3. ADQUIRIR O ACTUALIZAR NOCIONES BÁSICAS DE ESTADÍSTICA:

Objetivos:

1. Aprender a calcular medidas de centralidad y dispersión de una distribución estadística.
2. Aprender a calcular probabilidades absolutas y condicionales de un determinado suceso.
3. Aprender a realizar estimaciones sobre una población a partir de una muestra de la misma.
4. Aprender a obtener la regresión lineal de dos variables aleatorias.
5. Aprender a trabajar con posibilidades absolutas y condicionales.

Actividad 23. Una *variable aleatoria* (X) en un conjunto-población es cualquier variable que puede tener distintos valores (x) para los distintos elementos-individuos de la población. La *distribución estadística* de dichos valores no tiene en cuenta los individuos concretos para los que dicha variable tiene cada valor, sino cuántos la tienen, a lo que llamamos *frecuencia* de dicho valor en la población. Llamaremos *parámetro poblacional* a cualquier cantidad que sólo dependa de las frecuencias. Para caracterizar una distribución estadística, nos interesará conocer su *centralidad*, dada por un valor alrededor del cuál se agrupan los valores de la variable aleatoria, y su *dispersión*, para expresar el alejamiento de dichos valores entre sí.
Como medidas de **centralidad** 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 cantidades sumables, y que el *tamaño* (número de individuos) de la población sea finito, viene dada por la suma de los valores para todos los individuos de la población dividida por su tamaño.
Como medidas de **dispersión** 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; obsérvese 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 primer y el tercer cuartil.
La *amplitud*: suponiendo que además tengamos definida una distancia en el conjunto de valores, será la distancia entre los valores mínimo y máximo en la población.
La *desviación media*: suponiendo que los valores de la variable aleatoria sean cantidades sumables, y que el tamaño de la población sea finito, será la media del valor absoluto de las diferencias entre su valor para cada individuo y su valor medio
La *varianza* $\sigma^2(X)$: suponiendo que los valores de la variable aleatoria sean cantidades sumables, y que el tamaño de la población sea finito, será la media del cuadrado de las

diferencias entre su valor para cada individuo y su valor medio [$\sigma^2(X)=\mu((X-\mu(X))^2)$]. La varianza puede calcularse también como la media de los cuadrados menos el cuadrado de la media [$\sigma^2(X)=\mu(X^2)-\mu(X)^2$] . Para calcularla de esta forma, y en el caso de que la media sea mucho mayor que la amplitud, conviene restar a todos los valores una cantidad fija próxima a su valor mínimo, operación que no modificará la dispersión.
La *desviación típica* $\sigma(X)$: es la raiz cuadrada de la varianza.
Llamaremos *normalización* de una variable al resultado de restarle su media y dividir la diferencia por su desviación típica, $N(X)=(X-\mu(X))/\sigma(X)$; la media de la variable normalizada valdrá cero [$\mu(N(X))=0$] y su desviación típica valdrá uno [$\sigma(N(X))=1$].
Ejercicio 18: definir una variable aleatoria (por ejemplo, el número de calzado) en la población definida por los alumnos de la asignatura, y obtener todos los parámetros poblacionales definidos en esta actividad. Estudiar cómo simplificar el cálculo de algunos de ellos utilizando las frecuencias.

Actividad 24. Llamaremos *probabilidad* de un subconjunto de valores de una variable aleatoria al cociente entre el número de individuos que tienen alguno de los valores del subconjunto y el tamaño de la población. Si el conjunto de valores es finito, podremos calcular la probabilidad de cada valor como su frecuencia dividida por el tamaño de la población, y la probabilidad de un subconjunto de valores como la suma de las probabilidades de cada uno de ellos. En tal caso, la suma de las probabilidades de todos los valores será igual a 1: llamaremos *distribución probabilística* a cualquier aplicación que asigne un número real no negativo a cada valor de un conjunto de valores tal que su suma valga 1. Si el conjunto de valores no es finito, pero tenemos definida una medida sobre el mismo, llamaremos *distribución de densidad probabilística* a cualquier aplicación que asigne un número real no negativo a cada valor tal que su integral valga 1: la probabilidad de un subconjunto de valores vendrá dada por la integral de dicha aplicación sobre el mismo.
Ejercicio 19: estudiar cómo obtener la media de una variable aleatoria conociendo su distribución probabilística. Aplicarlo al caso del Ejercicio 18. ¿Cómo podríamos definir la media de una distribución de densidad probabilística?

