**PROBABILIDAD**

**Conceptos básicos y cálculo frecuencial**

Vamos a centrarnos en los datos del Titanic, y a emplear nuestras bibliotecas favoritas:

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

**import** numpy **as** np

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

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

En esta sesión vamos a empezar a engarzar ya la estadística inferencial y nuestro próximo objetivo final: tomar decisiones a partir de los datos para situaciones "nuevas" no necesariamente contempladas anteriormente.

Vamos a querer saber no ya cuánta gente murió en el Titanic o su ratio de supervivencia, sino si de repente nos dicen que hubo un pasajero X no contemplado en nuestros datos, ¿podríamos decir si murió o no?¿De qué puerto salió?¿Viajaba en primera o en tercera? Claro para hechos pasados parece menos interesante, salvo que tu campo sea precisamente el estudio del pasado.

Pero, si has hecho el EDA sobre el precio de las casas en los últimos 10 años, lo que te interesará es poder usar esos datos para poder decidir dónde comprar en los próximos tres años, por ejemplo. ¿Y cómo se responde a eso? De forma definitiva e inequívoca es imposible, pero a partir de los datos que has usado puedes inferir (si con estadística inferencial) una serie de características de lo que serán esas casas en los próximos tres años, cuánto son de fiables tus datos y tener una idea de que zonas tienen mejores posibilidades de ser buena inversión y cuáles no. Es decir, **estimar la probabilidad** para cada zona de que sean buenas o no o cuánto de buenas para la inversión y sobre eso tomar una decisión. Y ahí es dónde entra la probabilidad como herramienta para tomar decisiones antes situaciones cuya complejidad nos impide utilizar fórmulas matemáticas precisas.

**Preguntas guía**

Para guiarnos en esta sesión, vamos a plantearos algunas preguntas a resolver con nuestros datos:

1. ¿Cuál es la probabilidad de que una persona elegida al azar sobreviviese en el Titanic?
2. ¿Cuál es la probabilidad de que una persona elegida al azar se embarcase en Queenstown?
3. ¿Cuál es la probabilidad de que una persona elegida al azar fuese hombre o viajese en primera clase?
4. ¿Cuál es la probabilidad de que una persona elegida al azar viajese en primera clase y viajase sola?
5. ¿Cuál es la probabilidad de que una persona elegida al azar se salvase y viajase en tercera clase?

Lo primero será definir o formalizar el concepto de probabilidad con el que hemos jugado sin profundizar.

**¿Qué es la probabilidad?**

*Formalmente*: **La**[**probabilidad**](https://es.wikipedia.org/wiki/Probabilidad)**mide la mayor o menor posibilidad de que se dé un determinado resultado (suceso o evento) cuando se realiza un experimento aleatorio.**

Como experimentos aleatorios nos entran en la cabeza fácilmente la lotería o tirar un dado de varias caras. Pero un experimento aleatorio también es escoger una persona al azar de entre los pasajeros del titanic (al azar == no intencionadamente esa persona en concreto) y también (aunque ahora parezca contradictorio), escoger una persona al azar aunque con características determinadas [escoger una pasajero del Titanic de entre los que embarcaron en Southampton y tenían más de 20 años, siempre que una vez seleccionados los pasajeros con esas características escogieramos "sin mirar al pasajero en cuestión].

En este sentido nuestras preguntas entran dentro de lo que denominaríamos un experimento aleatorio.

Para **calcular la probabilidad** de un evento se toma en cuenta **todos los casos posibles** de ocurrencia del mismo; es decir, de cuántas formas puede ocurrir determinada situación. Los **casos favorables** de ocurrencia de un evento serán los que cumplan con la condición que estamos buscando. La [*probabilidad*](https://es.wikipedia.org/wiki/Probabilidad) toma valores entre 0 y 1 (o expresados en tanto por ciento, entre 0% y 100%). Esto es lo que se denomina un cálculo frecuencial y se conoce como **regla de Laplace**:

Texto

Descripción generada automáticamente con confianza media

Con esto ya estamos en disposición de poder dar contestación a nuestras dos primeras preguntas:

Para la primera pregunta buscamos el número de casos favorables (pasajeros que sobrevivieron) frente al total de pasajeros:

casos\_favorables **=**len(df\_titanic.loc[df\_titanic["alive"] **==** "yes"]) *# Casos favorables: el número de personas que sobrevivieron*

casos\_totales **=** len(df\_titanic)

print("p(sobrevivir de una persona elegida al azar) = ", casos\_favorables**/**casos\_totales)

p(sobrevivir de una persona elegida al azar) = 0.3838383838383838

Y también, aquí es donde tenemos que tener en cuenta nuestros intervalos de confianza y demás, porque debemos partir del caso de que nosotros trabajamos con los datos muestrales, con lo que nuestras probabilidades obtenidas con muestras deberían ser consideradas dentro de un rango (pero esto no lo vamos a respetar, life is hard...)

