Fundamentos de la estadística

Unidad 1

Rodrigo Barrera

Proceso estadístico

¿Qué es la estadística?

Estadística

El término estadística se deriva de la palabra latina status (que significa “estado”). Los primeros usos de la estadística implicaron la recopilación de datos y la elaboración de gráficas, para describir diversos aspectos de un estado o de un país. En 1662 John Graunt publicó información estadística acerca de los nacimientos y los decesos. Al trabajo de Graunt siguieron estudios de tasas de mortalidad y de enfermedad, tamaño de poblaciones, ingresos y tasas de desempleo.

¿Qué es una variable?

La RAE entrega la siguiente definición;

- Que varia o puede varias
- Que está sujeto a cambios frecuentes o probables

Variable aleatoria

Una variable aleatoria es una aplicación que asocia cada suceso elemental un número real. Conviene definir este concepto con precisión, puesto que es la idea fundamental que permite dar un tratamiento riguroso los fenómenos aleatorios.

Variable aleatoria: Sea $(\Omega, \mathcal{A}, \mathcal{P})$ un espacio probabilístico asociado un experimento aleatorio. Una variable aleatoria es una aplicación definida sobre que toma valores en el conjunto de los números reales $X : \Omega \to \mathbb{R} | a \to X(a) \in \mathbb{R}$ si se verifica que $\forall x \in \mathbb{R}$ el conjunto $\{a \in \Omega | X(a) \leq x\} \in \mathcal{A}$

Variable aleatoria

Discreta

Una variable aleatoria puede tomar un conjunto de valores finito, infinito nume rable una infinidad no numerable de valores reales. Será discreta continua, por tanto, según sea su contradominio imagen.

Variable aleatoria discreta: La variable aleatoria es discreta cuando toma un conjunto finito infinito numerable de valores reales.

La variable aleatoria definida por el número de caras que salen cuando se lanzan dos monedas es una variable discreta; su contradominio está formado por los puntos del conjunto: $\{0, 1, 2\}$

Variable aleatoria

Continuas

Hay variables aleatorias que tienen por imagen toda la recta real algún intervalo de la misma $(a, b), [a.b), (a, b), (-\infty, a), [a, \infty)$, ... este tipo de variable aleatoria, que no toma valores aislados, le llamamos varia ble aleatoria continua, esto es:

Variable aleatoria continua: La variable aleatoria es continua cuando toma valores en una escala continua.

Tipos de variable

Las variables pueden clasificarse en diferentes tipos dependiendo de los valores a los que dan lugar. Esta clasificación es importante porque determinará el tipo de técnicas de análisis que pueden utilizarse para su estudio.

Las variables se clasifican en dos grandes grupos: las variables cualitativas o categóricas y las variables cuantitativas. Las variables cualitativas no toman valores numéricos y pueden clasificarse en un determinado número de categorías o estados (sexo, nivel de estudios, etc.)

Tipos de variable

Las variables cuantitativas toman valores numéricos (edad, número de hijos, nivel de colesterol, etc.). Las variables cualitativas se clasifican a su vez en variables cualitativas ordinales y no ordinales, dependiendo de si sus categorías o estados pueden ordenarse o no. Así, sexo sería una variable cualitativa no ordinal, mientras que nivel de estudios sería una variable cualitativa ordinal. Las variables cuantitativas se subdividen a su vez en variables cuantitativas discretas y cuantitativas continuas, dependiendo de si toman un número finito o infinito numerable de valores (discretas) o infinito no numerable (continuas).

Tipos de variable

A las variables cuantitativas continuas también se las llama variables de razón o intervalo. Así, por ejemplo, las variables número de hijos, número de ingresos en un hospital, etc., serían variables cuantitativas discretas (obsérvese que entre 0 y 1 hijo no hay valores posibles), mientras que nivel de colesterol, edad o nivel de ácido úrico serían variables continuas, puesto que cualquier valor entre dos dados es posible (toman valores en un intervalo).

Medidas de tendencia central

Las medidas de tendencia central proporcionan información sobre la posición o localización de los datos observados. Entre las medidas de este tipo se encuentran la media, la mediana o la moda.

Una alternativa al cálculo de la media, no sensible a observaciones atípicas o extremas, la constituye la mediana. El valor de la mediana, para un conjunto de datos, se obtiene de forma que deja el mismo número de observaciones a su izquierda que a su derecha.

Medias

- Media aritmética $\sum \frac{x_i}{n}$
- Media ponderada $\frac{\sum x_i w_i}{\sum w_i}$
- Media geométrica $(\prod x_i)^{\frac{1}{n}}$
- Media armónica $\frac{n}{\sum_i^n \frac{1}{x_i}}$

Medidas de tendencia central

Mediana

Una desventaja de la media es su sensibilidad a cada valor, de tal forma que una puntuación excepcional puede afectarla de manera drástica. La mediana resuelve, en gran medida, esa desventaja. La mediana es un “valor intermedio”, ya que la mitad de los valores de los datos están por debajo de la mediana y la otra mitad por arriba de ella. La siguiente definición es más precisa.