Actividad 25: Si trabajamos con dos variables aleatorias X,Y combinadas, podemos obtener del mismo modo la probabilidad (absoluta) P(x,y) de que dichas variables tengan simultáneamente unos valores determinados contando el número de individuos en que ello se produzca y dividiéndolo por el tamaño de la población. Pero en determinadas ocasiones nos interesará conocer la *probabilidad condicional*, P(y|x), definida como la probabilidad de un valor (y) de una variable (Y) restringida a la subpoblación definida por un determinado valor (x) de otra variable (X). Para ello podemos contar el número de individuos que tienen el par de valores (x.y) y dividirlo por el número de individuos que tienen el valor (x) de la segunda variable, que será el tamaño de la subpoblación. Pero podemos también calcular la probabilidad condicional utilizando el **Teorema de Bayes**, que nos dice que $P(y|x)=P(x,y)/P(x)$. Si las dos variables aleatorias son *independientes*, la probabilidad condicional coincidirá con la probabilidad absoluta del valor de la primera variable, $P(y|x)=P(y)$, y se cumplirá que $P(x,y)=P(x)P(y)$.
Ejercicio 20: combinar la variable aleatoria estudiada en el Ejercicio 18 con otra variable definida en el conjunto de alumnos de la asignatura (por ejemplo, el sexo o la

edad). Obtener las probabilidades condicionadas de la primera variable respecto de la segunda. ¿Son independientes? A partir de la probabilidades ya calculadas, y fijando el valor de la primera variable, obtener la probabilidad condicional de la segunda variable respecto de dicho valor.

Ejercicio 21: en una ciudad multiétnica se comete un delito. Un testigo afirma que dicho delito ha sido cometido por una persona "de color". Pero reproduciendo la situación en las mismas condiciones de iluminación se encuentra que el testigo acierta el "color" en un 70% de los casos, tanto si es realmente "de color" como "blanca". Sabiendo que en la ciudad hay un 10% de personas "de color", calcular cuál es la probabilidad de que el delito haya sido cometido realmente por una persona "de color".

Actividad 26. Llamamos *muestra* (sin reemplazamiento) de una población a cualquier subconjunto de la misma. Una muestra con reemplazamiento se obtendrá extrayendo sucesivamente individuos de la población sin suprimirlos de la misma. Llamamos *estadístico* a cualquier parámetro poblacional restringido a una muestra, como la media muestral $\bar{X}$ o la desviación típica muestral s(X) [naturalmente, para los estadisticos se cumplen la mismas relaciones que para los parámetros poblacionales correspondientes, por ejemplo $s^2(X)=\bar{(X^2)}-(\bar{X})^2$]. Y llamamos *distribución muestral* a la distribución de un estadístico en el conjunto de todas las muestras de la población de un cierto tamaño n. Los parámetros poblacionales de la distribución muestral, como la media de las medias $\mu(\bar{X})$, la varianza de las medias $\sigma^2(\bar{X})$ o la media de las varianzas $\mu(s^2(X))$, son importantes para realizar estimaciones a partir del estadístico de una muestra sobre el correspondiente parámetro poblacional. A tal efecto, hay que tener en cuenta las siguientes relaciones:

$\mu(\bar{X})=\mu(X)$

$\sigma^2(\bar{X})=\sigma^2(X)/n$ si las muestras son con reemplazamiento o la población es infinita (aproximadamente, si es muy grande).