La segunda pregunta se resolvería de forma análoga:

casos\_favorables **=**len(df\_titanic.loc[df\_titanic["embark\_town"] **==** "Queenstown"]) *# Casos favorables: el número de personas que embarcaron en Queenstown*

casos\_totales **=** len(df\_titanic)

print("p(Queenstown de una persona elegida al azar) = ", casos\_favorables**/**casos\_totales)

p(Queenstown de una persona elegida al azar) = 0.08641975308641975

Para que pienses un poco, ¿se vería afectado este cálculo si en vez de partir de todos los datos, la población, lo hacemos de una muestra?

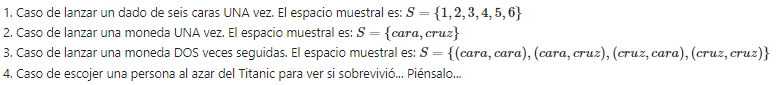
**Espacio muestral, Suceso, unión de sucesos e intersección de sucesos:**

Diagrama

Descripción generada automáticamente

**Espacio muestral**

El **espacio muestral** es el conjunto de todos los resultados posibles en un "experimento" aleatorio:



No, no es 

Aunque pueda parecerlo por los ejemplos anteriores el espacio muestral en el caso de la supervivencia del titanic son todos los valores posibles y en ese caso como estamos escogiendo un pasajero habrá tantos casos posibles como pasajeros (fijate en el caso de lanzar dos veces una moneda, el resultado no es cara o cruz sino el combinado):



Es decir el espacio muestral será:

df\_titanic["alive"].str.replace("yes","sobrevivió").replace("no","murió").values

array(['murió', 'sobrevivió', 'sobrevivió', 'sobrevivió', 'murió',

'murió', 'murió', 'murió', 'sobrevivió', 'sobrevivió',

'sobrevivió', 'sobrevivió', 'murió', 'murió', 'murió',

'sobrevivió', 'murió', 'sobrevivió', 'murió', 'sobrevivió',

'murió', 'sobrevivió', 'sobrevivió', 'sobrevivió', 'murió',

'sobrevivió', 'murió', 'murió', 'sobrevivió', 'murió', 'murió',

'sobrevivió', 'sobrevivió', 'murió', 'murió', 'murió',

…

'murió', 'sobrevivió', 'sobrevivió', 'murió', 'murió', 'murió',

'sobrevivió', 'sobrevivió', 'murió', 'murió', 'murió', 'murió',

'murió', 'murió', 'sobrevivió', 'murió', 'sobrevivió', 'murió'],

dtype=object)

Cuando trabajamos con datasets (conjuntos de datos) podemos asimilar el espacio muestral a dicho dataset para cada una de las características que tiene cada elemento del dataset o para su combinado.

**Suceso**

Un suceso es un subconjunto de un espacio muestral. Ejemplos de sucesos "sencillos" (subconjuntos de un único elemento):

1. En el caso de lanzar un dado... cada una de las caras que pueden salir. Es decir un suceso será "sacar un 4", otro "sacar un 1", etc
2. En el caso de lanzar una moneda, tendríamos dos sucesos sencillos ("sacar cara", "sacar cruz")

No todos los sucesos tienen por qué ser "sencillos". Piensa en el caso de lanzar una moneda dos veces seguidas, podemos definir el suceso sacar cara en alguna de las tiradas. ¿Como se definiría ese suceso?:



[Los sucesos o eventos se denotan normalmente por las letras A, B, C, etc]

Captura de pantalla de un celular

Descripción generada automáticamente

Y en el caso del titanic, tenemos dos posibles Sucesos (que corresponden con los valores de la variable categórica): "Sobrevivir", "Morir". Pero ninguno es sencillo. Cuando se da el caso de que dos o más eventos o sucesos no pueden ocurrir a la vez, se dice que son ***mutuamente excluyentes*** y la suma de sus probabiliddes debe ser igual a la unidad (o al 100% si trabajamos con porcentajes).

