**INFERENCIAL (I): POBLACION, MUESTRA Y SESGO**

**Población y muestra**

Empecemos por unos conceptos básicos que ya conocíamos:

Texto

Descripción generada automáticamente con confianza baja

Lo que suele ocurrir es que no tenemos acceso a toda la población, bien porque es muy grande o bien porque en realidad queremos considerar no sólo los elementos actuales de la población sino los futuros también para poder hacer predicciones sobre ellos (por ejemplo, si quiero estimar el resultado de los partidos de fútbol de la semana que viene o el impacto de un medicamento en los futuros pacientes).

En definitiva, a partir de ahora trabajaremos casi siempre con datos que consideraremos muestras de una población más amplia.

Para ilustrar estos conceptos y todos los de la unidad vamos a utilizar dos datasets:

**import** matplotlib.pyplot **as** plt

**import** pandas **as** pd

df\_titanic **=** pd.read\_csv("./data/titanic.csv")

df\_life\_satisfaction **=** pd.read\_csv("./data/lifesat\_full.csv")

df\_life\_satisfaction.head()

Tabla

Descripción generada automáticamente con confianza baja

len(df\_titanic), len(df\_life\_satisfaction)

(891, 36)

Realmente los dos son muestras de una población mayor, el dataset de Titanic tiene 891 registros de los dos mil y picos posibles (la población completa de todos los pasajeros que viajaron en él) y el dataset de nivel de satisfacción vital contiene datos de 36 países de los más de 190 países reconocidos en el mundo. En ese sentido ambos datasets son muestras de poblaciones mayores.

**Lo importante de una muesta es que sea "REPRESENTATIVA" de la población de la que ha sido extraída**. Esto quiere decir que lo que podamos estudiar y deducir de los datos de la muestra podamos aplicarlo al resto de la población. De esa manera podremos hacer estimaciones, deducciones e inferencias sobre la población en general sin necesidad de tener todos los datos o estudiar la población completa.

**Muestra aleatoria**

En general, cuando queremos estudiar una población de la que no podemos obtener todos los datos (por los motivos que sean) acudimos a obtener una muestra de la misma. La forma de obtener la muestra impacta en la representatividad de la misma y por lo tanto en la validez de las inferencias que hagamos de lo analizado sobre dicha muestra.

Para mantener en la muestra las propiedades de la población, es decir su representatividad, intentaremos obtener una muestra aleatoria (más adelante intentaremos que además sea estratificada, pero por ahora sólo que te suene). Es decir, escogiendo al azar los elementos que van a formar parte de nuestra muestra. Al proceso de obtención de una o varias muestras aleatorias se le denomina **muestreo aleatorio**.

Diagrama

Descripción generada automáticamente

Como muestra la figura es como hacer un sorteo y escoger a los ganadores para la muestra. Vamos a verlo con nuestros datasets, suponiendo que estos sean TODA la población y no muestras. Obtengamos una muestra de cada uno de ellos de forma aleatoria aprovechando el método sample de pandas:

num\_muestras\_titanic **=** int(len(df\_titanic) **\*** 0.2)

muestra\_titanic **=** df\_titanic.sample(num\_muestras\_titanic, random\_state **=** 42)

print(f"Poblacion titanic: {len(df\_titanic)} pasajeros")

print(f"Muestra titanic: {len(muestra\_titanic)} pasajeros")

num\_muestras\_paises **=** int(len(df\_life\_satisfaction) **\*** 0.2)

muestra\_paises **=** df\_life\_satisfaction.sample(num\_muestras\_paises, random\_state **=** 42)

print(f"Poblacion Indice Satifaccion: {len(df\_life\_satisfaction)} países")

print(f"Muestra Indice Satisfaccion: {len(muestra\_paises)} países")

Poblacion titanic: 891 pasajeros

Muestra titanic: 178 pasajeros

Poblacion Indice Satifaccion: 36 países

Muestra Indice Satisfaccion: 7 países

En este ejemplo hemos obtenido dos muestras aleatorias (escogidas al azar) del 20% del tamaño de la población. ¿Y el tamaño importa? Sí, pero no asegura la representatividad aunque la mejore.

Viendo los números de nuestro muestreo, ¿serán representativas nuestras muestras? Para poder contestar ahondemos en eso de la representatividad y en dos conceptos "nuevos": parámetros y estadísticos.

**Representatividad, parámetros y estadísticos**

Nuestras muestras aleatorias serían representativas si sus características son muy muy parecidas a las de la población a la que "representan". Es decir, si las medidas que obtengamos sobre las muestras son parecidas a las que obtendríamos en la población.

¿Y qué medidas? Pues las que vimos al tratar la estadística descriptiva y el análisis de datos en el módulo anterior. Sí, las medidas univariantes (frecuencias, medias, desviaciones), bivariantes, etc de una muestra deberían parecerse a las de la población para que podamos decir que esta es representativa.

Por ejemplo, comparemos el ratio de supervivencia en la muestra aleatoria y en la "población" total del dataset del titanic:

[6]:

fig,axs **=** plt.subplots(1,2,figsize **=** (10,8))

freq **=** df\_titanic["alive"].value\_counts(normalize **=** **True**)

axs[0].bar(x **=** freq.index, height **=** freq.values, label **=** "Poblacion")

axs[0].set\_title("superviviencia poblacion")

axs[0].set\_ylim(0,0.7)

freq\_m **=** muestra\_titanic["alive"].value\_counts(normalize **=** **True**)

axs[1].bar(x **=** freq\_m.index, height **=** freq\_m.values, label **=** "Muestra")

axs[1].set\_title("supervivencia muestra")

axs[1].set\_ylim(0,0.7)

[6]:

(0.0, 0.7)

Gráfico, Gráfico de barras

Descripción generada automáticamente

Se parecen pero hmmm, no del todo. Prueba tú a aumentar o disminuir el número de pasajeros en la muestra a ver qué ocurre.

