Apuntes Stats

Oskar Denis Siodmok

23 de noviembre de 2020

1 Introducción

1.1. Conceptos Básicos

- Individuos o elementos: Contienen la información a estudiar.
- Población: Conjunto de individuos o elementos que presentan la variable a estudiar.
- Muestra: Subconjunto representativo de una población.
- Variables: Propiedades de los elementos de la población que tendrán sus respectivos valores
- Clases: Conjunto de valores que cumplen una propiedad. Por definición un valor solo puede pertenecer a una clase (por ejemplo, intervalos en variables contínuas).
- Parámetro: Será una función que operará con las diferentes variables y valores sobre la población con un propósito.
- Estadístico: Será una aproximación al parámetro a partir de una muestra.

1.1.1. Variables Estadísticas

Las variables estadísticas, ya definidas en el anterior apartado, se denotarán mediante una letra mayúscula (por lo general con X o Y). Podrán tomar cualquier valor de cualquier conjunto. El dominio de la variable será el conjunto de todos los posibles valores de dicha variable.

1.1.2. Tipos de variables

- Variables Cuantitativas: Se expresan en cantidades numéricas o cualquier otro sistema que se pueda ordenar. A su vez se clasifican en:
  - Variables Discretas: Toman valores concretos de conjuntos finitos o infinitos.
  - Variables Continuas: Toman valores de conjuntos infinitos y no concretos (el dominio son valores continuos, como todos los reales en un intervalo).

  Para muchas variables resulta complicado distinguir el tipo. Por ejemplo, aunque la altura matemáticamente sea continua, en la vida real nadie va a determinar una altura por encima de 2 decimales.
- Variables Cualitativas: No se pueden medir, solo clasificar. Un tipo concreto de este tipo de variables son las *Variables Ordinales*, las cuales pese a no tener un valor numérico si que pueden tener relaciones de orden.

1.1.3. Representación de datos

Esta se puede realizar de varias formas:

- Tablas y Gráficos: representan información de forma rápida y visual.
- Medidas Descriptivas: describen la información de forma numérica.

1.2. Tablas de Frecuencias

Se pueden realizar sobre cualquier conjunto de datos y sobre cualquier variable. Por ejemplo, dada una población de n indivíduos que presenta la variable X se obtienen las clases $c = \{c_1, c_2, \ldots, c_k\}$ posibles. En este caso, n_i hará referencia al número de observaciones para $i \in \{1, \ldots, k\} \subseteq \mathbb{N}$. De esta forma, $n = \sum_i n_i$ será el número total de observaciones de nuestra variable, independientemente de la clase de cada observación. Por otro lado, la frecuencia relativa de una variable representará la frecuencia de una variable (n_i) sobre 1 y se calculará como $\frac{n_i}{n}$.

Clase c_i	c_1	c_2	$\ldots$	c_k	
Freq. Absoluta n_i	n_1	n_2	$\ldots$	n_k	$n = \sum_i n_i$
Freq. Relativa f_i	f_1	f_2	$\ldots$	f_k	$f = \sum_i f_i i$

Se podrían añadir las frecuencias acumuladas, las cuales se van sumando a los datos anteriores. Siendo N_i la frecuencia absoluta acoumulada y F_i la frecuencia relativa acumulada:

Clase c_i	c_1	c_2	$\ldots$	c_k	
Freq. Absoluta n_i	n_1	n_2	$\ldots$	n_k	$n = \sum_i n_i$
Freq. Relativa f_i	f_1	f_2	$\ldots$	f_k	$f = 1$
F. abs. acum. $N_i = \sum_{j=1}^{i} n_j$	N_1	N_2	$\ldots$	N_k	$N = N_k = n$
F. rel. acum $F_i = \sum_{j=1}^{i} f_j$	F_1	F_2	$\ldots$	F_k	$F = F_k = 1$

