PUCE Virtual

Estadística Aplicada

Principios de Probabilidad

Clase 3

Ingeniería en ciberseguridad

La excelencia no se improvisa

1. INTRODUCCIÓN DE LA CLASE

En esta clase, iniciaremos con el análisis de la teoría de probabilidades. Continuamente, los ejecutivos de una empresa se enfrentan a la toma de decisiones de negocios y, en general, sin importar la naturaleza del negocio, es claro que dichas decisiones se basarán en la probabilidad de que determinado resultado, positivo o negativo, se produzca. Por otro lado, como habíamos visto en clases anteriores, este tipo de análisis estará siempre sujeto a un cierto nivel de incertidumbre.

Esta clase inicia con la demostración de las distintas formas de medir la probabilidad de que un evento se presente. Estas varían desde situaciones simples a complejas, incluyendo casos que involucran variables numéricas y categóricas, así como situaciones que consideran una muestra o población.

3) PRINCIPIOS DE PROBABILIDAD

DEFINICIONES BÁSICAS

Población: La población estadística es el total de individuos o conjunto de ellos que presentan o podrían presentar el rasgo característico que se desea estudiar (economipedia). Algunos autores llaman a este concepto el universo. Veamos algunos ejemplos:

Si el gerente de una empresa desea investigar los problemas de producción, la empresa tiene 5 plantas o fábricas, dado que los resultados son específicos para esta empresa, podemos decir que la población será todas las 5 plantas

Si se desea estudiar el efecto de una vacuna en niños, entonces la población será el total de niños en los que se usaran la vacuna (no el total de niños que existen)

Si se desea estudiar la preferencia electoral se entrevistan a 10.000 personas, la población en este caso sería los ciudadanos con derecho a voto.

Muestra: Es una parte representativa de la población que se selecciona para ser estudiada. Hay varias razones por las cuales un trabajo de investigación no es posible hacerlo con toda la población: Por ejemplo:

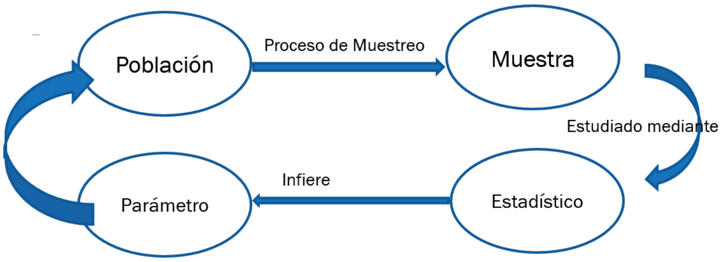
- ✓ Ejecutar prueba sobre el 100% de la población es muy costoso, como probar una vacuna en toda la población para medir el efecto de esta.
- ✓ La población está geográficamente dispersa, por lo que sería muy costoso movilizarse a muchos lugares.
- ✓ El análisis sobre cada elemento de la población termina dañando a la población. Por ejemplo, queremos medir la resistencia de un ladrillo a la presión, en este caso el experimento terminaría con las destrucciones del ladrillo. Si lo hacemos sobre el 100% de la producción de ladrillos de una fábrica, esta terminará sin inventario de ladrillos.

Parámetro y Estadístico:

Estos dos conceptos son similares, pero aplican a distintos grupos. Se define como parámetro a una medida descriptiva de la población total, mientras que el estadístico representa lo mismo, pero es el valor obtenido de la muestra. Por ejemplo, siguiendo con el ejemplo del gerente de la empresa que tiene 5 plantas de producción, el gerente ha seleccionado 2 de ellas para contabilizar la producción diaria. Esta contabilización sería mediante la medición de la producción diaria; a esto lo llamamos el estadístico, en base al cual podemos extrapolar la producción de las 5 plantas, que es el parámetro. La figura 1 muestra la relación entre estos dos conceptos:

Figura 1: Relación entre estadístico y parámetro

Creación autor Alfonso Prado



Error de Muestreo:

El error de muestreo se define como la diferencia entre el parámetro desconocido de la población y el estadístico de la muestra utilizado para calcular el parámetro. La exactitud de toda estimación depende en gran parte del muestreo y de que este sea representativo; sin embargo, como no podemos garantizar que la muestra sea 100% representativa, siempre habrá un error de muestreo.

