Probabilidad y Estadística con Numpy

Rodrigo Reyes M.

March 18, 2016

1 Introducción a la probabilidad

Este es un pequeño documento para recordar las nociones de probabilidad y estadística tipicamente vistas en la licenciatura.

Existen 3 axiomas principales postulados por el matemático ruso Kolmogorov:

1. La probabilidad siempre es expresada en un número real mayor o igual a cero y menor o igual a uno.

$$0 \leq P(A) \leq 1$$

2. La probabilidad de el espacio muestral es uno.Es decir no hay eventos ajenos al espacio muestral.

$$P(\Omega) = 1$$

3. La probabilidad conjunta (o la unión) de eventos mutuamente exclusivos o ajenos es la suma de sus probabilidades.

$$P(E_1 \cup E_2 \cup ..) = \sum_{i=1}^{\infty} P(E_i)$$

Además de los axiomas de Kolmogorov hay dos conceptos muy importantes.

Dos eventos son ajenos o independientes si su probabilidad no influye en la probabilidad del otro.

Ejemplos:

- Lanzar una moneda no afecta la probabilidad de obtener 6 en un dado por tanto son eventos independientes
- Si saco una pelote roja de una urna con 2 pelotas rojas y 100 negras la probabilidad de obtener nuevamente una pelota roja es menor por tanto NO SON EVENTOS INDEPENDIENTES.
- Aunque paresca poco intuitivo el haber obtenido 7 caras en una serie de lanzaminetos de moneda NO AFECTAN la probabilidad de obtener nuevamente cara por tanto son independientes.

2 Ejercicios

1. Si P(A) = 0.4 ¿Cuál es la probabilidad de $\neg P(A)$?
2. De los siguientes eventos indique con la letra I si son independientes y X si no lo son.
   - La probabilidad de obtener un rey de corazones si obtuve un rey de tréboles.
   - La probabilidad de obtener una reina de espadas si obtuve un 3 de corazones.
   - La probabilidad de que llueva si me gané la lotería.
3. Tenemos una urna con 5 pelotas de la siguiente forma, 2 son rojas , otras 2 son verdes y 1 es negra. Si realizo dos experimentos en donde saco una pelota y la vuelvo a insertar en la urna para sacar nuevamente una pelota.¿Cuál es la probabilidad de obtener una pelota negra?

3 Introducción a Numpy y Python

Python es un lenguaje interpretado con Orientación a Objetos creado por Guido van Rossum.En éste curso utilizaremos la biblioteca de Numpy que es ampliamente utilizada para análisis numérico y minería de datos.

Las funciones en Python se definen de la siguiente forma

```
define fun():
    print("hola")
```

El lenguaje Python utiliza Tabs o espacios para delimitar bloques de código, es decir identar en Python es equivalente a generar un bloque de { } en Java.

Cabe destacar que es recomendable utilizar uno de los dos métodos y no ambos (es decir no mezclar Tabs y espacios).Algunos editores de texto convierten automáticamente las tabs en cuatro espacios.

Para éste curso se requiere instalar IPython que es una herramienta que simplifica mucho el análisis de datos.

Ésta herramienta al ser instalada con SciPy instala automáticamente las dependencias que nos permiten utilizar tanto Numpy como Matplotlib (una herramienta para generar gráficas).

3.1 Funciones básicas de Numpy

Para comenzar con los siguientes ejercicios deberás ejecutar el comando

```
ipython --pylab
```

Una vez iniciado el ambiente, ejecuta el siguiente comando.

```
random.random_sample()
```

El objeto random tiene un método `random_sample` que al ser llamado sin argumentos genera un número entre 0 y 1 de manera aleatoria. (La muestra se toma de una distribución uniforme que se verá más adelante)

3.2 Ejercicios de código

1. Crea un archivo en el editor de tu preferencia que se llame `proba.py`
2. En ése archivo crea la función `lanzar_moneda` que retorne 0 si al llamar a `random.random_sample()` el número es menor o igual a 0.5 y 1 en otro caso.
3. Carga el archivo `proba.py` en IPython.
4. Revisa que puedas ejecutar `lanzar_moneda()`
5. Modifica `proba.py` ahora creando una funcion `simular(n)` que simule lanzar la moneda n veces y regrese un arreglo con la primera posición indique la suma de veces que `lanzar_moneda` regresó 0 y en la segunda posición la suma de veces que regresó 1.

Tip: Para crear un arreglo en Numpy se crea con el método array y con una tupla como argumento.

```
array([1,2,3]) # Arreglo con los elementos 1 , 2 ,3.
```

3.3 ¡Vamos a graficar!

