MÓDULO 2: MÉTODOS ESTADÍSTICOS PARA IA

OBJETIVO GENERAL: INTRODUCIR LOS FUNDAMENTOS ESTADÍSTICOS NECESARIOS PARA ANALIZAR DATOS Y EVALUAR MODELOS DE INTELIGENCIA ARTIFICIAL, FAVORECIENDO LA TOMA DE DECISIONES BASADAS EN EVIDENCIA.

- CLASE 1: NOCIONES DE PROBABILIDAD. ESTADÍSTICA DESCRIPTIVA.
- CLASE 2: TEST DE HIPÓTESIS.
- CLASE 3: REGRESIÓN.
- CLASE 4: ESTADÍSTICA NO PARAMÉTRICA.

MÓDULO 2: MÉTODOS ESTADÍSTICOS PARA IA

OBJETIVO GENERAL: INTRODUCIR LOS FUNDAMENTOS ESTADÍSTICOS NECESARIOS PARA ANALIZAR DATOS Y EVALUAR MODELOS DE INTELIGENCIA ARTIFICIAL, FAVORECIENDO LA TOMA DE DECISIONES BASADAS EN EVIDENCIA.

- CLASE 1: NOCIONES DE PROBABILIDAD. ESTADÍSTICA DESCRIPTIVA.
- CLASE 2: TEST DE HIPÓTESIS.
- CLASE 3: REGRESIÓN.
- CLASE 4: ESTADÍSTICA NO PARAMÉTRICA.

Clase 2 - Parte 3: Métodos Estadísticos para IA

Test de Hipótesis sobre la proporción.

Objetivos de esta presentación

► Si p_0 es la proporción de una población que cumple con cierto suceso (denominado éxito), estudiaremos la distribución que tiene una muestra de tamaño n. En particular, observaremos el caso en que n sea grande.

► Utilizando el item anterior construiremos un test de hipótesis sobre la proporción.

► Finalmente, comentaremos el test de diferencia de proporciones.

Resumen

- ✓ Distribución de la proporción para muestras grandes
- ✓ Test de hipótesis sobre una proporción
- ✓ Diferencia de proporciones

Resumen

- ✓ Distribución de la proporción para muestras grandes
- ✓ Test de hipótesis sobre una proporción
- ✓ Diferencia de proporciones

Ejemplo motivador

- Como ejemplo motivador, utilizaremos datos extraídos del INDEC[1]. En el INDEC se muestran datos de todo tipo, tomaremos una encuesta denominada: “Encuesta Nacional sobre Prevalencias de Consumo de Sustancias Psicoactivas”[2].

- Contaremos la **proporción** de personas en la región de Gran La Plata que beben al menos tres cervezas por día durante el fin de semana.

- Para saber el nombre de las variables se debe leer el instructivo detallado por INDEC (ya están seleccionados los datos en el archivo ejemplo15.ipynb).

[1]https://www.indec.gob.ar/

[2]https://www.indec.gob.ar/ftp/cuadros/menusuperior/enprecosp/bases_enprecosp2011.rar

Ejemplo 15: carga de datos

- Pasamos la base de datos a un archivo .csv y se pueden cargar con Python:

```
# Carga de datos
archivo = "BaseUsuarioENPreCoSP-2011.csv"
datos = pd.read_csv(archivo)

# Selección de datos de Gran La Plata
datos_granlaplata = datos[datos['AGL_URB'] == 6]

# Selección de personas que beben cerveza durante el fin de semana
cerveza_fin_semana = datos_granlaplata['BIBA09_01']
cerveza_fin_semana.replace({888: 0.0}, inplace=True)  # Reemplazar 888 (no bebe alcohol) con 0
```

Código/ejemplo_15.ipynb

- Note que reemplazamos el valor “888” por 0 cervezas porque, según el instructivo, no toma alcohol.
- Por otro lado, sacamos los valores “NA” porque suponemos que no se les preguntó esa parte de la encuesta.

```
# Eliminación de valores NA
cerveza_fin_semana_limpia = cerveza_fin_semana.dropna()
n = len(cerveza_fin_semana_limpia)
print("Número de datos sin NA:", n)
```

Ejemplo 15: cálculo de las proporciones

- Si ejecutan el script podrán ver que $n = 204$ (una muestra grande). Una vez que tenemos los datos podemos calcular la proporción de interés:

```
# Proporción de personas que toman al menos 3 cervezas por fin de semana
x = np.sum(cerveza_fin_semana_limpia > 2)
p_obs = x / n
```

Código/ejemplo_15.ipynb

- A este valor lo denominamos **proporción observada**:

$$\hat{p} = \frac{x}{n} = \frac{87}{204} \approx 0{,}43 \tag{1}$$

- donde x cuenta la cantidad de sucesos "éxito". En nuestro caso, contamos un éxito cuando la persona seleccionada toma al menos tres cervezas[3].