Sesgo Muestral:

También conocido como “bias,” es la tendencia a favorecer la selección de ciertos elementos de muestra en lugar de otros. Esto ocurre cuando hay una situación, intencional o no, que induce a seleccionar determinados elementos de la muestra. Por ejemplo, estimar la preferencia de voto en localidades donde nuestro candidato es favorito.

Teoría del Muestreo:

En base a lo indicado anteriormente, tenemos 2 conceptos que debemos balancear. El primero es que la muestra debe ser representativa de la población y el segundo es que debemos mantener una aleatoriedad en la selección de la muestra. De acuerdo con Otzen y Manterola (2017), “La representatividad de una muestra permite extrapolar y, por ende, generalizar los resultados observados en esta a la población accesible; y, a partir de esta, a la población blanco. Por ende, una muestra será representativa o no solo si fue seleccionada al azar, es decir, que todos los sujetos de la población blanco tuvieron la misma posibilidad de ser seleccionados en esta muestra y ser incluidos en el estudio y, por otro lado, que el número de sujetos seleccionados represente numéricamente a la población que le dio origen respecto de la distribución de la variable en estudio en la población, es decir, la estimación o cálculo del tamaño de la muestra”.

Técnicas de muestreo

Describe los distintos mecanismos de muestreo

Enlace

Precisión

De acuerdo con www.usc.gal Es la proximidad entre las indicaciones o los valores medidos, obtenidos en mediciones repetidas de un mismo objeto, bajo condiciones especificadas. La precisión se puede expresar numéricamente mediante medidas de dispersión tales como la desviación típica, varianza o el coeficiente de variación bajo las condiciones especificadas. La precisión se utiliza para definir la repetibilidad de medida.

Ejemplo: quiero medir la temperatura de un vaso de agua. Realizo 20 mediciones y no todas arrojan el mismo valor; hay décimas de grado de diferencia debido a la precisión del instrumento de medida.

Muchas veces, la precisión del instrumento de medida puede afectar el tipo de variable que se definirá. Por ejemplo, si en la medición de la temperatura del agua utilizamos un instrumento cuya precisión es de 1 grado, entonces la temperatura ya no se puede considerar como variable continua sino discreta.

Exactitud

En primer lugar, vale la pena puntualizar que la exactitud, a diferencia de la precisión, no se expresa en forma numérica, sino que más bien es la diferencia entre el valor obtenido y algo que se considera verdadero. Por ejemplo, mido la temperatura del aire con un termómetro y la comparo con la temperatura publicada en una página web que se considera "correcta." Por lo tanto, si lo comparo con otra página web, mi medición podría no ser tan correcta como pensaba.

Incertidumbre

De acuerdo con www.usc.gal, la incertidumbre es el parámetro asociado con el resultado de una medición, que caracteriza la dispersión de los valores que podrían ser razonablemente atribuidos al valor a medir. El valor de incertidumbre incluye componentes procedentes de efectos sistemáticos en las mediciones, debido a componentes que se calculan a partir de distribuciones estadísticas de los valores que proceden de una serie de mediciones y valores que se calculan a partir de funciones de densidades de probabilidad basadas en la experiencia u otra información.

¡Definición un tanto complicada! Más sencillo: cuando se desea medir un parámetro de la población, lo haremos mediante un muestreo, como ya se ha visto. El nivel de variabilidad entre una muestra y otra es la incertidumbre. Por ejemplo, tengo un dataset con personas y datos relacionados a ellas. Realizo un muestreo y obtengo una media de la edad de la muestra, digamos 35 años; luego realizo otro muestreo y obtengo 45 años; luego realizo otra medida y obtengo 40 años; puedo concluir que la incertidumbre está alrededor de 10 años.

3.1) Probabilidad condicional

En el lenguaje formal de la incertidumbre, que es la base de la inferencia, el problema básico que estudiamos en probabilidad es: dado un proceso generador de datos, ¿cuáles son las propiedades del resultado? Wasserman (2015). Más específicamente, cuán seguido podemos esperar obtener un valor en particular si repetimos una medición muchas veces y se denota como P(X).