Ahora comprobemos las medidas de tendencia central, posición y dispersión de la edad y el precio del ticket, por comparar variables numéricas:

[7]:

df\_titanic\_medidas **=** df\_titanic[["age","fare"]].describe()

df\_muestra\_medidas **=** muestra\_titanic[["age","fare"]].describe()

df\_titanic\_medidas.merge(df\_muestra\_medidas, left\_index **=** **True**, right\_index**=** **True**, suffixes**=** ["\_pob","\_muestra"])

Tabla

Descripción generada automáticamente

Aquí hay más parecido "visual", pero todavía no podemos asegurarnos... Y eso que aquí podemos comparar muestra y población... Pero, lo normal no será tener la población (en cuyo caso no necesitamos la muestra), para esas situaciones para saber si las medidas son representativas acudiremos a los test de hipótesis. Y lo veremos en las siguientes unidades. Antes apliquemos esto mismo al dataset de satisfacción vital y hablemos del sesgo.

[8]:

df\_lifesat\_medidas **=** df\_life\_satisfaction.describe()

df\_muestra\_medidas **=** muestra\_paises.describe()

df\_lifesat\_medidas.merge(df\_muestra\_medidas, left\_index**=** **True**, right\_index**=** **True**, suffixes **=** ["\_pob","\_muestra"]).loc["mean"].T

[8]:

GDP per capita (USD)\_pob 42620.982007

Life satisfaction\_pob 6.599421

Life expectancy\_pob 79.455556

Homicide rate\_pob 3.741667

Employment rate\_pob 68.250000

GDP per capita (USD)\_muestra 56824.405352

Life satisfaction\_muestra 7.072510

Life expectancy\_muestra 81.600000

Homicide rate\_muestra 1.557143

Employment rate\_muestra 69.857143

Name: mean, dtype: float64

Este es incluso más difícil, porque hay medias que se parecen y otras no mucho... Prueba de nuevo a aumentar los valores de muestras.

Cuando las medidas hacen referencia a la población (nuetros "\_pob" en los ejemplos anteriores) se les llama **parámetros**. Cuando son las medidas hechas sobre las muestras, se les denomina **estadísticos**. Además para distinguirlos se les designa por letras diferentes:

Tabla

Descripción generada automáticamente

NOTA: La mediana de una muestra se representa por �~ y la proporción en la población por �, en cuyo caso, la proporción en la muestra se representa por �^

**Sesgos**

Se trata de una **distorsión no aleatoria que introducimos en la muestra como consecuencia de la selección de dicha muestra**. Si existe sesgo y no se controla, entonces, en general, haremos inferencias y deducciones erróneas a partir de la muestra.

Por ejemplo, si realizamos una encuesta sobre cómo va el país, no será lo mismo si todos los encuestados son de una clase social determinada, o de un partido político concreto.

Para verlo, supongamos que escogemos nuestra muestra del titanic pero sin darnos cuenta lo hacemos únicamente sobre pasajeros de Cherbourgh, por ejemplo porque sólo teníamos datos de esa ciudad:

[9]:

muestra\_sesgada **=** df\_titanic.loc[df\_titanic.embark\_town **==** "Cherbourg"]

Si consideramos los estadísticos de nuestra como parámetros, veamos que error habríamos cometido, por ejemplo con el ratio de supervivencia:

[10]:

fig,axs **=** plt.subplots(1,2,figsize **=** (10,8))

freq **=** df\_titanic["alive"].value\_counts(normalize **=** **True**)

axs[0].bar(x **=** freq.index, height **=** freq.values, label **=** "Poblacion")

axs[0].set\_title("superviviencia poblacion")

axs[0].set\_ylim(0,0.7)

freq\_m **=** muestra\_sesgada["alive"].value\_counts(normalize **=** **True**)

axs[1].bar(x **=** freq\_m.index, height **=** freq\_m.values, label **=** "Muestra")

axs[1].set\_title("supervivencia muestra")

axs[1].set\_ylim(0,0.7)

[10]:

(0.0, 0.7)

Gráfico, Gráfico de barras

Descripción generada automáticamente

Como ya sabíamos hay una gran diferencia, tanto como para convertir al Titanic de una catastrofe a un accidente grave pero no tanto...

Por eso a la hora de seleccionar nuestra muestra debemos controlar mucho los posibles sesgos. Las implicaciones pueden ser mucho más impactantes incluso...

**Otro ejemplo de sesgo: El sesgo de supervivencia**

**El sesgo de supervivencia** (nada que ver con la supervivencia del Titanic) es un sesgo muy difícil de controlar ya que es el sesgo que se produce por tratar con los datos que tenemos y no con todos los posibles, porque son los datos que nos han llegado (que han sobrevivido). Dos ejemplos característicos son el estudio de los castillos y el de los aviones británicos en la segunda guerra mundial:

Imagen que contiene foto, diferente, computadora, tablero

Descripción generada automáticamente

**Estadística Inferencial**

Para terminar recordar qué es lo que buscamos con la estadística inferencial frente a la estadística descriptiva:

Word

Descripción generada automáticamente con confianza baja