**Unión e interscción de sucesos**

* **Unión:** La unión de varios sucesos sencillos crea un evento compuesto que **ocurre si uno o más de los eventos ocurren** (es como un "OR"). La unión de A y B se escribe A∪B y significa "Ya sea A o B, o ambos A y B."
* **Intersección:** La intersección de dos o más eventos simples crea un evento compuesto que **ocurre sólo si ocurren todos los eventos simples** (es como un "AND"). La intersección de A y B se escribe A∩B y significa "A y B."

Ojo, a partir de la definición anterior puede haber unión e intersección de cualquier tipo de suceso y la probabilidad siempre se puede calcular igual: cardinalidad del resultado dividida por cardinalidad total (casos favorables entre casos posibles)

Con esto ya casi podemos obtener la respuesta a nuestras tres últimas preguntas iniciales, pero antes un par de ejemplos para ilustrar los conceptos anteriores:

**Ejemplo con monedas**

Supongamos que tenemos el "experimento" lanzar una moneda dos veces seguidas. Recuerda que su espacio muestral era:

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

**Ejemplo con dos variables diferentes**

Imagina que tu espacio muestral son 9 personas, de los cuales tienes 5 ciudadanos españoles (A), y 4 de pelo castaño (B).

Forma, Círculo

Descripción generada automáticamente

Escogiendo una persona al azar del espacio muestral:

* ¿Cuál es la probabilidad que sea un ciudadano español? P(A) = 5/9
* ¿Cuál es la probabilidad de que tenga el pelo castaño? P(B) = 4/9
* ¿Cuál es la probabilidad de que sea español o tenga el pelo castaño? P(A∪B) = 8/9
* ¿Cuál es la probabilidad de que sea español y tenga el pelo castaño? P(A∩B) = 1/9
* ¿Cuál es la probabilidad de que no sea español? P(∼A) = 4/9 (también se le dice suceso complementario de A)

**Aplicación a nuestro dataset**

Y ahora podemos aplicarlo directamente a nuestro dataset y a las preguntas:

1. ¿Cuál es la probabilidad de que una persona elegida al azar fuese hombre o viájese en primera clase?

Se trataría de la unión de dos sucesos (fíjate en el "o" de la pregunta):

Texto

Descripción generada automáticamente

En la siguiente sesión veremos como obtener las probabilidades sucesos unión a partir de sus sucesos originales pero por ahora lo que vamos a hacer es seguir aplicando casos favorables dividido por casos totales:

df\_titanic.columns

Index(['sex', 'age', 'sibsp', 'parch', 'fare', 'class', 'who', 'adult\_male', 'embark\_town', 'alive', 'alone'],

dtype='object')

*# Como es una unión es como aplicar un "OR" a las dos condiciones:*

ser\_hombre **=** df\_titanic.who **==** "Male"

viajar\_primera **=** df\_titanic["class"] **==** "First"

casos\_favorables **=** len(df\_titanic[ser\_hombre **|** viajar\_primera]) *# Esto equivale a calcular la cardinalidad del conjunto unión*

casos\_totales **=** len(df\_titanic)

print(f"p(ser hombre o viajar en primera) = {round(casos\_favorables**/**casos\_totales **\*** 100,2)}%")

p(ser hombre o viajar en primera) = 24.24%



1. ¿Cuál es la probabilidad de que una persona elegida al azar viajese en primera clase y viajase sola?

En este caso estamos hablando de la intersección de dos sucesos (fíjate en el "y"):

Texto

Descripción generada automáticamente

Buscamos:



*#Como es una intersección es aplicar un AND:*

viajar\_sola **=** df\_titanic.alone

viajar\_primera **=** df\_titanic["class"] **==** "First"

casos\_favorables **=** len(df\_titanic[viajar\_sola **&** viajar\_primera]) *# Esto equivale a calcular la cardinalidad del conjunto INTERSECCION*

casos\_totales **=** len(df\_titanic)

print(f"p(de escoger a una persona que viajara en primera, sola) = {round(casos\_favorables**/**casos\_totales **\*** 100,2)}%")

p(de escoger a una persona que viajara en primera, sola) = 12.23%



Y para terminar: ¿Cuál es la probabilidad de que una persona elegida al azar se salvase y viajase en tercera clase?