La probabilidad se define como un valor numérico que estará entre 0 y 1 y representa cuál es la posibilidad de que un evento se presente; se escribe de la siguiente forma: $P(x) = 0 \leq x \leq 1$, donde x representa el evento. Por ejemplo, la probabilidad de pasar este curso sin estudiar sería P(sin estudiar) = 0.

En general, podemos establecer la siguiente fórmula para calcular una probabilidad:

Figura 2

Fórmula de la probabilidad

$$P(E) = \frac{\text{número de resultados favorables}}{\text{número total de posibles resultados}} = \frac{n(E)}{n(S)}$$

Al número de resultados posibles también se le denomina el espacio muestral, por ejemplo, para el lanzamiento de un dado los posibles resultados son $\{1,2,3,4,5,6\}$. La probabilidad de que uno de los resultados del espacio muestral se presente es 1 y lo denotamos de la siguiente forma: $\sum P(E_i)=1$

Existen básicamente 3 formas de cálculo de probabilidades:

- ✓ Modelos de frecuencia relativa o a-posteriori, basados en información histórica
- ✓ Modelo clásico o a priori, basado en posibles resultados
- ✓ Modelos subjetivos, cuando asignamos una probabilidad a algo que no sucedido

FRECUENCIA RELATIVA

Este modelo utiliza los datos de las observaciones que se han registrado y, en base a estos, calcula la frecuencia con la que se ha presentado dicho evento y, en base a esta frecuencia, se calcula la probabilidad.

Figura 3

Fórmula de frecuencia relativa

$$P(E) = \frac{Numero\ de\ veces\ que\ ha\ ocurrido\ el\ evento}{Número\ total\ de\ observaciones}$$

Este modelo es también llamado a posteriori, dado que la data ya ha sido recabada. Por ejemplo, utilizando el dataset BrCa que tiene 2982 observaciones, se ha estimado la cantidad de observaciones de la variable "size" y su probabilidad relativa en base a los distintos niveles.

Tabla 1

Ejemplo Probabilidad Relativa del dataser BrCa

size	Cuenta	Probabilidad Relativa
<=20 mm	1387	0.465
>20-50 mm	1291	0.433
>50 mm	304	0.102

Este modelo puede presentar algunos problemas. Por ejemplo, de la fórmula podemos deducir que, si las observaciones no incluyen una o más posibilidades, será imposible calcular la probabilidad. Por otro lado, si se cuenta con pocas observaciones, los resultados pueden ser engañosos.

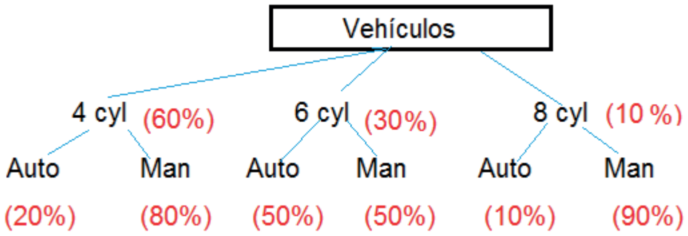
Adicionalmente es común que se pregunte por la probabilidad combinada de varias observaciones, para lo cual distinguiremos dos tipos de combinaciones:

La multiplicativa: Denotada por: $P(A \cap B) = P(A) * P(B)$

Por ejemplo, del dataset mtcars, se desea obtener la probabilidad de sacar un vehículo de 4 cyl que sea automático.