$\mu(s^2(X))=\sigma^2(X)\cdot(n-1)/n$

Diremos que un estadístico es *insesgado* cuando la media de su distribución muestral es igual al correspondiente parámetro poblacional. Para estimar un parámetro poblacional a partir de un estadístico, éste debe ser insesgado.

Ejercicio 22: ¿la media y la varianza muestrales son estadísticos insesgados? En caso de que alguna de ellas no lo sea, ¿cómo podríamos obtener un estadístico corregido que sí fuera insesgado? Tomando el conjunto de alumnos de la asignatura como una muestra del conjunto de alumnos de la licenciatura, realizar una estimación sobre este conjunto de la media y de la varianza de la variable aleatoria del Ejercicio 18. Estimar también la varianza de la distribución muestral de las medias.

Actividad 27. Diremos que $[S_1,S_2]$ es un *intervalo de confianza* del $100\alpha\%$ para un parámetro poblacional Ω si la probabilidad de que Ω se encuentre en dicho intervalo es igual a α. Para poder obtener intervalos de confianza necesitaremos conocer la distribución muestral de un estadístico insesgado S de dicho parámetro poblacional. Definimos la distribución *normal* como la distribución de densidad probabilística dada por la fórmula

$$p(x) = \exp(-(x-\mu)^2/(2\sigma^2))/(\sigma(2\pi)^{1/2})$$

Se demuestra que si la población es infinita y las muestras son grandes, la distribución muestral de las medias se aproxima a la distribución normal. Trabajando con una

variable aleatoria normalizada $N(X)=(X-\mu(X))/\sigma(X)$, la distribución muestral de sus medias se aproximará a la distribución normal tipificada, que es la que tiene una media $\mu=0$ y una desviación típica $\sigma=1$,

$$p(x) = \exp(-x^2/2)/(2\pi)^{1/2}$$

Si las muestras son pequeñas pero la distribución de la variable aleatoria X en la población corresponde a una distribución normal, entonces la distribución muestral de las medias de la correspondiente variable aleatoria normalizada,

$$t = (\bar{X}-\mu(\bar{X}))/\sigma(\bar{X}) ,$$

se ajusta a una distribución **t** *de Student*, definida por

$$p_\nu(t) = p_\nu(0)\cdot(1+t^2/\nu)^{-(\nu+1)/2}$$

con $\nu=n-1$ grados de libertad, siendo n el tamaño de las muestras y escogiendo $p_\nu(0)$ de modo que la integral entre menos infinito y más infinito de $p_\nu(t)$ valga la unidad, de modo que sea una distribución de densidad probabilística. En la práctica, puede utilizarse como aproximación la distribución **t** de Student si el tamaño de la población es mucho mayor que el de la muestra. Se demuestra que si ν tiende a infinito la distribución **t** de Student tiende a la distribución normal tipificada. Llamaremos *coeficiente de confianza* $t_\alpha(\nu)$ al valor de t tal que la probabilidad de encontrarse entre $-t_\alpha(\nu)$ y $t_\alpha(\nu)$ dada por una distribución **t** de Student con ν grados de libertad sea igual a α.

Ejercicio 23: para obtener un intervalo de confianza del 80% para la media $\mu(X)$ de la variable aleatoria del Ejercicio 18 en el conjunto de alumnos de la licenciatura a partir de la media $\bar{X}$ y de la desviación típica s(X) en el conjunto de alumnos de la asignatura, comenzaremos estimando la desviación típica $\sigma(X)$ de la población mediante la desviación típica corregida ^s(X) para ser insesgada. A partir del valor estimado de $\sigma(X)$ calcularemos la desviación típica $\sigma(\bar{X})$ de la distribución muestral mediante la fórmula de la Actividad 26. De la tabla de la distribución t de Student obtendremos el coeficiente de confianza $t_{0'80}(\nu)$ (deberemos fijarnos en la figura que encabeza la tabla para determinar cuál es la columna que corresponde a dicho coeficiente de confianza). Y finalmente utilizaremos la expresión de la media t de la correspondiente variable aleatoria normalizada y la condición $-t_{0'80}(\nu) < t < t_{0'80}(\nu)$ para obtener el intervalo de confianza del 80% para $\mu(X)$.