*#Como es una intersección es aplicar un AND:*

superviviente **=** df\_titanic.alive **==** "yes"

viajar\_tercera **=** df\_titanic["class"] **==** "Third"

casos\_favorables **=** len(df\_titanic[superviviente **&** viajar\_tercera]) *# Esto equivale a calcular la cardinalidad del conjunto INTERSECCION*

casos\_totales **=** len(df\_titanic)

print(f"p(sobrevivir y viajar en tercera) = {round(casos\_favorables**/**casos\_totales **\*** 100,2)}%")

p(sobrevivir y viajar en tercera) = 13.36%



Ojo NO hemos calculado la probabilidad de que sabiendo que viajaba en tercera se salvase. Eso es lo que se llama probabilidad condicionada y la veremos en la siguiente sesión. Pero para que no te quedes con las ganas y lo vayas pensando, sería (de nuevo considerando casos favorables y casos posibles):

*# Al poner la condición de que viaja en tercera, lo que hacemos es reducir el espacio muestral y por tanto el denominador (al reducir la cardinalidad y por tanto el número de casos posibles)*

casos\_favorables **=** len(df\_titanic[superviviente **&** viajar\_tercera]) *# Igual que antes son todos los que viajaban en tercera y se salvaron*

casos\_totales **=** len(df\_titanic.loc[df\_titanic["class"] **==** "Third"]) *# Aquí está la diferencia, como sabemos que va en tercera, los casos posibles son solo los que van en tercera*

print(f"p(habiendo viajado en tercera, sobrevivir) = p(sobrevir de los pasajeros de tercera) = {round(casos\_favorables**/**casos\_totales **\*** 100,2)}%")

p(habiendo viajado en tercera, sobrevivir) = p(sobrevir de los pasajeros de tercera) = 24.24%

**Unión, intersección, probabilidad condicional e independencia**

Para esta sesión seguimos tratando con el dataset del titanic y sus variables categóricas:

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

**import** numpy **as** np

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

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

Existen otras formas de obtener las probabilidades de la unión de sucesos y de la intersección de sucesos que no son necesariamente aplicar la regla de Laplace. Empecemos con la unión.

**Obtención de probabilidades de unión de sucesos**

Distinguimos dos casos, cuando los suscesos a unir son mutuamente excluyentes y cuando no.

**Unión de eventos mutuamente excluyentes**

Recuerda que mutuamente excluyentes es que no pueden darse a la vez (matemáticamente la intersección de los conjuntos de sucesos excluyentes es el conjunto vacio porque no se pueden dar los dos eventos a la vez).

Diagrama de Venn

Descripción generada automáticamente con confianza baja

Si los eventos son *mutuamente excluyentes* entonces para calcular la [probabilidad](https://es.wikipedia.org/wiki/Probabilidad) de su unión, simplemente sumamos sus [probabilidades](https://es.wikipedia.org/wiki/Probabilidad) individuales.

Imagen de la pantalla de un celular con letras

Descripción generada automáticamente con confianza baja

Como ejemplo, obtengamos la probabilidad del suceso "ser mujer o niño" al escoger al azar un pasajero del titanic. Siguiendo la fórmula anterior:

Texto

Descripción generada automáticamente

Obtenemos P(E) y P(F) usando la aproximación frecuencial:

*# p(E) frecuencia relativa de ser mujer:*

p\_E **=** df\_titanic["who"].value\_counts(normalize**=** **True**)["woman"]

p\_F **=** df\_titanic["who"].value\_counts(normalize **=** **True**)["child"]

print("p(E) = ", p\_E)

print("p(F) = ", p\_F)

p(E) = 0.3041526374859708

p(F) = 0.0931537598204265

Y ahora la probabilidad de la unión P(E∪F) como la suma:

p\_E\_U\_F **=** p\_E **+** p\_F

print("p(escoger una pasajero que sea mujer o niño) = p(E U F) = ", p\_E\_U\_F)

p(escoger una pasajero que sea mujer o niño) = p(E U F) = 0.39730639730639733

Y comprobamos haciendo un calculo frecuencial del suceso combinado como lo hicimos anteriormente, al ser unión es un or de las condiciones:

casos\_favorables **=** len(df\_titanic.loc[(df\_titanic.who **==** "child") **|** (df\_titanic.who **==** "woman")])

casos\_totales **=** len(df\_titanic)

print(casos\_favorables**/**casos\_totales)

