Introducción a Learning Analytics con ejemplos prácticos

UD 01. Anexo II. Repaso de conceptos de estadística

Actualizado Septiembre 2020

Licencia

Nomenclatura

A lo largo de este tema se utilizarán distintos símbolos para distinguir elementos importantes dentro del contenido. Estos símbolos son:

📖 **Importante**

❕ **Atención**

💬 **Interesante**

ÍNDICE DE CONTENIDO

UD01. ANEXO II - REPASO DE CONCEPTOS DE ESTADÍSTICA

1. INTRODUCCIÓN

En este anexo, haremos un repaso rápido de algunos de los conceptos de estadística descriptiva utilizados a lo largo del curso. Desde aquí intentaremos repasar sobre todo qué es cada medida y qué usos puede tener.

No es necesario hacer hincapié en como realizar los cálculos estadísticos, ya que para ello podemos utilizar gran cantidad de utilidades libres, como LibreOffice Calc https://es.libreoffice.org/ o PSPP https://www.gnu.org/software/pspp/

> ❕ **Atención 1:** el objetivo de este anexo es simplemente servir de guía rápida de repaso. **No es necesario saber calcular las medidas utilizadas, ni ser un gran experto en estadística**.

> ❕ **Atención 2:** Si al repasar estos conceptos, alguno se os atraganta, no os preocupéis. Los ejemplos que trabajaremos durante el curso nos ayudarán a entender todo :)

2. CONCEPTOS A REPASAR

2.1 Media aritmética y ponderada

La media aritmética es un valor que se obtiene a partir de la suma de todo los valores de un conjunto de datos siendo esta dividida entre el número de sumandos.

La media ponderada, es similar a la media aritmética, solo que los valores pueden tener distinta ponderación (cada valor tiene un peso distinto en la media).

La media en general suele darnos información descriptiva sobre un conjunto de datos y en algunas ocasiones (en las que no hay alta variabilidad de datos) suele estar cerca del centro de la distribución.

Hay que tener en cuenta que por sus características es susceptible a tener desviaciones por valores muy altos o muy bajos, por lo cual debe complementarse por otras medidas.

Más información https://es.wikipedia.org/wiki/Media_aritm%C3%A9tica

2.2 Mediana

La mediana es un valor que se obtiene al ordenar un conjunto de datos cuantitativos y seleccionar el dato “de enmedio”.

El algoritmo utilizado para seleccionar la mediana, siendo N el número de elementos:

- Si N es impar, se toma el valor en la posición (N+1)/2. Dicha posición la ocupa el llamado valor central.
- Si N es par, se toma la media entre los dos valores centrales, es decir la media entre el elemento en la posición N/2 y el elemento de la posición (N/2) +1.

La mediana en general suele darnos información de que valores hay en la zona media de la distribución, estando menos afectada por variaciones de los valores extremos que la media.

Más información https://es.wikipedia.org/wiki/Mediana_(estad%C3%ADstica)

2.3 Moda

La moda es el valor que se repite con más frecuencia en un conjunto de datos. Al tratarse de frecuencias, además de poder usarse en datos cuantitativos, también puede utilizarse para medir datos cualitativos (Ejemplo: frecuencia de una palabra en un texto cualitativo)..

Aunque es una variable de fácil cálculo e interpretación, hay que tener en cuenta que es muy susceptible a variar entre muestras (al ignorar a veces gran parte de los datos) y tampoco tiene porqué estar cerca del centro de la distribución.

Mas información https://es.wikipedia.org/wiki/Moda_(estad%C3%ADstica)

2.4 Varianza

La varianza es una medida de dispersión respecto de la media de un conjunto de datos. Esta medida se calcula como la esperanza del cuadrado de la desviación de dicha variable respecto a su media.

La varianza tiene como unidad de medida el cuadrado de la unidad de medida de la variable. **Ejemplo**: si la variable se expresa en metros, la varianza se expresa en metros al cuadrado.

Para calcular la varianza:

- En primer lugar, calculamos la media del conjunto de datos.
- De cada elemento, calculamos el valor del elemento menos la media. El valor que obtenemos lo elevamos al cuadrado.
- Sumamos todos los valores que hemos obtenido y los dividimos por el número de elementos.

Los valores obtenidos al calcular la varianza no tiene una magnitud absoluta y dependen de la escala usada en el conjunto de datos. Podemos decir, que cuanto menor es el valor de la varianza, hay menor dispersión de datos, y viceversa.

Mas información https://es.wikipedia.org/wiki/Varianza

2.5 Desviación típica

La varianza por su propia naturaleza, es un valor cuadrático (se podría decir, que es la desviación cuadrática promedio de la media) que se expresa en diferente unidad de medida que la utilizada en el conjunto de datos para la que se ha calculado.

La desviación típica se calcula como la raíz cuadrada de la varianza, expresándose en la misma unidad de medida que el conjunto de datos original.