Figura 4

Distribución de probabilidades de variables cyl y am en mtcars



Entonces P(4cyl) * P(automático) =0.6*0.2=12%

Por otro lado, las probabilidades aditivas resultan de la suma de dos probabilidades, es decir, al investigador le da igual que cualquiera de los dos resultados ocurran. La probabilidad aditiva de A o B denotado como P(A U B) y su fórmula es: $P(A \cup B) = P(A)+P(B)-P(A\cap B)$

Por ejemplo: La probabilidad de sacar un as o una de las 13 cartas de corazones sería P(as U corazones) =(4/52)+(13/52)-(1/52) =16/52

De acuerdo con Webster (2017) los eventos A y B no necesariamente son excluyentes, es decir si ambos pueden ocurrir al mismo tiempo caeremos en doble conteo, en cuyo caso restaremos las posibilidades de este doble conteo ya que existe una carta que es as de corazones. Pero si los eventos son excluyentes entonces la fórmula quedaría así: $P(A \cup B) = P(A)+P(B)$

Probabilidad Condicional:

Con frecuencia se desea determinar la probabilidad de algún evento (A), dado que antes otro evento ya haya ocurrido (B) y se denota como P(A|B) y su fórmula es:

Figura 5

Probabilidad condicional

$P(A \mid B)=P(A\cap B)/P(B) =P(A)*(P(B \mid A))/P(B)$

Donde para el caso de condiciones excluyentes

$P(A \cap B)=P(A)*P(B|A)$

La probabilidad condicionada permite calcular la probabilidad de la intersección de dos sucesos, es decir, la probabilidad de que se den ambos sucesos A y B, de esta forma, la probabilidad de que tanto A como B ocurran es igual a la probabilidad de que A ocurra dado que B haya ocurrido multiplicado por la probabilidad de que B ocurra. En este caso, la probabilidad de que A ocurra no está afectada por la ocurrencia o no ocurrencia de B y se dice que los dos sucesos son independientes.

Por ejemplo, calcular cual es la probabilidad de obtener una J dado que sabemos que es una figura F, denotado como P(J|F)

Aplicando la fórmula anterior tendríamos

P(J) = 4/52 dado que existen 4 Js

P(F|J) =1 porque todas las J son figuras

P(F) = 12/52 porque existen 12 figuras.

Por lo tanto, P(J|F)=(4/52)*1/(12/52)= 4/12

Este es el concepto de probabilidad condicional del evento A dado que se conoce que el evento B ya ocurrió.

3.2) Teorema de Bayes

Supongamos que tenemos un conjunto completo de sucesos Ai , i = 1, . . . , n y un suceso B cualquiera del espacio muestral. A veces es necesario conocer la probabilidad de uno de los sucesos Aj condicionada a que haya ocurrido B. Esto se puede hacer por el Teorema de Bayes.

Veamos un ejemplo concreto:

Una fábrica tiene 2 equipos que producen la misma mercadería, la maquina 1 es más nueva y por lo tanto trabaja más rápido y produce el 60% de la producción, la maquina 2 es más antigua y produce el 40% restante. La máquina 1 además tiene apenas un 2% de producto descartado, mientras que máquina 2 tiene un 4% de producto descartado. De esta descripción podemos concluir lo siguiente tabla

Tabla 2

Cálculo de probabilidad del ejemplo

Probabilidad	Fórmula
Máquina 1 Producto OK	P (M1 ∩ OK)=P(A)*P(OK | A) =0.6*0.98=0.588
Maquina 1 Producto descartado	P(M1 ∩ Des)=P(A)*P(Des | A) =0.6*0.02=0.012
Maquina 2 producto Ok	P (M2 ∩ OK)=P(M2)*P(OK | M2)=0.4*0.96=0.384
Máquina 2 Producto descartado	P(M2 ∩ Des)=P(M2)*P(Des | M2)=0.4*0.96=0.016

Entonces conociendo en que máquina fue producido un ítem podemos obtener su probabilidad que este sea OK o Des. Sin embargo, es posible que deseamos hacer el análisis inverso: Dado un ítem que sabemos ha sido descartado, cual es la probabilidad de que haya sido producido en la maquina 1. Para lo cual aplicamos la fórmula de probabilidad condicional.

