- Métodos computacionales:
  Alejandro Segura

- **Estadística**

  a) Incluir el código Notebook (.ipynb).

  b) Guardar la información en una carpeta llamada Semana12_Nombre1_Nombre2

  c) **Hacer una sola entrega por grupo**.

Contents

List of Figures

1 Estadística

1.1 Tablas de Contingencia (Categorical variables)

1. La universidad de los Andes realiza una encuesta a $N = 100$ estudiantes sobre la intención de voto de tres candidatos a la presidencia, los candidatos son: A, B y C. Adicionalmente, se tuvo en cuenta el género de cada estudiante. Los resultados se pueden resumir en la siguiente Tabla:

	A	B	C
Mujeres	11	15	34
Hombre	23	6	11

Table 1: Tabla de contingencia asociada a la encuesta realizada por la universidad.

Los resultados mostrados en la Tabla [1] representan la frecuencias observadas (f_o) de la intención de voto de cada candidato. La universidad desea saber si las variables género y candidato son o no independientes con un nivel de confianza del 95%. Para tal propósito, planteamos las siguientes hipótesis:

a) H_0 : Las variables género y candidato son independientes. (Hipótesis nula)

b) H_1 : Las variables género y candidato no son independientes. (Hipótesis alternativa)

Note que la hipótesis alternativa es la negación de la hipótesis nula $H_1 = \neg H_0$. **Suponemos que la hipótesis nula es verdadera** para calcular las frecuencias esperadas (f_e). Entonces:

a) Si las variables son independientes entonces $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$. Entonces, deberá calcular las probabilidades marginales y luego calcular el valor de la probabilidad de cada entrada en la tabla. Usando el producto de Kronecker sería muy eficiente y claro.

b) La frecuencia esperada en cada posición es: $f_e^i = \mathbb{P}(A \cap B)^i \cdot N$, i.e, la probabilidad de cada evento multiplicado por el tamaño de muestra.

La matriz de probabilidad $\mathbb{P}(x, y) = \mathbb{G}(x) \otimes \mathbb{H}(y)$, puede ser escrita como el producto de Kronecker de los vectores de probabilidad marginal (sería interesante que implementaran esta operación en su código). Finalmente, para medir la discrepancia entre las frecuencias esperadas y observadas se tiene el estadístico de prueba (muy conocido) χ^2 dado por:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_o^i - f_e^i)^2}{f_e^i} \tag{1}$$

Este valor se compará con la distribución $\chi^2_{\alpha,(r-1)(c-1)}$. En esta definición, $\alpha = 0.05$ se refiere al nivel de significación de la prueba, r es el número de filas y c el número de columnas. $(r-1)(c-1)$ son los grados de libertad (conceptos por definir) asociados a nuestra prueba estadística; entonces para nuestro ejemplo, $\chi^2_{0.05,2} \approx 5.991$ nos definirá el valor crítico del estadístico a un nivel de confianza del 95%.

En `Python`:

```
from scipy.stats import chi2
chi2.ppf(0.95,df=2)
```

ppf devuelve el percentil asociado a un (p-value) $p = 0.05$, para una distribución χ^2 de dos grados de libertad. Calcule el $\chi^2 \approx 16.50$, que es un valor más extremo que el valor crítico, por lo cuál, debemos rechazar la hipótesis nula con un nivel de confianza del 95%. Las variables son dependientes estadísticamente.

Remark:

Efectivamente, como muestra la tabla, las mujeres prefieren el candidato C y los hombres el candidato A.