1.1. Aprender a obtener medidas de centralización y dispersión en una distribución estadística:

Objetivos:

1. Comprender la noción de distribución estadística y su no dependencia de las propiedades específicos de individuos específicos.
2. Aprender a comparar diferentes distribuciones estadísticas por el valor alrededor del cual se agrupan sus valores.
3. Aprender a comparar diferentes distribuciones estadísticas por la dispersión de sus valores.
4. Entender en qué medida varían las medidas de centralización y dispersión de una distribución estadística al sumar, restar, multiplicar o dividir sus valores por una cantidad fija.
5. Aprender a calcular las medidas de centralización y dispersión de forma que se simplifiquen los cálculos y se eviten los errores de cancelación.
6. Aprender a normalizar las distribuciones estadísticas a fin de hacerlas comparables más allá de sus medidas de centralización y dispersión.

Actividad 1.2. Una *variable aleatoria* (X) sobre un conjunto-población U es cualquier variable que puede tener distintos valores (x) para los distintos elementos-individuos de la población La distribución *estadística* de estos valoras no tiene en cuenta los individuos concretos para los que esta variable tiene cada valor, sino cuántos la tienen, lo que denominamos *frecuencia* f(x) de este valor en la población. Llamaremos *parámetro poblacional* a cualquier cantidad que solamente dependa de las frecuencias. Dos variables aleatorias serán estadísticamente *equivalentes* cuando tengan la misma distribución de frecuencias
Ejercicio 1.1: tomar una variable aleatoria sobre el alumnado asistente a la clase, por ejemplo el hecho de llevar o no llevar gafas, y realizar un experimento sencillo a fin de comprobar que el número de los que llevan gafas es un parámetro poblacional.

Actividad 1.3.
Ejercicio 1.2: representar gráficamente en diagramas de barras la distribución estadística de las frecuencias del número de calzado y de la edad en el alumnado asistente a clase.

Actividad 1.4. Como medidas de *centralidad* (valor alrededor del cual se agrupan los valores de la variable aleatoria) podemos tomar:
La *moda*: aquel valor que tenga la máxima frecuencia en la población.
La *mediana*: suponiendo que el conjunto de valores de la variable aleatoria esté ordenado, será un valor que tenga tantos individuos con un valor inferior como con un valor superior.
La *media* $\mu(X)$: suponiendo que los valores de la variable aleatoria sean números reales, y que el *tamaño* (número de individuos n(U)) de la población sea finito, viene dada por la suma de los valores X_i para todos los individuos i de la población dividida por su tamaño, $\mu(X) = \sum_i X_i / n(U)$.
Teorema 1.1: $\sum_x f(x) = n(U)$, $\mu(X) = \sum_x x \cdot f(x) / n(U)$.

Problema 1.1: calcular las diferentes medidas de centralización para las distribuciones estadísticas de la Actividad 1.3. ¿Cómo podemos utilizar las frecuencias para simplificar los cálculos?

Actividad 1.5. Para justificar que el cálculo de la media es una operación lineal, demostrar los siguientes teoremas:
Teorema 1.2: si tenemos dos variables aleatorias X, Y con valores numéricos reales sobre la misma población U, $\mu(X+Y)=\mu(X)+\mu(Y)$.
Teorema 1.3: si tenemos una variable aleatoria sumable X y un número real constante c, $\mu(c\cdot X)=c\cdot\mu(X)$.

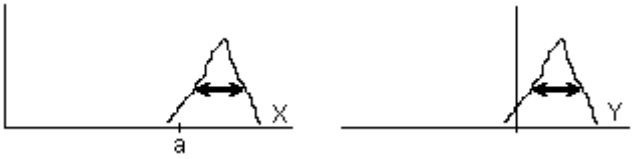
Actividad 1.6. A partir de la linealidad del cálculo de la media expresada en los dos teoremas anteriores, y teniendo en cuenta que la media de una constante es igual a la misma constante, demostrar
Teorema 1.4: si tenemos una variable aleatoria X con valores numéricos reales y un número real $a\in R$, entonces $\mu(X)=a+\mu(X-a)$.
Teorema 1.5: si tenemos una variable aleatoria X con valores numéricos reales y dos números reales $a,c\in R$, y tomamos $Y=(X-a)/c$, entonces $\mu(X)=a+c\cdot\mu(Y)$.
Los teoremas anteriores se pueden utilizar para simplificar el cálculo de la media. Aplicarlo para la resolución del
Problema 1.2: medir la longitud de la propia mano con una precisión de 0'5 cm y calcular la media del conjunto del alumnado asistente a clase.