[3] Aunque no estemos de acuerdo con lo literalmente expresado en esta frase :(

Distribución Binomial

- Si consideramos la v.a. definida como la **cantidad de éxitos** X (ahora con mayúscula porque es una v.a.) cuando extraemos una muestra de tamaño n sobre una población y suponiendo que nuestro éxito tiene una proporción poblacional p_0 entonces X tiene distribución Binomial con parámetros n y p_0. Lo denotamos así $X \sim \mathrm{B}\left(n, p_0\right)$.
- En el ejemplo que estamos manejando $n = 204$, un número relativamente grande.
- Alternativamente podemos pensar a X como una suma de n variables aleatorias Y_i (debemos asegurarnos de que sean estadísticamente iid) donde cada $Y_i \sim \mathrm{B}\left(1, p_0\right)$ (experimento de Bernoulli). Y_i cuenta "1" para un éxito con probabilidad p_0 y "0" para fracaso con probabilidad $1 - p_0$. Entonces:

$$X = \sum_{i=1}^{n} Y_i \qquad (2)$$

Aproximación Normal a la Binomial

- La Ec. (2) es una m.a. por lo tanto se cumple el Teorema Central del Límite, entonces vale la aproximación:

$$X = \sum_{i=1}^{n} Y_i \sim \mathbf{N}\left(np_0, np_0\left(1 - p_0\right)\right) \quad (3)$$

- Note que:

$E\left\{Y_i\right\} = 0\left(1 - p_0\right) + 1p_0 = p_0 \Rightarrow E\left\{X\right\} = \sum_{i=1}^{n} E\left\{Y_i\right\} = np_0$

$V\left\{Y_i\right\} = 0^2\left(1 - p_0\right) + 1^2 p_0 - p_0^2 = p_0\left(1 - p_0\right) \Rightarrow V\left\{X\right\} = \sum_{i=1}^{n} V\left\{Y_i\right\} = np_0\left(1 - p_0\right)$

donde $E\left\{\cdot\right\}$ y $V\left\{\cdot\right\}$ denotan la esperanza y la varianza, respectivamente.

- Concluimos entonces que X tiene una **distribución aproximadamente Normal**. En la práctica esta aproximación vale cuando np_0 es grande y p_0 no está muy cerca de 0 o de 1.

Distribución del estimador de la proporción

- Podemos pensar la Ec. (1) como una estimación puntual del parámetro p_0. Ahora visto como una v.a.:

$$\hat{P} = \frac{X}{n} \tag{4}$$

que por lo mencionado antes tendrá una distribución aproximadamente Normal, pero cambiarán su media y varianza.

$$E\left\{\hat{P}\right\} = \frac{1}{n}E\left\{X\right\} = p_0$$

$$V\left\{\hat{P}\right\} = \frac{1}{n^2}V\left\{X\right\} = \frac{p_0\left(1-p_0\right)}{n}$$

Resumen

- ✓ Distribución de la proporción para muestras grandes
- ✓ Test de hipótesis sobre una proporción
- ✓ Diferencia de proporciones

Construcción del test de hipótesis

Continuando con el ejemplo15.ipynb:

- Una pregunta que se podría hacer es: ¿Hay evidencia suficiente para decir que más del 35% de la población en el Gran La Plata toma al menos 3 cervezas por día el fin de semana?

- Según lo que vimos en la *presentación 2.1*, una vez que ya se diseñó el experimento y se conoce la hipótesis nula (en nuestro ejemplo podría ser $H_0\colon p = p_0 = 0{,}35$), podemos seguir un procedimiento para el diseño del test de hipótesis:

1. Especificar el estadístico de prueba
2. Definir si es un test unilateral (una cola) o bilateral (dos colas)
3. Definir cuál será la significancia

Diseño del test de hipótesis

1. Aplicamos el estadístico Z utilizando el resultado de la Ec. (4).

$$Z_0 = \frac{\frac{X}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Si p_0 es verdadero sabemos que este estadístico tiene distribución $\mathbf{N}(0,1)$.

2. Definimos que es un test unilateral (una cola):

$$H_0: p = p_0 = 0,35$$

$$H_1: p > p_0$$

3. Podemos directamente concluir con el p-valor.

Computando el test en Python

- La probabilidad que debemos calcular para obtener el p-valor es:

$$\text{p−valor} = P\{Z_0 > z_0 | H_0 \text{ verdadero}\} = 1 - \Phi(z_0)$$

donde:

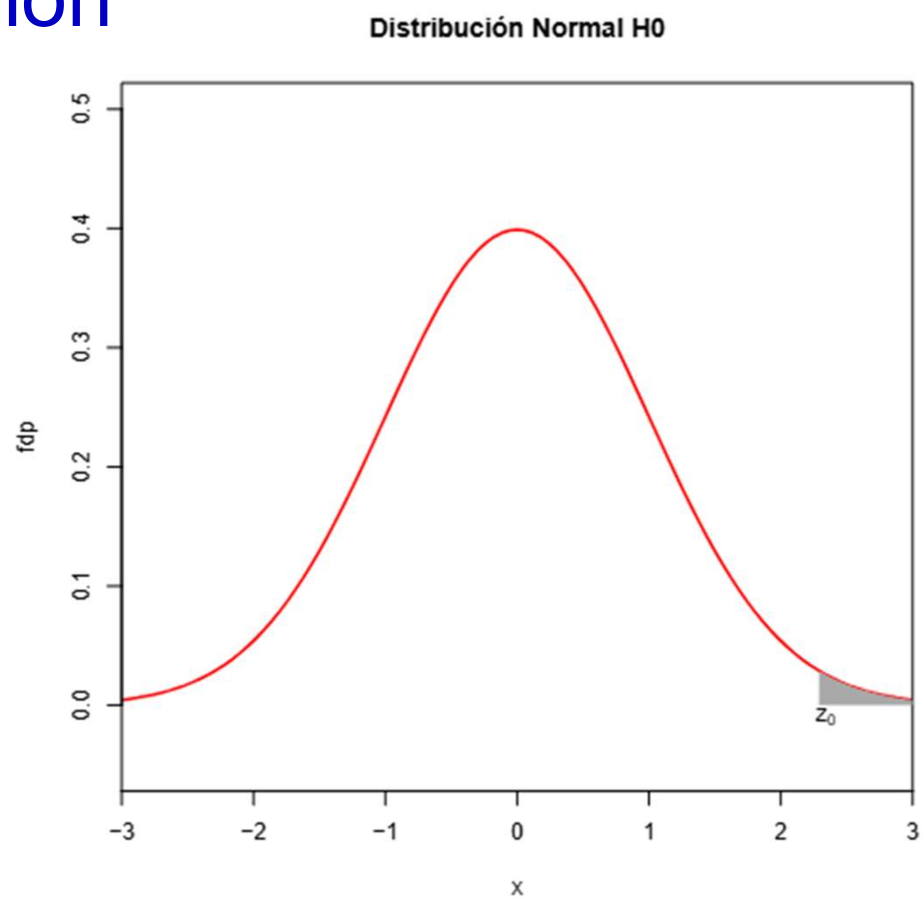
$$z_0 = \frac{\frac{x}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$ es la observación del estadístico.

```python
# Aproximación Normal a la Binomial
p0 = 0.35
sigma = np.sqrt(p0 * (1 - p0) / n)
z_0 = (p_obs - p0) / sigma
p_value = 1 - stats.norm.cdf(z_0)
```

Código/ejemplo_15.ipynb

- Esto da un p-valor de 0,011.

Computando el test en Python



► La figura muestra los datos estandarizados de la probabilidad que se calculó.

► Python también tiene una herramienta para calcular la probabilidad un poco más aproximada (da como resultado p-valor = 0,01418):

```
# Test de hipótesis binomial
p_value_bin = stats.binomtest(x, n=n, p=0.35, alternative='greater')
```

Código/ejemplo_15.ipynb

Conclusiones preliminares

- Respecto del ejemplo, debe remarcarse que el p-valor rechaza H_0 para una significancia de $0,05$. Es decir, existe evidencia para decir que la proporción de los que beben al menos 3 cervezas por día el fin de semana es superior al 35 %.
- En general, la aproximación Normal a la Binomial es interesante porque:
  - permite realizar un cómputo rápido.
  - es útil para calcular el tamaño de muestra para conseguir una potencia determinada (de manera similar a como hicimos en diapostiva 15 en la presentación 2.1). Para nuestro ejemplo (test de una cola):

$$n = \left[\frac{z_\alpha \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p(1-p)}}{p - p_0} \right]^2$$

    donde p es el valor "verdadero" (el valor de la H_1)[4].
  - Si tenemos pocos datos, tendremos que utilizar la distribución Binomial (en Python la función **binomtest** nos facilita el cálculo).

[4]En el ejemplo 15 para p = 0,44 el tamaño de muestra resultante fue n ≈ 250.

Test de hipótesis sobre una proporción

Aquí vamos a formalizar lo visto en esta sección:

Supongamos que consideramos una muestra aleatoria $(X_1, X_2, \ldots, X_n)$ de tamaño n, donde X_i tiene **distribución binomial** con parámetros 1 y p: $X_i \sim B(1, p)$.

Sabemos que $X = X_1 + X_2 + \ldots + X_n$ es una variable aleatoria con **distribución binomial** con parámetros n y p: $X \sim B(n, p)$.