0.39730639730639733

**Unión de eventos que no son mutuamente excluyentes**

Si los eventos no son *mutuamente excluyentes* entonces debemos corregir la fórmula anterior para incluir el efecto de la superposición de los eventos. Esta superposición se da en el lugar de la *intersección* de los eventos; por lo tanto la formula para calcular la [probabilidad](https://es.wikipedia.org/wiki/Probabilidad) de estos eventos es:

****

**Ejemplo**

Volviendo al Titanic, pensemos en obtener la probabilidad de que un pasajero al azar "sobreviviese o viajase solo". Ahora los sucesos no son excluyentes ya que hubo pasajeros que viajaban sólos y sobrevivieron. Usando la formula:

P(E)=P(sobrevivir):

casos\_favorables **=** len(df\_titanic.loc[df\_titanic.alive **==** "yes"])

casos\_totales **=** len(df\_titanic)

p\_E **=** casos\_favorables**/**casos\_totales

print("p(E) = ", p\_E)

p(E) = 0.3838383838383838

P(F)=P(viajar solo)

[6]:

casos\_favorables **=** len(df\_titanic.loc[df\_titanic.alone])

casos\_totales **=** len(df\_titanic)

p\_F **=** casos\_favorables**/**casos\_totales

print("p(F) = ", p\_F)

p(F) = 0.6026936026936027

P(E∩F)=P(viajar solo y sobrevivir) que obtenemos frecuencialmente como hicimos en la sesión anterior

casos\_favorables **=** len(df\_titanic.loc[(df\_titanic.alive **==** "yes") **&** (df\_titanic.alone)])

casos\_totales **=** len(df\_titanic)

p\_E\_x\_F **=** casos\_favorables**/**casos\_totales

print("p(E x F) = ", p\_E\_x\_F)

p(E x F) = 0.1829405162738496

p\_E\_U\_F **=** p\_E **+** p\_F **-** p\_E\_x\_F

print("p(E U F) = ", p\_E\_U\_F)

p(E U F) = 0.803591470258137



**Sucesos Independientes y Probabilidad Condicional**

Antes de ver como obtener la probabilidad de la intersección de sucesos o eventos, es bueno introducir dos conceptos. Por un lado el de sucesos independientes y por otro el de probabiliad condicional:

**Sucesos Independientes**

*"En términos de probabilidad, los sucesos independientes son dos o más eventos cuya ocurrencia no afecta la probabilidad de ocurrencia de los otros. Esto significa que la ocurrencia de un evento no tiene influencia sobre la ocurrencia del otro."* (por ejemplo: la tiradas de un dado, el resultado de un partido de baloncesto y que llueva esa tarde)

Nota: En el caso del Titanic es realmente difícil encontrar sucesos o variables completamente independientes porque la gran mayoría están relacionadas como ya vimos al hacer su mini-EDA

**Probabilidad Condional**

Con frecuencia queremos conocer la [probabilidad](https://es.wikipedia.org/wiki/Probabilidad) de algún evento, dado que otro evento ha ocurrido. **Esto se expresa simbólicamente como  P(E|F) y se lee como "la**[**probabilidad**](https://es.wikipedia.org/wiki/Probabilidad)**de E dado F".** El segundo evento se conoce como la *condición* y el proceso se refiere a veces como "condicionamiento en F". La [probabilidad condicional](https://es.wikipedia.org/wiki/Probabilidad_condicionada) es un concepto importante de estadística, porque a menudo estamos tratando de establecer que un factor tiene una relación con un resultado, como por ejemplo, que las personas que fuman cigarrillos tienen más [probabilidades](https://es.wikipedia.org/wiki/Probabilidad) de desarrollar cáncer de pulmón. La [probabilidad condicional](https://es.wikipedia.org/wiki/Probabilidad_condicionada) también se puede usar para definir la [independencia](https://es.wikipedia.org/wiki/Independencia_(probabilidad)). **Dos variables se dice que son**[**independientes**](https://es.wikipedia.org/wiki/Independencia_(probabilidad))**si la siguiente relación se cumple:**



Probabilidad de ser fumador si eres asmático

**Calculando la probabilidad condicional**  
Para calcular la probabilidad del evento E dada la información de que el evento F ha ocurrido utilizamos la siguiente formula:



Esta fórmula es la que seguimos al final de la sesión anterior para calcular la probabilidad de que los pasajeros de tercera sobrevivieran, es decir la probabilidad de escoger un pasajero que sobrevivió condicionado a que fuera un pasajero de tercera:

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

 que obteníamos empleando el método frecuencial para numerador y denominador y simplificando. Vamos a hacerlo ahora calculando las dos probabilidades por separado:

P(E∩F) la obtuvimos como ejemplo de intersección de sucesos:

*#Como es una intersección es aplicar un AND:*

superviviente **=** df\_titanic.alive **==** "yes"

viajar\_tercera **=** df\_titanic["class"] **==** "Third"

casos\_favorables **=** len(df\_titanic[superviviente **&** viajar\_tercera]) *# Esto equivale a calcular la cardinalidad del conjunto INTERSECCION*

casos\_totales **=** len(df\_titanic)

p\_E\_x\_F **=** casos\_favorables**/**casos\_totales

P(F) se obtiene de calcular la frecuencia relativa de pasajeros de tercera, pero vamos a obtenerla como una "probabilidad":

casos\_favorables **=** len(df\_titanic.loc[df\_titanic["class"] **==** "Third"])

casos\_totales **=** len(df\_titanic)

p\_F **=** casos\_favorables**/**casos\_totales

Y la probabiliad condicional:

p\_condicionada **=** p\_E\_x\_F **/** p\_F

print("p(E|F) = ", p\_condicionada)

p(E|F) = 0.24236252545824846

Que, como era de esperar, es el mismo valor que nos salió en la píldora anterior.

**Obtención de probabilidades de intersección de sucesos**

Para ver formas alternativas de obtener la probabilidad de la intersección de sucesos o eventos los vamos a dividir en dos tipos también, pero ahora tratamos sucesos independientes y sucesos dependientes.

**Intersección de eventos independientes**

Para calcular la [probabilidad](https://es.wikipedia.org/wiki/Probabilidad) de que ocurran varios eventos (la intersección de varios eventos), se multiplican sus [probabilidades](https://es.wikipedia.org/wiki/Probabilidad) individuales. La fórmula específica utilizada dependerá de si los eventos son [independientes](https://es.wikipedia.org/wiki/Independencia_(probabilidad)) o no. Si son [independientes](https://es.wikipedia.org/wiki/Independencia_(probabilidad)), la [probabilidad](https://es.wikipedia.org/wiki/Probabilidad) de E y F se calcula como:

****

**Ejemplo**

Tiro tres dados. ¿Cuál es la probabilidad de que en las tres tiradas salga un 4? Se trata de tres eventos independientes, y la probabilidad de ocurra cada uno de ellos por separado es de 1/6.

Por tanto, la probabilidad de que en las tres tiradas salga un 4, será de (1/6) \* (1/6) \* (1/6) = (1/216) = 0.46 %

**Intersección de eventos no independientes**

Si dos eventos no son [independientes](https://es.wikipedia.org/wiki/Independencia_(probabilidad)), debemos conocer su [probabilidad condicional](https://es.wikipedia.org/wiki/Probabilidad_condicionada) para poder calcular la [probabilidad](https://es.wikipedia.org/wiki/Probabilidad) de que ambos se produzcan. La fórmula en este caso es:



**Ejemplo**

Saco dos cartas de una baraja española (sin devolverlas al mazo), ¿cuál es la probabilidad de que ambas cartas sean reyes?

Si el evento E es "obtener un rey la primera vez que sacamos carta", P(E) = 4/40. Si no devuelvo la carta al mazo, se quedaría con 39 cartas. Ahora necesito calcular la probabilidad del segundo evento, es decir, la probabilidad de que la segunda vez que saque carta sea un rey, codicionado al evento anterior, por tanto, P(F/E) = 3/39.