Si la variable es cualitativa entonces las clases serán nominales. Si la variable es discretas las clases serán valores numéricos dentro del rango y si es continua serán intervalos $(l_{i-1}, l_i]\ \forall i$. Si las clases son intervalos habrá varios parámetros y conceptos de interes:

- Amplitud del intervalo: $a_i = l_i - l_{i-1}$.
- Marca de clase: Será el valor representativo del intervalo. Por ejemplo, el punto medio: $c_i = \frac{l_i + l_{i-1}}{2}$. Podrá representarse en la segunda columna o fila de la tabla.
- Número de intervalos: Se realizará mediante aproximación pues se pueden plantear los intervalos que se quiera. Una elección típica es:

$$k = \begin{cases} \sqrt{n} & , n \text{ no es muy grande} \\ 1 + \log_2(n) & \end{cases}$$

- Normalización: Cuando el intervalo tiene muchos decimales conviene redondear hacía arriba para que los datos sean más legibles. Notar que si se redondea a la baja se altera el número de intervalos, por lo cual no conviene.
- Límites de los intervalos: Por definición de clase, un valor de una variable solo puede pertenecer a una clase. Debido a ello hay que tener en cuenta los límites de los intervalos prestando atención a que ningún par de clases ccontenga valores repetidos.

1.3. Gráficos

1.3.1. Diagrama de Barras

En el eje X se representan las clases y en el Y las frecuencias absolutas o relativas (ya que son proporcionales la escala se mantendría). Busca en google como son que no tengo ningún conjunto de datos interesante. Si el diagrama de barras se hace a partir de una variable continua y sus intervalos entonces será un histograma. Ante intervalos de mimas amplitud, el gráfico resultante será proporcional a su correspondiente historgrama de frecuencias relativas, esta característica se tendrá que cumplir siempre. Otra vez, hay que tener cuidado con el número de intervalos seleccionados para que el histograma muestre la información de forma óptima. Según la forma del histograma puede ser:

- Distribución unimodal simétrica: Los datos tienen forma de campana de gauss.
- Distribución bimodal simétrica: Hay cierta simetría respecto al eje que divide los datos por la mitad pero no es necesariamente gaussiana.

- Distribución asimétrica a la derecha: Hay escasez de datos a la derecha.
- Distribución asimétrica a la izquierda: Hay escasez de datos a la izquierda.

1.3.2. Diagrama de Sectores

Se divide un círculo en sectores proporcionales a las frecuencias. El ángulo de cada clase se calcularía con una simple regla de tres:

$$\frac{n}{n_i} = \frac{2\pi}{x_i}$$

1.4. Medidas Descriptivas de Centralización

Las medias descpritivas de centralización aproximan el valor central respecto al cual los datos se ordenan. De entre estas medidas destacan la media, moda y mediana.

1.4.1. Media Aritmética

Esta medida será la suma total de todos los datos dividida por el número total de observaciones.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Donde x_i hará referencia a cada observación. En el caso de que los datos se den en forma de tabla de frecuencia:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i f_i$$

En el caso de ser una variable contínua solo habrá que cambiar x_i por el número de observaciones dentro del intervalo, o sea, c_i. Dependiendo de la a_i, o sea, la amplitud de los intervalos, habrá una mayor o menor pérdida de precisión. Por otro lado, la linealidad de la media será una función $\bar{y} = a + b\bar{x}$ de la misma forma que $Y = a + bX$.

Este parámetro presenta una serie de inconvenientes:

- No se puede calcular con variables cualitativas o nominales.
- Es muy sensible a los valores extremos.
- Centraliza los datos de forma subóptima ante distribuciones de datos asimétricas.
- Ante variables contínuas y tablas de frecuencia su valor depende de los intervalos.
- Ante una variable discreta, la media puede no pertenecer al dominio de la variable.

1.4.2. Moda

Se trata del valor mas frecuente dentro de las observaciones. En el caso de que se trabaje con una variable continua se considerará el intervalo modal como al intervalo de mayor frecuencia. Este parámetro tiene la ventaja de ser muy fácil de calcular pero puede no ser única.