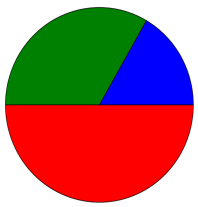
Es hora de ver los resultados del pequeño simulador de lanzamiento de monedas que hemos creado.Muchas veces los ananlistas pueden deducir muchas propiedades sobre los datos al ser graficados.

La forma más sencilla de ver la información es mediante una gráfica de pastel o pie plot.

En IPython es extremadamente fácil crear uno.Simplemente ejecuta la siguiente linea de código.

```
pie([10,20,30])
```

El resultado debe ser similar a la siguiente figura



Matplotlib permite modificar muchas cosas respecto a una gráfica desde los colores, las etiquetas e incluso hacer animaciones.

Si deseas saber más acerca de los parámetros que se pueden modificar el siguiente link te ayudará `http://matplotlib.org/1.2.1/api/pyplot_api.html?highlight=hist#matplotlib.pyplot.pie`

3.4 Ejercicios de código

1. Modifica `proba.py` creando una funcion `graficar(n)` que muestre una gráfica utilizando a la n como parametro para la función `simular(n)`.Es decir la gráfica de Pie debe mostrar la proporcion de cara o cruz en los lanzamientos virtuales de moneda.
2. Ejecuta la función `graficar(5)`
3. Ejecuta la función `graficar(10)`
4. Ejecuta la función `graficar(10000)`

4 Estadística

4.1 Teorema de Límite Central

Una de los pilares de la estadística moderna se centra en el Teorema de Límite Central.Formalmente nos dice:

Sea $(X_1, X_2, ..., X_n)$ un conjunto de variables aleatorias independientes e idénticamente distribuidas con media μ y varianza $0 < \sigma^2 < \infty$.Entonces para una n lo suficientemente grande la media se distribuye normal con media μ y varianza σ^2.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2)$$

Lo importante de el teorema es que nos permite que al tener una muestra grande modelamos utilizando la distribución normal o gaussiana cuyas propiedades son bien conocidas y ampliamente estudiadas.

4.2 Ley de los grandes números

Es importante conocer la relación entre la probabilidad y la estadística.En los ejercicios previos se pudo observar que las simulaciones computacionales pueden aproximarse a su probabilidad. Por ejemplo la siguiente es una gráfica realizada a partir de una simulación de lanzamientos de dos dados.

A este tipo de gráficos se les llama histogramas y nos muestra visualmente la distribución de los datos. Se puede observar claramente que al realizar el experimento 10000 veces los datos toman una forma muy peculiar, similar a la de una campana.También es fácil ver que el 7 es el valor más común.Veamos que la probabilidad de obtener una suma de 7 al tirar dos dados es 0.16 y se