Actividad 28. Si la distribución de una variable aleatoria X en una población infinita se ajusta a una distribución normal, la distribución muestral del estadístico

$$\chi^2 = n\cdot s^2(X)/\sigma^2(X)$$

se ajusta a una distribución *Chi-cuadrado*, definida por

$$p_\nu(\chi^2) = K_\nu\cdot(\chi^2)^{(\nu-2)/2}\cdot\exp(-\chi^2/2)$$

con $\nu=n-1$ grados de libertad, siendo n el tamaño de las muestras y escogiendo K_ν de modo que la integral entre menos infinito y más infinito de $p_\nu(\chi^2)$ valga la unidad, de modo que sea una distribución de densidad probabilística. Llamaremos *coeficiente crítico* $\chi^2_p(\nu)$ al valor de χ^2 tal que la probabilidad de encontrar un valor mayor o igual en una distribución Chi-cuadrado con ν grados de libertad sea igual a p.

Ejercicio 24: para obtener un intervalo de confianza del 80% para la varianza $\sigma^2(X)$ de la variable aleatoria del Ejercicio 18 en el conjunto de alumnos de la licenciatura a partir de la varianza $s^2(X)$ en el conjunto de alumnos de la asignatura buscaremos en la tabla de la distribución Chi-cuadrado dos coeficientes críticos $\chi^2_p(\nu)$ y $\chi^2_{1-p}(\nu)$ tales que la probabilidad de que el estadístico χ^2 se encuentre entre ellos sea de 0'80, y utilizaremos la expresión de dicho estadístico para obtener el intervalo de confianza del 80% para

$\sigma^2(X)$. Obtener el correspondiente intervalo de confianza del 80% para la desviación típica $\sigma(X)$ de la población y comprobar que la desviación típica corregida ^s(X) de la muestra pertenece a dicho intervalo.

Actividad 29. Si tenemos una variable aleatoria con k valores i=1,2..k, una hipótesis que les asigna probabilidades p(i), y una muestra de tamaño n, llamaremos *frecuencia esperada* del valor i en dicha muestra a $e_i=n\cdot p(i)$, y frecuencia observada o_i a su frecuencia en la muestra. Se demuestra que si todas las frecuencias esperadas son igual o mayor que 5, entonces la distribución del estadístico χ^2 obtenido sumando $(o_i-e_i)^2/e_i$ para todos los valores de la variable aleatoria se aproxima a la distribución Chi-cuadrado con grados de libertad $\nu=k-1$ (prueba Chi-cuadrado). Si dicho estadístico es superior a $\chi^2_\beta(\nu)$, diremos que la hipótesis probabilística es *rechazada* por la muestra con un nivel de significación β. Si dicho estadístico es inferior a $\chi^2_{1-\beta}(\nu)$, diremos hay *concordancia* entre la hipótesis y la muestra con un nivel de significación β. Si el estadístico se encontrara entre dichos dos valores, diremos que la muestra *no es definitoria* para la hipótesis con ese nivel de significación.
Ejercicio 25: arrojar un dado 30 veces, anotar el número de veces que se obtiene cada cara y realizar realizar estimaciones sobre hipótesis probabilísticas en relación al dado.