1.4.3. Mediana

La mediana será el parámetro que divida al número de observaciones al 50 % por los dos lados, o sea, será la observación central de todas las observaciones. Ante n observaciones:

$$M_e = \begin{cases} x_{(n+1)/2} & \iff n \text{ impar} \\ \frac{x_{n/2} + x_{(n/2)+1}}{2} & \iff n \text{ par} \end{cases}$$

Si los datos se ordenan en una tabla de frecuencia, M_e será el primer valor con $F_i \geq 1$ o con $N_i \geq n$. Entre las propiedades de este parámetro destaca:

- No le afectan las observaciones extremas pues solo depende del orden de los valores de la variable. Debido a esto la mediana se usa mucho en distribuciones asimétricas.

- No se puede calcular con variables cualitativas o nominales, al igual que la media.
- Su mayor defecto es que tiene propiedades muy complicadas en las que se profundizará más adelante. Debido a esta abstracción es difícil de usar en inferencia estadística.

1.5. Medidas Descriptivas de Posición

Una posición genérica es un cuantil, al igual que la mediana informa sobre el valor medio de todas las observaciones, un cuantil informa sobre la posición en concreto. De esta forma, $p \in \{n \in \mathbb{R} : 0 < n < 100\}$ será un percentil e informará sobre el porcentaje de datos que tendrá por debajo de si mismo denotado como P_p. A partir de estos percentiles se pueden definir los cuartiles.

- Primer Cuartil: $Q_1 = P_{25}$.
- Segundo Cuartil: $Q_2 = P_{50} = M_e$.
- Tercer Cuartil: $Q_3 = P_{75}$

Estos ofrecen información muy genérica. Por ejemplo, para un conjunto de notas, si $Q_3 = 6$ sabremos que un 7.79 se encuentra entre el 25 % mejores notas de la clase. De la misma forma se definirán los deciles como P_p con $p \in \{10n : n \in [1, 9] \subseteq \mathbb{N}\}$.

1.6. Medidas Descriptivas de Dispersión

Ante datos muy dispersos las medidas de centralización pierden su efectividad. Para esto existen las medidas de dispersión las cuales trabajarán con la sensibilidad de los datos.

1.6.1. Rango

Se define como $R = max\{x_1, \ldots, x_n\} - min\{x_1, \ldots, x_n\}$. Este parámetro es fácil de calcular y comparte unidades con los datos. Entre las desventajas de esta medida destacan:

- No se utilizan todos los datos (solo 2).
- Puede variar mucho ante datos extremos.
- Ante el aumento de observaciones el rango o aumenta o se queda igual, nunca baja.

De la misma forma se puede definir el rango intercuartílico como $RQ = Q_3 - Q_1$. Tanto el rango intercuartílico como el rango general son de gran importancia para realizar diagramas de cajas.

1.6.2. Diagramas de Cajas

Este sintetizará múltiples parámetros de posición y, a su vez, de dispersión. Presentan información sobre la simetría y los datos atípicos de la dispersión. El proceso de creación de un diagrama de cajas es bastante más específico que los métodos del resto de tipo de gráficos. El proceso se puede sintetizar en múltiples pasos:

- Primero se realiza un rectángulo que pasa por los cuartiles. Este rectángulo contendrá una línea perpendicular que lo dividirá en dos y representará la mediana. De esta forma, dentro de la caja quedaría representado el 50 % de los datos.
- A continuación se representarán 2 líneas perpendiculares al eje de los datos a $1.5RQ$ de los 2 cuartiles, tanto a la derecha como a la izquierda.
- Para terminar la representación de estos límites se conectarán las líneas a la caja con línea punteada. Así, cualquier dato fuera de estos límites se podría considerar atípico y se representaría como puntos concretos.