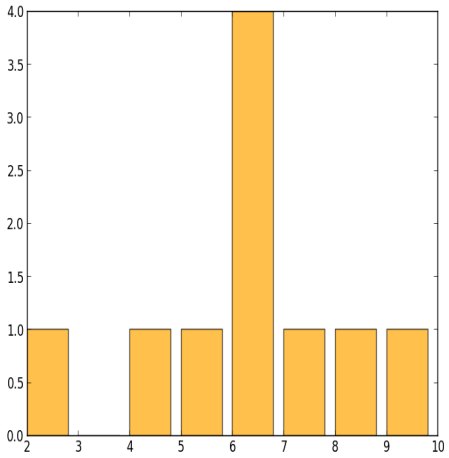


Figure 1: Histograma de lanzamiento de dos dados con 10 experimentos

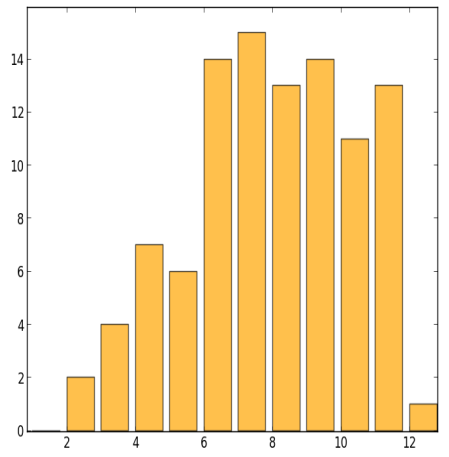


Figure 2: Histograma de lanzamiento de dos dados con 100 experimentos

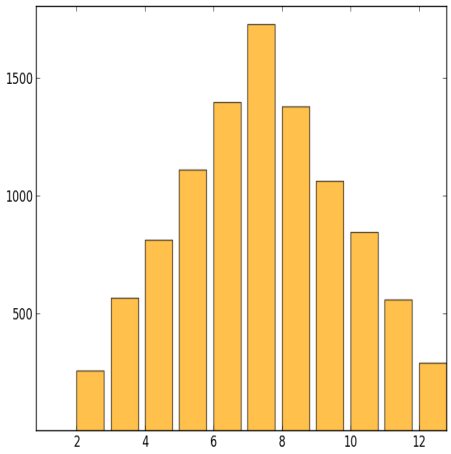


Figure 3: Histograma de lanzamiento de dos dados con 10000 experimentos

obtuvieron aproximadamente 1740 sietes.Ésto nos dá una proporción de 0.174 que es un número muy cercano a 0.16.(Incluso se ejecutó una simulación con 10000000 de experimentos y se obtuvieron 167129 que es 0.167129 todavía más cercano a 0.16).Entonces formalmente la ley de grandes números nos dice (En este caso la ley fuerte, la ley débil es ligeramente distinta y se aplica en otros casos.)

$$P(\lim_{n\to\infty} \bar{X} = \mu) = 1$$

Es decir la media muestral se parece cada vez más a la media global. Incluso utilizando el teorema previo de límite central los datos se pueden modelar con una distribución normal con $\mu = 7$ Sin embargo, hay dos cosas importantes a destacar, como se puede ver en el histograma los datos no tenian una distribución aparente cuando el número de experimentos es pequeño.Ésto es de enorme importancia cuando se trabaja con datos debido a que se pueden cometer errores al tener conjuntos muy pequeños de datos que no representen correctamente al total de la población.Esto hay que tomarlo como pauta para el último tema que son las falacias estadísticas más comunes y cómo reducir o detectarlas.

4.3 Falacias estadísticas

Es importante poder identificar las falacias estadísticas.Una frase célebre de Mark Twain *There are three kinds of lies: lies, dammed lies and statistics*, muchas veces la pseudociencia y algunas investigaciones mal hechas utilizan datos estadísticos como único sustento a teorías que no son las correctas.

La falacia más común es *Correlación no implica causalidad*, esto intenta sustentar una generalización a partir de una correlación estadística.

Por ejemplo supógase que se realiza un estudio y se determina que estadísticamente los alumnos obtienen mejores calificaciones en el semestre cercano a el invierno.

De ésto se concluye que el aprendizaje está ligado a las estaciones o al clima.

Es fácil ver que a pesar que existe una aparente correlación no hay suficiente evidencia que permita sustentar que el invierno cause mejoría en los estudios.Bien se puede deber a otros factores *escondidos o no aparentes*, tales como

1. El semestre es más corto por lo que las clases se reducen.
2. Los profesores tienden a ausentarse más debido a que hay mayor cantidad de fechas festivas.
3. etc..

En muchos estudios serios se utilizan grupos de control para poder reducir conclusiones erróneas y sobretodo se hace mención de el error que tienen las mediciones y la magnitud de los muestreos.

Otra falacia muy común es comunmente vista con una anécdota. Un turista visita Texas y descubre que en un granero hay una serie de circulos concéntricos de color rojo y blanco a manera de objetivos.Y en cada uno de los círculos hay un agujero causado por una bala.El turista inmediatamente felicita al granjero que porta un rifle suponinedo que resulta ser un experto en tiro al blanco, mientras su esposa le dice *En relidad es muy malo en tiro al blanco, sin embargo es muy bueno pintando blancos en los hoyos que dejan las balas.* Ésta falacia se debe a que amoldamos las teorías ó hipótesis para que embonen con los datos.

4.4 Ejercicios

1. Investiga la función de densidad para la distribución normal o gaussiana.
2. ¿Que significa μ y σ ? (En la función de densidad normal)
3. ¿Qué es la desviación estándar?¿Cómo está relacionada con la varianza?
4. Encuentra o inventa una falacia estadística.Describe ¿porqué es falacia?

4.5 Ejercicios de código

1. Crea una simulación de una urna con 5 pelotas: 2 rojas, 2 verdes y una negra.Tal cómo se describe en el ejercicio 3 de la sección 2.
2. Realiza la simulación 10000 veces y grafica el resultado en un gráfico de pastel con los colores correspondientes a el color de las pelotas.
3. Indica la proporción que se obtuvo.¿Qué tan cercana es a el resultado que se obtuvo utilizando probabilidad?¿Porqué?