Actividad 30. Si tenemos dos variables aleatorias numéricas X e Y, llamaremos *covarianza* de las mismas a

$$c_{XY}=\mu(X\cdot Y)-\mu(X)\cdot\mu(Y)$$

y diremos que y=a+bx es la *recta de regresión* de Y sobre X si la suma de los cuadrados de las diferencias entre a+bX e Y es la mínima posible. Se demuestra que la recta de regresión pasa por el punto $(\mu(X), \mu(Y))$, y que si la varianza de X es mayor que cero, entonces dicha recta se obtiene tomando

$$b=c_{XY}/\sigma^2(X),\ y=\mu(Y)+b\cdot(x-\mu(X)).$$

Si la varianza de Y es también mayor que cero, definimos el *coeficiente de correlación* entre X e Y por

$$\rho_{XY} = c_{XY}/(\sigma(X)\sigma(Y))$$

Se demuestra fácilmente que si X e Y son independientes entonces $c_{XY}=0$, y por tanto $\rho_{XY}=0$ (la recíproca no es cierta).
Teniendo en cuenta que $\mu(\ )$ es lineal y que $\sigma(a+bX)=|b|\sigma(X)$ se demuestra también fácilmente que si los puntos (X,Y) están alineados sobre la recta de regresión, es decir Y=a+bX, entonces $\rho_{XY}=\pm1$ (según cuál sea el signo de b).
El coeficiente de correlación mide el grado de ajuste de la recta de regresión: si vale 0 diremos que no hay correlación lineal (aunque puede haber correlación no lineal), y si vale ±1 diremos que la correlación lineal es perfecta. Si $\rho_{XY}>0$ diremos que la correlación lineal es positiva, y si $\rho_{XY}<0$ diremos que es negativa.
Ejercicio 26: comprobar que si X e Y son independientes entonces $c_{XY}=0$ y que si Y=a+bX entonces $\rho_{XY}=\pm1$
Ejercicio 27: estudiar la correlación lineal en el caso

X	1	2	3
Y	1	2	1

¿X e Y son independientes?
Ejercicio 28: obtener la recta de regresión y estudiar la correlación lineal entre dos variables aleatorias (por ejemplo, la edad y el número de calzado) en la población definida por los alumnos de la asignatura.

Actividad 31. Puede darse el caso de que no conozcamos las frecuencias relativas (es decir, las probabilidades) de los distintos valores de una variable aleatoria X, y debamos limitarnos a estimar su *posibilidad* entre 0 y 1. Dado que la variable deberá tener algún valor, deberá haber algún valor x con posibilidad $\pi(x)=1$. Así, llamaremos *distribución posibilística* a cualquier aplicación que asigne un número real no negativo a cada valor de un conjunto E de valores, de modo que su máximo sea 1. Para todo subconjunto A de valores, su posibilidad $\pi(A)$ vendrá dada por el máximo de dicha aplicación sobre el mismo, y definiremos su *necesidad* como $\nu(A)=1-\pi(E-A)$; obtendremos la necesidad de un valor x mediante $\nu(x)=1-\pi(E-\{x\})$.
Si trabajamos con dos variables aleatorias X,Y combinadas, podemos estimar también la posibilidad (absoluta) $\pi(x,y)$ de que dichas variables tengan simultáneamente unos valores determinados. En tal caso, $\pi(x)$ será la posibilidad del conjunto $\{(x,Y)\}$, es decir, el máximo para todo y de $\pi(x,y)$. A su vez, $\pi(y)$ será la posibilidad del conjunto $\{(X,y)\}$, es decir, el máximo para todo x de $\pi(x,y)$.Y definiremos la *posibilidad condicional* $\pi(y|x)$ mediante

$\pi(y|x) = \pi(x,y)$ si $\pi(x)>\pi(x,y)$
$\pi(y|x) = 1$ si $\pi(x)=\pi(x,y)$

A partir de la posibilidad condicional $\pi(y|x)$ puede obtenerse la posibilidad absoluta $\pi(x,y)$, que será el mínimo de $\pi(x)$ y $\pi(y|x)$.
Ejercicio 29: dividir la clase en 2 grupos; el primer grupo definirá dos variables aleatorias X, Y, estimará los valores de las posibilidades absolutas $\pi(x,y)$ y calculará y hará públicas los valores de la posibilidad condicional $\pi(y|x)$ y de la posibilidad $\pi(x)$; a partir de ellos, el segundo grupo calculará los valores de la posibilidad condicional $\pi(x|y)$ y de la necesidad $\nu(x)$.