1.6.3. Desviación Media

Otra forma bastante eficiente de calcular que tan atípico es un valor x_i es con la diferencia de este con la media. De esta forma $x_i - \bar{x}\ \forall i$ serían las desviaciones de cada valor respecto a la media aritmética. Una forma de condensar toda esta información sobre cada observación podría ser con la media de todas las desviaciones:

$$D_m = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i - n\bar{x}$$

El problema aparece teniendo en cuenta que en función de si la observación está a la derecha o a la izquierda el valor de la desviación es o positivo o negativo, de forma que la desviación media se acercaría a 0, o sea, $D_m \approx 0$. Solucionar este problema es tan sencillo como usar el valor absoluto de la diferencia para cada dato. Finalmente, la desviación media se definirá como:

$$D_m = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

1.6.4. Varianza y Desviación Típica

La esencia será la misma que la de D_m pero para forzar el símbolo positvo se usará el cuadrado de la diferencia. Esto implicará que el parámetro será mucho más sensible.

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Hay que tener en cuenta que la magnitud de este parámetro será el cuadrado de la magnitud de los datos. Para solucionar esto aparece el parámetro de *Desviación típica*, la cual se definirá como la raiz de la varianza:

$$s = \sqrt{s^2}$$

Además, con un simple desarrollo se comprueba que:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

De la misma forma que la media, para una tabla de frecuencia:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - \bar{x}^2 = \sum_{i=1}^{k} x_i^2 f_i - \bar{x}^2$$

Entre las propiedades de estas medidas destaca:

- Las dos son sensibles a los valores extremos.
- Se comprueba que para s mínimo el 75 % de los datos se encuentra en el intervalo $(\bar{x} - 2s,\ \bar{x} + 2s)$.
- No se puede usar para variables nominales (de la misma forma que la media).

Por último, carbría destacar la cuasi-varianza la cual se dieferencia de la varianza por dividir el resultado entre $n - 1$ y no n. Este valor tiene mejores propiedades que la varianza para estimar por lo cual se usa mucho más en Inferencia Estadística.

$$s_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{ns^2}{n-1}$$

1.6.5. Tipificación

Este proceso consistirá en convertir una variable a una media de 0 y una desviación estándar $s_z = 1$. Esta nueva variable se llamará variable tipificada y se denotará como:

$$Z = \frac{X - \bar{x}}{s}$$

Tipificar a variables con estas propiedades es útil para poder hacer comparaciones entre diferentes conjuntos de datos con diferentes dominios

1.6.6. Coeficiente de Variación

$$CV = \frac{s}{\bar{x}} 100$$

Esta medida se usa para comparar variables entre 2 conjuntos de datos con diferente media o unidades. También sirve para determinar si la media es consistente con una varianza o para comprobar la variabilidad entre grupos de datos tomados por personas distintas. Todo esto se debe a que este parámetro se mide en porcentajes. Hay que tener cuidado co la tipificación, pues una media de 0 podría implicar una operación ilegal.

1.7. Medidas de Forma

Al estudiar un conjunto de datos también resulta interesante ver si estos de distribuyen siguiendo una simetría o no. Una vez conocido este valor vendría bien preguntarse si el histograma es más o menos apuntado, característica que se medirá a partir de la frecuencia de la normal. De esta forma, si una variable es continua simétrica y unimodal entonces la media, mediana y moda coincidirán.

Además, si se presenta una asimetría hay que considerar si es positiva o negativa, esto implicaría frecuencias más altas a la izquierda o a la derecha, respectivamente. Los diferentes parámetros de forma son:

- Momento de orden p para $p \in \mathbb{N}$: $\mu_p = \frac{1}{n} \sum_{i=1}^{n} x_i^p$.
- Momento central de orden p para $p \in \mathbb{N}$: $m_p = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^p$.
- Coeficiente de asimetría: $\gamma_1 = \frac{m_3}{m_2 \sqrt{m_2}} = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$.
- Coeficiente de curtosis o apuntamiento: $\gamma_2 = \frac{m_4}{s^4} - 3$.