Por sus características, se podría decir, que la desviación típica es cuánto esperas que se desvie un valor del conjunto de datos de la media de dicho conjunto.

El valor obtenido al calcular la desviación típica, al igual que la varianza, no tiene una magnitud absoluta y depende de la escala usada en el conjunto de datos. Cuanto menor es el valor de la desviación típica, hay menor dispersión de datos, y viceversa.

Más información https://es.wikipedia.org/wiki/Desviaci%C3%B3n_t%C3%ADpica

2.6 Unidad tipificada (Standard score, Z-score)

En este punto hablaremos de la Unidad tipificada, también llamada “Standard score” o “Z-score”.

A veces, en estadística es necesario comparar datos que, siendo similares, proceden de distintos conjuntos de datos. A veces la comparación absoluta de estos datos no tiene sentido o aporta poca información.

En esos casos, se requiere un proceso de normalización para realizar una comparación que tenga en cuenta aspectos concretos de cada conjunto y nos proporcione información más útil.

Una forma de normalizar estas variables, es normalizar cada valor usando la **unidad tipificada:**

- Para calcular la unidad tipificada de un valor X de un conjunto de datos
  - Tomamos un valor X de un conjunto de datos.
  - Tomamos la media M del conjunto de datos.
  - Tomamos la desviación típica S del conjunto de datos.
  - Aplicando la siguiente fórmula obtenemos la unidad tipificada: **(X - M)/S.**

Se puede entender mejor con un ejemplo:

En clase se hizo en un instante dado la actividad A (una prueba tipo test). Posteriormente en el tiempo, se hizo en clase la actividad B (una actividad práctica). Queremos tomar un alumno y comparar en cuál de las actividades ha rendido mejor:

- En la actividad A, un alumno saca un 8.5. La media de su clase ha sido 7 y la desviación típica ha sido de 0.8. Su unidad tipificada sería (8.5 - 7)/0.8= 1.87.
- En la actividad B, un alumno saca 8. La media de su clase ha sido 6.7 y la desviación típica ha sido de 0.5. Su unidad tipificada sería (8- 6.7)/0.5=2.6.

Si comparamos el rendimiento del alumno usando su unidad tipificada, en la actividad B habría obtenido un mejor rendimiento que en la actividad A, pese a que en la actividad A en términos absolutos obtuvo una mejor calificación.

Se debe tener en cuenta en análisis que:

- La unidad tipificada puede tomar valor negativo.
- La varianza de todas las puntuaciones tipificadas es 1.
- La media de todas las unidades tipificadas es 0
    - Esto nos indica que valores más cercanos a 0 están más cercanos a la media y valores más alejados de cero, indicando que está alejado de la misma.
    - De ahí surge su otro nombre “Z-score” o “Zero score”.

Más información: https://es.wikipedia.org/wiki/Unidad_tipificada

2.7 Cuartiles

Los cuartiles son una medida de posición que nos indica en qué posiciones se divide en 4 partes un conjunto de datos. Generalmente, son utilizados para hacer análisis descartando valores demasiado altos o bajos que puedan sesgar la percepción del conjunto de datos.

Existen distintos métodos para calcular los cuartiles, incluso con diferentes resultados. El método más utilizado para N elementos ordenados es:

- **Primer cuartil**: se calcula la posición con la fórmula (N+1)/4.
    - Se podría decir que el primer cuartil, es la mediana de la primera mitad de los datos.
- **Segundo cuartil:** es la posición donde está la mediana.
- **Tercer cuartil**: se calcula la posición con la fórmula 3*(N+1)/4
    - Se podría decir que el tercer cuartil, es la mediana de la segunda mitad de los datos.
- Si se obtiene un número decimal, se considera la posición el siguiente entero.

Estas 3 referencias (primer cuartil, mediana, y tercer cuartil) nos dan una mejor perspectiva de los datos que observar únicamente la mediana.

Además, la posición del primer y tercer cuartil nos pueden servir para delimitar un subconjunto de datos entre esas posiciones. Con ese subconjunto podemos trabajar otras medidas estadísticas considerando que posiblemente tiene un menor sesgo que el conjunto original.

Además conociendo los cuartiles, podemos obtener dos medidas de dispersión interesantes:

- **Rango intercuartílico**: se obtiene de restar el valor de la posición del tercer cuartil al valor de la posición del primer cuartil.
    - Esta medida nos dice cuánto cambian los datos entre esas dos posiciones, ignorando posibles datos extremos, situados antes del primer cuartil o después del tercero.
- **Desviación cuartil**: se conoce a esta medida de dispersión como la mitad del rango intercuartílico (o dicho de otro modo, la media entre el primer cuartil y el tercer cuartil).
- Estas medidas de dispersión en general están menos afectadas por los sesgos que varianza y desviación típica, por lo cual son útiles para conjuntos de datos sesgados.

