.067	14.067	14.067	14.067	14.067	14.067	14.067	14.067	14.067	7
8	15.508	15.508	15.508	15.508	15.508	15.508	15.508	15.508	15.508	15.508	15.508	15.508	15.508	8
9	16.919	16.919	16.919	16.919	16.919	16.919	16.919	16.919	16.919	16.919	16.919	16.919	16.919	9
10	18.307	18.307	18.307	18.307	18.307	18.307	18.307	18.307	18.307	18.307	18.307	18.307	18.307	10
11	19.675	19.675	19.675	19.675	19.675	19.675	19.675	19.675	19.675	19.675	19.675	19.675	19.675	11
12	21.026	21.026	21.026	21.026	21.026	21.026	21.026	21.026	21.026	21.026	21.026	21.026	21.026	12
13	22.364	22.364	22.364	22.364	22.364	22.364	22.364	22.364	22.364	22.364	22.364	22.364	22.364	13
14	23.685	23.685	23.685	23.685	23.685	23.685	23.685	23.685	23.685	23.685	23.685	23.685	23.685	14
15	25.000	25.000	25.000	25.000	25.000	25.000	25.000	25.000	25.000	25.000	25.000	25.000	25.000	15
16	26.296	26.296	26.296	26.296	26.296	26.296	26.296	26.296	26.296	26.296	26.296	26.296	26.296	16
17	27.577	27.577	27.577	27.577	27.577	27.577	27.577	27.577	27.577	27.577	27.577	27.577	27.577	17
18	28.845	28.845	28.845	28.845	28.845	28.845	28.845	28.845	28.845	28.845	28.845	28.845	28.845	18
19	30.104	30.104	30.104	30.104	30.104	30.104	30.104	30.104	30.104	30.104	30.104	30.104	30.104	19
20	31.353	31.353	31.353	31.353	31.353	31.353	31.353	31.353	31.353	31.353	31.353	31.353	31.353	20
21	32.671	32.671	32.671	32.671	32.671	32.671	32.671	32.671	32.671	32.671	32.671	32.671	32.671	21
22	33.924	33.924	33.924	33.924	33.924	33.924	33.924	33.924	33.924	33.924	33.924	33.924	33.924	22
23	35.178	35.178	35.178	35.178	35.178	35.178	35.178	35.178	35.178	35.178	35.178	35.178	35.178	23
24	36.421	36.421	36.421	36.421	36.421	36.421	36.421	36.421	36.421	36.421	36.421	36.421	36.421	24
25	37.652	37.652	37.652	37.652	37.652	37.652	37.652	37.652	37.652	37.652	37.652	37.652	37.652	25
26	38.885	38.885	38.885	38.885	38.885	38.885	38.885	38.885	38.885	38.885	38.885	38.885	38.885	26
27	40.115	40.115	40.115	40.115	40.115	40.115	40.115	40.115	40.115	40.115	40.115	40.115	40.115	27
28	41.337	41.337	41.337	41.337	41.337	41.337	41.337	41.337	41.337	41.337	41.337	41.337	41.337	28
29	42.557	42.557	42.557	42.557	42.557	42.557	42.557	42.557	42.557	42.557	42.557	42.557	42.557	29
30	43.778	43.778	43.778	43.778	43.778	43.778	43.778	43.778	43.778	43.778	43.778	43.778	43.778	30
40	55.758	55.758	55.758	55.758	55.758	55.758	55.758	55.758	55.758	55.758	55.758	55.758	55.758	40
50	67.565	67.565	67.565	67.565	67.565	67.565	67.565	67.565	67.565	67.565	67.565	67.565	67.565	50
60	79.320	79.320	79.320	79.320	79.320	79.320	79.320	79.320	79.320	79.320	79.320	79.320	79.320	60
70	91.164	91.164	91.164	91.164	91.164	91.164	91.164	91.164	91.164	91.164	91.164	91.164	91.164	70
80	103.001	103.001	103.001	103.001	103.001	103.001	103.001	103.001	103.001	103.001	103.001	103.001	103.001	80
90	114.839	114.839	114.839	114.839	114.839	114.839	114.839	114.839	114.839	114.839	114.839	114.839	114.839	90
100	126.675	126.675	126.675	126.675	126.675	126.675	126.675	126.675	126.675	126.675	126.675	126.675	126.675	100
Z _α	-2.58	-2.33	-1.96	-1.64	-1.28	-0.874	0.000	0.674	1.282	1.645	1.96	2.33	2.58	Z _α

$$\mathbb{P}(\chi^2 > 5,17) \in [0,01, 0,025]$$

Y por tanto

$$\mathbb{P}(\chi^2 > 5,17) < 0,05 = \alpha$$

¿Intuición?

A mayor χ^2 , mayor es la diferencia entre las probabilidades 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.

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!