Actividad 1.7. Como medidas de dispersión (para expresar el alejamiento entre sí de los valores de una variable aleatoria) podemos tomar:
Los *cuartiles* primero y tercero: suponiendo que el conjunto de valores de la variable aleatoria esté ordenado, los cuartiles serán tres valores que dividan al conjunto de valores en cuatro subconjuntos de valores que correspondan al mismo número de individuos; observamos que el segundo cuartil coincidirá con la mediana. Si tenemos definida una distancia en el conjunto de valores, podemos medir la dispersión como la distancia entre el primero y el tercer cuartil.
La *amplitud*: suponiendo que los valores estén ordenados y tengamos definida una distancia entre ellos, será la distancia entre los valores mínimo y máximo en la población.
La *desviación media*: suponiendo que los valores de la variable aleatoria sean números reales y que el tamaño de la población sea finito, será la media del valor absoluto de las diferències entre su valor para cada individuo y la media de estos valores, $\mu(|X-\mu(X)|)$
La *varianza* $\sigma^2(X)$: suponiendo que los valores de la variable aleatoria sean números reales y que el tamaño de la población sea finito, será la media del cuadrado de las diferencias entre su valor para cada individuo y la media de estos valores, $\sigma^2(X)=\mu((X-\mu(X))^2)$.
La *desviación típica* $\sigma(X)$: es la raíz cuadrada de la varianza.
Demostrar el
Teorema 1.6: $\sigma^2(X)=\mu(X^2)-\mu(X)^2$ (la variança es igual a la media de los cuadrados menos el cuadrado de la media).
Este teorema proporciona una forma más cómoda de calcular la varianza.
Problema 1.3: calcular las diferentes medidas de dispersión para las distribuciones estadísticas de la Actividad 1.3.

Problema 1.4: calcular la varianza de este conjunto de valores: (1000000'1 , 1000000'2 , 1000000'2, 1000000'3).

Actividad 1.8. Cómo habremos visto al intentar resolver el Problema 1.4, si la media de una distribución estadística es mucho más grande que su amplitud, entonces la media del cuadrado y el cuadrado de la media tendrán muchas cifras significativas coincidentes, que pueden incluso superar la precisión de nuestros instrumentos de cálculo; en este caso, obtendríamos erróneamente cero como su diferencia, produciéndose así un "error de cancelación". A fin de poder evitarlo utilizando las propiedades de la varianza, demostrar el

Teorema 1.7: si tenemos una variable aleatoria X con valores numéricos reales y un número real $a \in R$, y tomamos $Y=X-a$, entonces $\sigma^2(Y)=\sigma^2(X)$, es decir, la varianza es invariant ante traslaciones, como se puede entender fácilmente observando la figura adjunta.



Por lo tanto, podremos evitar el error de cancelación restando a todos los valores una cantidad fija próxima a su valor mínimo. Aplicarlo a la resolución del Problema 1.4.

Actividad 1.9. Demostrar el

Teorema 1.8: si tenemos una variable aleatoria X con valores numéricos reales y dos números reales $a,c \in R^+$, y tomamos $Y=(X-a)/c$, entonces $\sigma(X)=c\cdot\sigma(Y)$.

Aplicarlo por simplificar la resolución del

Problema 1.5: calcular la varianza de la distribución estadística del Problema 1.2.

Actividad 1.10. Para comparar la forma de distribuciones estadísticas con diferentes medias y varianzas podemos transformarlas en otras distribuciones estadísticas con medias y varianzas coincidentes. Llamaremos así *normalización* de una variable aleatoria X al resultado de restarle su media y dividir la diferencia por su desviación típica, $N(X)=(X-\mu(X))/\sigma(X)$. Demostrar el

Teorema 1.9: $\mu(N(X))=0$ y $\sigma(N(X))=1$.

Ejercicio 1.3: Representar gráficamente en la misma figura la normalización de las distribuciones estadísticas de la Actividad 1.3.