Más información:

- Cuartiles https://es.wikipedia.org/wiki/Cuartil
- Rango intercuartilico https://es.wikipedia.org/wiki/Rango_intercuart%C3%ADlico

2.8 Percentil

El percentil, al igual que el cuartil, es una medida de posición calculada para dividir un conjunto de datos, solo que la posición se fija indicando el porcentaje de elementos que quieren incluirse.

Puede indicarse si es inferior (el porcentaje se cuenta desde los primeros elementos) o superior (el porcentaje se cuenta desde los últimos elementos).

Por ejemplo, el percentil 10 inferior, la posición es la que marca el 10% de elementos del conjunto de datos.

Para calcular el percentil, siendo P el valor del percentil y N el número de elementos del conjunto de datos, se calcula como (P*N)/100. Si se obtiene un resultado decimal, se pasa al siguiente entero.

Por ejemplo, el percentil 25 de un conjunto de 50 elementos se calcula como (25*50)/100= 12.5, por lo cual consideramos el siguiente entero a 12-5 (el número 13) como el que fija el percentil 25.

Relacionándolo con los cuartiles:

- La posición del primer cuartil es equivalente al percentil inferior 25.
- La posición del segundo cuartil es equivalente al percentil inferior 50.
- La posición del tercer cuartil es equivalente al percentil inferior 75.

Mas información en https://es.wikipedia.org/wiki/Percentil

2.9 Coeficientes de asimetría (Fisher, Pearson y Bowley-Yule)

Los coeficientes de asimetría son indicadores que permiten establecer el grado de asimetría que presenta un conjunto de datos, sin tener que hacer su representación gráfica.

Sin entrar en detalles de su cálculo, entramos en los 3 principales coeficientes utilizados. Al final los 3 nos dan perspectivas numéricas de posibles asimetrías desde distintas perspectivas:

- **Fisher**: se basa en evaluar la proximidad de los datos a la media.
- **Pearson**: se basa en la diferencia entre la media y la moda respecto a la dispersión del conjunto de datos.
- **Bowley-Yule**: se basa en la suposición de que el primer cuartil y el tercer cuartil deben estar a una distancia igual a la mediana, así que los cambios indican asimetrías.

Los valores de estos coeficientes se interpretan de la siguiente forma:

- **Valores inferiores a 0**: la distribución tiene asimetría negativa (los valores se concentran en la primera mitad).
- **Igual a cero**: la distribución es simétrica.
- **Valores superiores a 0**: la distribución tiene asimetría positiva (los valores se concentran en la segunda mitad.

> 📖 **Importante:** En nuestros análisis **¿Por qué no simplemente representar gráficamente los datos y ver con nuestros ojos el nivel de asimetría y curtosis?** Si es posible, es una buena solución :) Aun así la representación gráfica no implica que estas medidas no pueden sernos útiles para comparar distribuciones, automatizar cálculos basados en ella, etc...

Más información https://es.wikipedia.org/wiki/Asimetr%C3%ADa_estad%C3%ADstica

2.10 Curtosis

La curtosis es una medida que nos permite, sin usar una representación gráfica, intuir la forma de la curva de una distribución (si la mayoría de los elementos están cerca de la media, si están cerca de los extremos, etc...).

Dicho de otra forma, el coeficiente de curtosis nos indica la cantidad de datos cercanos a la media.

Según los valores del coeficiente curtosis, podemos interpretar:

- **Coeficiente de curtosis = 0**: los datos siguen una distribución normal.
- **Coeficiente de curtosis > 0**: este valor indica que los datos del conjunto están distribuidos muy cerca de la media. A mayor valor de curtosis, más cerca de la media.
- **Coeficiente de curtosis < 0**: este valor indica que los datos del conjunto están distribuidos lejos de la media, en los extremos. A menor valor de curtosis, más alejados de la media.

Mas información en https://es.wikipedia.org/wiki/Curtosis

3. MATERIAL ADICIONAL

[1] Intro to statistics [Udacity.com]
https://classroom.udacity.com/courses/st101
[2] Media, mediana, moda y otras medidas de resumen
https://reporterodedatos.com/media-mediana-moda-y-otras-medidas-de-resumen/
[3] Medidas de posición central: media y mediana
https://www.universoformulas.com/estadistica/descriptiva/medidas-posicion-central/
[4] Medidas de dispersión
https://www.universoformulas.com/estadistica/descriptiva/medidas-dispersion/
[5] Medidas de posición no central: cuartiles y percentiles
https://www.universoformulas.com/estadistica/descriptiva/medidas-posicion-no-central/
[6] Asimetría y curtosis
https://www.universoformulas.com/estadistica/descriptiva/asimetria-curtosis/

4. BIBLIOGRAFÍA

[1] Statistics in a nutsell [Sarah Boslaugh, 2015]
[2] La estadística en comic [Larry Gonick, Woollcott Smith, 1993]
[3] The Manga Guide to Statistics [Shin Takahashi, 2008]