Por tanto, nuestra probabilidad de obtener dos reyes en dos cartas consecutivas es de: (4/40) \* (3/39) = 12/1560

12**/**1560

0.007692307692307693

**Variables continuas**

Para esta sesión nos cambiamos al dataset de la satisfacción vital:

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

**import** numpy **as** np

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

**import** seaborn **as** sns

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

**Probabilidades de variables contínuas**

Al igual que hicimos al introducir el tema de la probabilidad con el titanic, planteemos una pregunta pero para el dataset, por ejemplo cuál es la probabilidad de que escogido un país al azar su valor de satisfacción vital sea 5.8

En teoría, con lo que sabemos aplicariamos la regla de Laplace es decir:

Texto

Descripción generada automáticamente con confianza media

Y la obtendríamos... Pues hagámoslo

casos\_favorables **=** len(df\_lifesat.loc[df\_lifesat["Life satisfaction"] **==** 5.8])

casos\_totales **=** len(df\_lifesat)

print("p(Pais con satisfaccion igual a 5.8)", casos\_favorables**/**casos\_totales)

p(Pais con satisfaccion igual a 5.8) 0.0

Hmmm, vaya, nos sale cero. ¿Por qué? Pues sencillamente porque no hay un país con exactamente un valor de 5.8, y en general puedes probar con varios valores que te saldrá siempre una probabilidad de 0. Es cierto que en algunos casos puede que obtengas una probabilidad no nula porque porque se hayan repetido algunos valores, pero en general es muy raro porque las variables numéricas continuas tienen infinitos valores posibles y por tanto sólo al limitar la precisión y el rango podremos encontrarnos útil el usar la frecuencia.

Por eso cuando trabajamos con variables numéricas contínuas de las que queremos saber su probabilidad trabajamos con rangos de valores, y obtenemos la probabilidad de que nuestra variable continua tenga su valor en ese rango. Por ejemplo, sí tiene sentido preguntarnos "escogido un país al azar, ¿cuál es la probabilidad de que su satisfacción vital este por debajo de 6?" o "que esté entre 6.5 y 7.5".

¿Y cómo obtenemos esos valores? Esta vez tenemos que acudir a la función densidad de probabilidad que vimos cuando analizamos variables contínuas en el módulo anterior.

**Función densidad de probabilidad**

La función de densidad de probabilidad de una variable numérica (aleatoria) nos permite calcular la probabilidad de un rango de valores calculando la integral de la función entre esos dos valores. Por ejemplo, si obtenemos el histograma y la función densidad de probabilidad "estimada" para la satisfacción vital de nuestro dataset, recordando un poco de seaborn:

sns.histplot(x **=** "Life satisfaction", data **=** df\_lifesat, kde **=** **True**)

<Axes: xlabel='Life satisfaction', ylabel='Count'>

Gráfico, Histograma

Descripción generada automáticamente

Si queremos contestar a la pregunta ¿cuál es la probabilidad de que un país tenga una "Life satisfaction" por debajo de 5.5, tendremos que integrar la curva estimada entre −∞ y 5.5:

[¿Por qué menos infinito y no el mínimo de los valores, porque estamos usando una curva aproximada entre esos rangos y no los valores exactos. Sí eso quiere decir que obtendremos un valor aproximado de la probabilidad]



Pues antes de que te asustes, utilizaremos las bibliotecas de Python para hacer la estimación de esas integrales por ti.

**Estimación de probabilidades de rangos de variables continuas**

El procedimiento es análogo al comentado pero usando Python:

[4]:

*#1 Obtenemos la estimación de la función densidad de probabilidad de la variable numérica*

**from** scipy.stats **import** gaussian\_kde *# Usamos este estimador, ojo no estima Kde gaussianos usa un metodo basado en gaussianas, que no es lo mismo*

f\_densidad **=** gaussian\_kde(df\_lifesat["Life satisfaction"])

*#2 Definimos los rangos de intregación*

rango\_min **=** **-**np.infty

rango\_max **=** 5.5

*#3 Integramos*

**from** scipy.integrate **import** quad *# Utiliza un método de integración numérico sobre el que no vamos a profundizar :-/*

probabilidad,\_ **=** quad(f\_densidad, rango\_min, rango\_max)

*#4 mostramos el resultado*

print(probabilidad)

0.11246323841562425



Que es un valor bastante consistente viendo la distribución de la figura anterior.

Para la otra pregunta, "probabilidad de que escogiendo un país al azar su satisfación esté entre 6.5 y 7.5"



Aprovechamos que ya hemos estimados su función densidad y simplemente tenemos que cambiar los rangos y volver a integrar:

[8]:

rango\_min **=** 6.5

rango\_max **=** 7.5

probabilidad,\_ **=** quad(f\_densidad, rango\_min, rango\_max)

print(probabilidad)

0.4060194849627096