Figura 6

Fórmula de Probabilidad Condicional

P(A ┤| B)=P(A∩B)/P(B) =P(A)*(P(B | A))/P(B)

Que en nuestro ejemplo sería:

Pero el problema es que P(Des) no se conoce. Aquí entra el teorema de Bayes que establece:

Figura 7

Teorema de Bayes

P(B)=P(A)*(P(B | A))/P(A|B) P(B)=P(A ∩B)/(P(A ∩D)+P(B∩A))

Que en nuestro ejemplo sería

P(Des)= 0.012/(0.012 +0.016) = 0.429

Teorema de Bayes

Describe el teorema usos y ejemplos

Enlace

Definición de los términos citados en la Clase

Probabilidad	Probabilidad significa posibilidad. Es una rama de las matemáticas que estudia la ocurrencia de un evento aleatorio. El valor se expresa de cero a uno. La probabilidad se ha introducido en las matemáticas para predecir la probabilidad de que ocurran eventos. El significado de probabilidad es básicamente el grado en el que es probable que algo suceda. Esta es la teoría básica de la probabilidad, que también se utiliza en la distribución de probabilidad, donde aprenderá la posibilidad de resultados para un experimento aleatorio. Para encontrar la probabilidad de que ocurra un solo evento, primero debemos saber el número total de resultados posibles.
Probabilidad acumulativa	La probabilidad acumulada se refiere a la probabilidad de que el valor de una variable aleatoria se encuentre dentro de un rango determinado. Por ejemplo, $\Pr(a \leq X \leq b)$ Donde X es una variable aleatoria y a y b son los límites del rango. Con frecuencia, se utiliza para calcular la probabilidad de que una variable aleatoria sea menor o igual a un valor especificado: $\Pr(X \leq b)$
Probabilidad clásica a-priori	La probabilidad a priori, también conocida como probabilidad clásica, es una probabilidad que se deduce del razonamiento formal. En otras palabras, la probabilidad a priori se deriva del examen lógico de un evento. La probabilidad a priori no varía de persona a persona (como lo haría una probabilidad subjetiva) y es una probabilidad objetiva.
Probabilidad condicional	Las probabilidades condicionales se escriben como P(A|B), que puede leerse como “la probabilidad de que A ocurra DADO que b ha ocurrido”. Si conocemos probabilidades como P(A), P(B) y P(A|B), podemos hallar otras probabilidades como P(B|A).
Probabilidad Total	La regla de probabilidad total (también conocida como ley de probabilidad total) es una regla fundamental en estadística relacionada con las probabilidades condicionales y marginales. La regla establece que, si se desconoce la probabilidad de un evento, se puede calcular utilizando las probabilidades conocidas de varios eventos distintos.
Teorema de Bayes	El teorema de Bayes es una fórmula matemática sencilla que se utiliza para calcular probabilidades condicionales. Ocupa un lugar destacado en los enfoques subjetivistas o bayesianos de la epistemología, la estadística y la lógica inductiva. Los subjetivistas, que sostienen que la creencia racional está regida por las leyes de la probabilidad, se apoyan en gran medida en las probabilidades condicionales en sus teorías de la evidencia y sus modelos de aprendizaje empírico. El teorema de Bayes es fundamental para estas empresas, tanto porque simplifica el cálculo de las probabilidades condicionales como porque aclara características significativas de la posición subjetivista. De hecho, la idea central del teorema —que una hipótesis se confirma por cualquier conjunto de datos que su verdad haga probable— es la piedra angular de toda la metodología subjetivista.

Referencias Citadas

Economipedia, Población estadística: Qué es, tipos y ejemplos, recuperado de https://economipedia.com/definiciones

Webster A., (2000), Estadística aplicada a los negocios, Irwin Professional Publishing

Wasserman, L. (2010). All Statistics: A Concise Course in Statistical Inference. Springer Publishing Company.

Otzen T. Manterola c. (2017) ,Técnicas de Muestreo sobre una Población a Estudio, Int. J. Morphol. vol.35 no.1 Temuco, recuperado de http://dx.doi.org/10.4067/S0717-95022017000100037

Conceptos de estadística

Alvarez H., Conceptos de estadística, Universidad Santiago de Compostela, Recuperado de https://www.usc.gal/genp/docencia/ConceptosDeEstadistica.pdf

Profundización Clase 1.

Cálculo de probabilidades

Explica como calcular varios tipos de probabilidades en R

Enlace

La excelencia no se improvisa

síguenos

MDASC - M5 - A2