Medidas de tendencia central

Moda

La moda se define, para un conjunto de datos, como el valor más frecuente, es decir, el valor que más veces se repite.

- Cuando dos valores se presentan con la misma frecuencia y ésta es la más alta, ambos valores son modas, por lo que el conjunto de datos es bimodal.
- Cuando más de dos valores se presentan con la misma frecuencia y ésta es la más alta, todos los valores son modas, por lo que el conjunto de datos es multimodal.
- Cuando ningún valor se repite, se dice que no hay moda.

Medidas de tendencia no central

En estadística descriptiva, las medidas de posición no central permiten conocer otros puntos característicos de la distribución que no son los valores centrales. Entre las medidas de posición no central más importantes están los cuantiles.

El término cuantil fue usado por primera vez por Maurice Kendall en 1940. El cuantil de orden p de una distribución (con $0 < p < 1$) es el valor de la variable x_p que marca un corte de modo que una proporción p de valores de la población es menor o igual que x_p. Por ejemplo, el cuantil de orden 0.36 dejaría un 36% de valores por debajo y el cuantil de orden 0.50 se corresponde con la mediana de la distribución.

Medidas de dispersión

Las medidas de dispersión indican el grado de concentración de los valores de la variable alrededor de una medida de posición central, dando, a su vez, una idea de la representatividad de esta medida de centralización como resumen global de la variable.
Las medidas de dispersión más utilizadas son: la varianza, la desviación típica y el coeficiente de variación.

Medidas de dispersión

- Rango
  - $R = Max_X - Min_X$
- Desviación estandar $\sigma = \sqrt{\frac{\sum(x_i - \hat{x})^2}{n}}$
- Rango intercuartílico $R_Q = Q_3 - Q_1$

Medidas de dispersión

Coeficiente de variación

La desviación típica proporcionaba una medida resumen de las distancias de cada dato a la media (desviaciones) en las mismas unidades que la variable original y, por tanto, depende de dichas unidades de medida. ¿Son comparables las desviaciones típicas de dos conjuntos de datos? ¿Puede afirmarse, en general, que a mayor desviación típica mayor dispersión?

Medidas de dispersión

Coeficiente de variación

Las unidades de la desviación estándar son las mismas que las unidades de los datos originales, es más fácil comprender la desviación estándar que la varianza. Sin embargo, esta misma propiedad dificulta comparar la variación de valores tomados de distintas poblaciones. Como el resultado es un valor libre de unidades de medida específicas, el coeficiente de variación resuelve esta desventaja.

$$CV = \frac{S}{\overline{x}} \tag{1}$$

Desviación media

- $D_{\overline{x}} = \frac{\sum |x_i - \overline{x}| f_i}{n}$
- $D_{M_e} = \frac{\sum |x_i - M_e| f_i}{n}$
- $D_{M_o} = \frac{\sum |x_i - M_o| f_i}{n}$

Tablas de frecuencia

La estadística exploratoria recomienda comenzar por el análisis de la estructura de los datos. Se clasifican éstos de acuerdo con la modalidad del carácter que pertene ce cada uno de los individuos se ordenan, anotando sus resultados en una tabla. La ordenación de los datos en la tabla, acompañados de las frecuencias corres pondientes, es lo que se llama distribución de frecuencias.

Tablas de frecuencia

# de estrellas	Frecuencia	Frecuencia relativa	Frecuencia relativa acumulada
2	8	0,095	
3	45	0,536	
4	23	0,274	
5	8	0,095	

Tablas de frecuencia

- Encuentre el máximo y el mínimo en el conjunto de datos.
- Seleccione el número de subintervalos, de igual largo, resguardando que no se produzca sobreposición. Estos se denominan intervalos de clases. La regla de Sturges sugiere $c = 1 + \log_2(N)$
- Cuante cuantas observaciones pertenecan a cada intervalo de clase. Asi se obtiene la frecuencia de la clase.
- Determine la frecuencia relativa dividiendo la frecuencia de la clase por el número totla de observaciones.

Covarianza

La covarianza es el valor que refleja en qué cuantía dos variables aleatorias varían de forma conjunta respecto a sus medias.

Nos permite saber cómo se comporta una variable en función de lo que hace otra variable. Es decir, cuando X sube ¿Cómo se comporta Y? Así pues, la covarianza puede tomar los siguiente valores:

Covarianza (X, Y) es menor que cero cuando X sube e Y baja. Hay una relación negativa.

Covarianza (X, Y) es mayor que cero cuando X sube e Y sube. Hay una relación positiva.

Covarianza

$$Cov(X, Y) = \frac{\sum_1^n (x_i - \overline{x})(y_i - \overline{y})}{n} \tag{2}$$

Muestreo

Conceptos

- **Población objetivo:** es la coleccion completa de todas las unidades que se quieren estudiar.
- **Muestra:** es un subconjunto de la población
- **Unidad de muestreo:** es el objeto a ser seleccionado en la muestra que permitirá el acceso a la unidad de observación.
- **Unidad de observación:** es el objeto sobre el que finalmente se realiza la medición
- **Variable de interés:** es la característica propia de los individuos sobre la que se realiza la inferencia para resolver los objetivos de la investigaci´on

Muestreo

Conceptos

Todo procedimiento de muestreo probabilístico requiere de un dispositivo que permita identificar, seleccionar y ubicar a todos y cada uno de los objetos pertenecientes a la población objetivo y que participarán en la selección aleatoria. Este dispositivo se conoce con el nombre de marco de muestreo. En investigaciones por muestreo se consideran dos tipos de objetos:

- **Elementos:** las unidades básicas e individuales sobre las que se realiza la medición.
- **Conglomerado:** agrupación de elementos cuya característica principal es que son homogéneos dentro de sí, y heterogeneos entre sí.