Siendo este tercer parámetro el más importante, pues nos informa sobre la asimetría de forma directa. Si $\gamma > 0 \longrightarrow$ asimetría positiva y si $\gamma < 0 \longrightarrow$ asimetría negativa. De la misma forma, γ_2 cuantifíca que tan apuntada es una distribución. La referencia de este parámetro es el apuntamiento de la distribución gaussiana o normal para la cual $\frac{m_4}{s^4} = 3$. De esta forma, si $\gamma > 0$ la distribución será leptocúrtica (más apuntada que la normal) y si $\gamma < 0$ platicúrtica (más aplastada que la normal).

2 Datos Bivariantes

2.1. Distribución de dos variables

Dada una población estadística de n individuos y 2 variables X e Y se define:

- Frecuencia Total: Número total de indivíduos n.
- Frecuencia absoluta del par (x_i, y_j): Se refiere al número de observaciones total para cada observación de las 2 variables a la vez. Se denota por n_{ij}.
- Frecuencia relativa del par (x_i, y_j): Exactamente igual que las frecuencias relativas para una variables. Se denota como $f_{ij} = \frac{n_{ij}}{n}$.

2.2. Tablas de frecuencia

Estos pares de observaciones para las 2 variables se pueden organizar en tablas de frecuencia bivariantes. En estas tablas se pueden mostrar tanto frecuencias absolutas como relativas. En el caso de que las variables sean cualitativas, entonces la tabla se denominará como tabla de contingencia.

$X \setminus Y$	y_1	y_2	$\dots$	y_l	
x_1	n_{11}	n_{12}	$\dots$	n_{1l}	$n_{1.}$
x_2	n_{21}	n_{22}	$\dots$	n_{2l}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\dots$	n_{kl}	$n_{k.}$
	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.l}$	n

2.3. Distribuciones Marginales

Una frecuencia absoluta o relativa margianl se refiere al número de veces que se repite un x_i sin tener en cuenta el valor de Y, como si con una distribución de una solo variable se estuviera trabajando. De esta forma, en base a la tabla del apartado anterior:

$$n_{i.} = \sum_{j=1}^{l} n_{ij} \ , \ \ n_{.j} = \sum_{i=1}^{k} n_{ij} \ , \ \ f_{i.} = \frac{n_{i.}}{n} \ , \ \ f_{.j} = \frac{n.j}{n}$$

2.4. Distribuciones Condicionales

Una distribución de Y sabiendo que $X = X_i$ se denota como $(Y|X = x_i)$. La misma notación se usa para distribuciones de X con un Y específico.

$Y\|X = x_i$	y_1	y_2	$\dots$	y_l	
n_{ij}	n_{i1}	n_{i2}	$\dots$	n_{il}	$n_{i.}$

2.5. Variables Independientes

Dos variables son estadísticamente independientes entre ellas si la varianza de una de ellas no afecta a los valores de la otra. De esta forma, ocurre que la distribución de una de las variables condicionada por la otra no afecta a los valores de esos valores. Además, X, Y independientes $\iff f_{ij} = f_{i.}f_{.j}$.

2.6. Gráficos

Estos datos se pueden representar de diferentes formas, por ejemplo, si el dominio de una de las variables corresponde a solo 2 valores, entonces se podría hacer 2 representaciones gráficas de la segunda variable condicionada por los 2 valores de la primera. Son de interes los diagramas de barras, los cuales sepueden representar apilados o agrupados.

Para datos multivariables aparece un tipo de representación nueva: los diagramas de dispersión. Para datos bivariantes, por ejemplo, se usa cada pareja de observaciones (x_i, y_i) como un punto del plano. Una vez representados todos los puntos, se pueden buscar relaciones entre las variables de forma bastante cómoda.