Entonces, la variable aleatoria $\hat{P}$ definida por $\hat{P} = \frac{X}{n}$ representa la proporción de individuos de la muestra que verifican la propiedad de interés.

Además: $E(\hat{P}) = p_0$ y $V(\hat{P}) = \frac{p_0(1-p_0)}{n}$

Podemos tomar como **estadístico de prueba** a $Z_0 = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.

Test de hipótesis sobre una proporción

Caso 1: Supongamos que planteamos la **hipótesis nula** y la **hipótesis alternativa bilateral** siguientes:

$$H_0: p = p_0 \quad \text{contra} \quad H_1: p \neq p_0$$

donde p_0 es una constante específica.

La **regla de decisión** es:

- Rechazar H_0 si $|Z_0| > z_{\frac{\alpha}{2}}$
- Aceptar H_0 si $|Z_0| \leq z_{\frac{\alpha}{2}}$

La lógica sigue siendo la misma, si el estadístico de prueba toma un valor inusual, entonces se considera que hay evidencia en contra de H_0 y se rechaza la hipótesis nula. Si $np_0 > 10$ y $n(1-p_0) > 10$ por el **Teorema Central del Límite**, podemos usar la **aproximación normal a la binomial** (el estadístico $Z_0 = \frac{\hat{P}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx N(0,1)$).

Test de hipótesis sobre una proporción

Caso 2: Supongamos que planteamos la **hipótesis nula** y la **hipótesis alternativa unilateral** siguientes:

$$H_0\colon p = p_0 \quad \text{contra} \quad H_1\colon p > p_0$$

donde p_0 es una constante específica.

La **regla de decisión** es:

- Rechazar H_0 si $Z_0 > z_\alpha$
- Aceptar H_0 si $Z_0 \leq z_\alpha$

Test de hipótesis sobre una proporción

Caso 3: Supongamos que planteamos la **hipótesis nula** y la **hipótesis alternativa unilateral** siguientes:

$$H_0: p = p_0 \quad \text{contra} \quad H_1: p < p_0$$

donde p_0 es una constante específica.

La **regla de decisión** es:

- Rechazar H_0 si $Z_0 < -z_\alpha$
- Aceptar H_0 si $Z_0 \geq -z_\alpha$

Test de hipótesis sobre una proporción

Cálculo del p-valor:

Caso 1: Si las hipótesis son $H_0: p = p_0$ contra $H_1: p \neq p_0$ entonces:

$$\boxed{\boldsymbol{p\text{−}valor = 2[1 - \Phi(|Z_0|)]}}$$

Caso 2: Si las hipótesis son $H_0: p = p_0$ contra $H_1: p > p_0$ entonces:

$$\boxed{\boldsymbol{p\text{−}valor = 1 - \Phi(Z_0)}}$$

Caso 3: Si las hipótesis son $H_0: p = p_0$ contra $H_1: p < p_0$ entonces:

$$\boxed{\boldsymbol{p\text{−}valor = \Phi(Z_0)}}$$

Expresiones similares al Test de hipótesis para la media con varianza conocida

Resumen

- ✓ Distribución de la proporción para muestras grandes
- ✓ Test de hipótesis sobre una proporción
- ✓ Diferencia de proporciones

Diferencia de proporciones de muestras grandes

El interés en este problema es comparar proporciones (p_1 y p_2) en dos poblaciones. Podremos utilizar lo visto hasta ahora.

- Como la idea general es la misma que lo desarrollado anteriormente aquí presentaremos el Estadístico de prueba y solo comentaremos el caso de **muestras grandes**.
- Suponiendo que tenemos dos muestras sobre dos poblaciones X_1 y X_2 de tamaños n_1 y n_2, respectivamente, que representan el número de veces que ocurre el suceso “éxito”. Entonces las estimaciones de las proporciones son: $P_1 = X_1/n_1$ y $P_2 = X_2/n_2$ y tendrán aproximadamente distribución Normal, por lo antes visto.

Estadístico de prueba

- Si el test de hipótesis es:

$$H_0 : p_1 = p_2 = p \text{ Hipótesis nula}$$
$$H_1 : p_1 \neq p_2 \text{ Hipótesis alternativa} \tag{6}$$

entonces el estadístico de prueba será:

$$Z_0 = \frac{(P_1 - P_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \tag{7}$$

como p no la conozco, entonces si la hipótesis nula es verdadera un estimador del denominador de la Ec. (7) es:

$\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, donde $P = (X_1 + X_2) / (n_1 + n_2)$.

- Si H_0 es veradera, entonces el estadístico tendrá distribución **aproximadamente Normal** $(\mathrm{N}(0,1))$.

Bibliografía

- D. C. Montgomery and G. C. Runger, “Applied Statistics and Probability for Engineers Third Edition”, John Wiley & Sons, Inc.

- Peter Dalgaard, “Introductory Statistics with R”, Second Edition, Springer.