Tipos de muestreo

Muestreo probabilístico

Las distintas técnicas de muestreo se clasifican en probabilísticas o aleatorias y no probabilísticas. La diferencia estriba fundamentalmente en que, en las primeras, cada uno de los individuos o elementos de la población tiene una probabilidad conocida y distinta de 0 de ser incluido en la muestra, mientras que en las últimas esta cuestión se desconoce. La principal consecuencia del conocimiento o desconocimiento de la probabilidad de inclusión de un individuo o elemento en la muestra es que las técnicas probabilísticas serán las únicas que harán posible tanto la cuantificación del error muestral como la aplicación de las técnicas inferenciales (construcción de intervalos de confianza y contrastes de hipótesis).

Tipos de muestreo

Muestreo aleatorio simple

El muestreo aleatorio simple es un método de selección de n unidades sacadas de N, de tal manera que cada una de las muestras tiene la misma probabilidad de ser elegida.

En la práctica una muestra aleatoria simple es extraída de la siguiente forma:

Se numeran las unidades de la población del 1 al N, y por medio de una tabla de números aleatorios o colocando los números 1 a N en una urna, se extraen sucesivamente n números. Las unidades que llevan estos números constituyen la muestra.

Tipos de muestreo

Muestreo estratificado

En este tipo de muestreo, la población de N unidades es dividida en subpoblaciones de $N_1, N_2, ..., N_L$ unidades, respectivamente. Estas subpoblaciones no se sobreponen y juntas forman la totalidad de la población, por lo que

$N_1 + N_2 +, ..., +N_L = N$

Las subpoblaciones son llamados estratos. Una vez que han sido determinados los estratos, se saca una muestra de cada uno, la extracción se realiza de forma independiente en cada estrato. Los tamaños de la muestra dentro de los estratos son representados por $n_1, n_2, ..., n_L$.

Tipos de muestreo

Muestreo estratificado

Si se toma una muestra aleatoria simple en cada estrato, el procedimiento completo es conocido como muestreo estratificado aleatorio.
La estratificación es una técnica común. Hay muchas razones para realizarla; las principales son:

- Si se desea cierta precisión en alguna subdivisión, es necesario tratarla como si fuera una “población” por sí sola.
- La estratificación puede dar lugar a una ganancia en precisión de los estimadores de la población. Esto ocurre cuando una población heterogénea es dividida en subpoblaciones cada una de las cuales es internamente homogénea.

Tipos de muestreo

Muestreo sistemático

- Conseguir un listado ordenado de los N elementos de la población.
- Determinar el tamañoo muestral n.
- Definir el tamaño del salto sistemático k dado por $k = N/n$
- Elegir un número aleatorio δ entre 1 y k (δ =arrranque aleatorio). Este numero permite obtener la primera unidad muestral.

Tipos de muestreo

Muestreo por conglomerados

El muestreo por conglomerados consiste en dividir la población en conjuntos sin solapamiento, y exhaustivos. De manera que cada uno de ellos represente toda la variabilidad posible.
En el muestreo por conglomerados, por tanto, lo que hacemos es crear grupos más pequeños de una población, los cuales tengan todas las características de esta.
Así, una vez los tenemos, podemos elegir algunos de ellos como muestra y analizarlos de forma más sencilla.

Factores de expansión

Es un concepto relacionado con la probabilidad de selección y se interpreta como la cantidad de unidades en la población que representa una unidad en la muestra, llámese personas, viviendas, áreas económicas o agrícolas, etcétera, dicho factor permite dar conclusiones sobre la población total.

DEFF

Un efecto de diseño (DEFF) es un ajuste realizado para hallar el tamaño de la muestra de una encuesta, debido a que un método de muestreo (por ejemplo, muestreo por conglomerados o muestreo estratificado) da lugar a tamaños de muestra mayores (o intervalos de confianza más amplios) de lo que cabría esperar con un muestreo aleatorio simple (MAS). El DEFF indica la magnitud de estos aumentos.

DEFF

El efecto del diseño es la relación entre la varianza real y la varianza esperada con el MAS. Puede expresarse más sencillamente como el tamaño real de la muestra dividido por el tamaño efectivo de la muestra (el tamaño efectivo de la muestra es lo que se esperaría si se utilizara MAS). Por ejemplo, supongamos que utiliza el muestreo por conglomerados. Un DEFF de 2 significa que la varianza es el doble de lo que cabría esperar con MAS. También significa que si utilizara el muestreo por conglomerados, tendría que utilizar el doble del tamaño de la muestra.