2.7. Medidas Descriptivas de Dependencia Lineal

Dos variables serán dependientes linealmente cuando el aumento de una implica el aumento o disminución de la otra. Existen diferentes medidas para calcular esta dependencia.

2.7.1. Covarianza

Se define como:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

El razonamiento para esta cuenta es el mismo que el de la varianza para una variable. Ahora bien, se dirá que si las variables no están relacionadas entonces la covarizanda tiende a cero, en cambio, si la varianza es negativa la relación entre las variables será negativa. Haciendo un poco de álgebra para la expresión de la covarianza se puede llegar a que:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

Expresión que, computacionalmente, es mucho más óptima.

Finalmente, hay que tener en cuenta que una covarianza de 0 no implica necesariamente una independencia en los datos. Por ejemplo, si X depende de Y de forma parabólica, la covarianza se acercará a 0, pero los datos estarán claramente relacionados.

2.7.2. Vector de Medias y Matriz de Covarianza

Será una medida de interes el vector de medias, el cual, como indica, contendrá los valores de las medias de todas las variables.

Por otro lado, la matriz de covarianza se define como:

$$S = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$$

Donde, recordando, $s_{x_i}^2$ hace referencia a la varianza de una variable X_i.

2.7.3. Coeficiente de Correlación Lineal

Aunque la covarianza sea una buena medida para medir correlaciones, está se ve afectada por las unidades: si una variable es dependiente del otra al 100 % pero están relacionadas por una recta de mucha pendiente, entonces la covarianza será muy grande. Para evitar esto hace falta platear una nueva medida que informe sólamente del nivel de dependencia de las variables, esta es el coeficiente de correlación lineal, que se define como:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

De esta forma, si la correlación es absoluta, $r_{xy} = (1 \vee -1)$. En caso contrario, si la correlación lineal es nula, $r_{xy} \to 0$. Hay que tener en cuenta que, en el caso de que una de las variables sea constante, independientemente del valor de la otra, la correlación lineal será de 0, aunque gráficamente parezca que existe una correlación. Finalmente, hay que tener en cuenta que un coeficiente de correlación cercano a 1 no implica una relación de causalidad entre las variables, podría tratarse sólamente de una coincidencia. A este fenómeno se le conoce como correlación espuria.

2.7.4. Recta de Regresión Lineal

Una forma interesante de expresar una relación lineal es, redundantemente, con una recta. Esta recta a su vez servirá de modelo para poder predecir resultados de observaiones de una solo variable. Esta recta se representa como:

$$Y = a + bX - \epsilon$$

Donde a y b son los coeficientes de la regresión y ϵ es el error.

Una forma óptima de calcular esta recta y sus parámetros es minimizando el error, o sea, que la distancia de cada uno de los puntos a la recta sea la mínima posible. Así, para cada par de datos (x_i, y_i), el error de esa observación respecto a la recta será:

$$\epsilon_i = y_i - (a + bx_i)$$

Ajusto por Mínimos Cuadrados

Como se ha mencionado en el anterior párrafo, el punto de la cuenta de la recta será minimizar el error según los valores de a y b. O sea, se quiere minimizar la función:

$$E(a, b) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Haciendo las cuentas se obtiene que:

$$b = \frac{s_{xy}}{s_x^2} \ , \quad a = \bar{y} - b\bar{x}$$

Una vez el modelo es óptimo, se puede usar la recta $Y = a + bX$ para hacer otras predicciones. Hay que considerar que es peligroso hacer observaciones fuera del rango de las variables. Se cumple también que, una vez optimizado el modelo, la media de los errores tiende a 0.

Varianza Residual y Coeficiente de determinación

Dada una recta de regresión lineal ajustada mediante mínimos cuadrados, la varianza residual se define como:

$$s_\epsilon^2 = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 = s_y^2 (1 - r_{xy}^2)$$

Donde aparece el coeficiente de determinación $R^2 = r_{xy}^2$. Este valor hace referencia a la proporción de la variación total en la variable dependiente Y.