ESTADÍSTICA Y PROBABILIDAD
(Very, very important)

Este documento pretende ser un resumen de la información que tienes ya en el libro, y lo usaremos junto con este y los vídeos del Aula Virtual. Ahora sólo falta que tu pongas el esfuerzo necesario para que todo salga bien. Eso si es very, very, very important.

1.- Estadística unidimensional

- En la estadística es muy importante no confundir el *número de datos distintos*, con el *número total de datos* **(QUE ES LA SUMA DE TODAS LAS FRECUENCIAS ABSOLUTAS).** Éste último lo utilizamos prácticamente en todo momento. Que vas a calcular la tabla de frecuencias, pues para las relativas lo necesitas. Que vas a calcular la media, pues también, ... En una tabla, ¿qué es?

Peso	Nº de personas
20-25	7
25-30	5
30-35	9
35-40	11
Totales:	**32**

Yo no puedo decir que hay 4 datos, por mucho que haya 4 datos distintos, pues hay 32 en total.

- **Tablas de frecuencias:** Yo puedo hacer de dos maneras las tablas de frecuencias.

  · Calcular la frecuencia relativa a partir de la absoluta, después el tanto por ciento y por último, para obtener las acumuladas ir acumulando las correspondientes.

  · Calcular las normales igual, acumular la frecuencia absoluta y a partir de ahí calcular las acumuladas como has hecho las simples.

  Vídeo: Tablas de frecuencias.

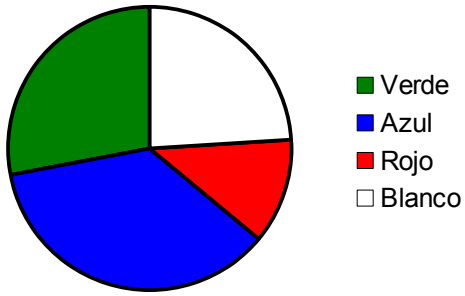
- **Gráficos:** La mayoría son muy claros y sólo suelen hacerse con las frecuencias absolutas, aunque podemos usar cualquiera. Quizás merece la pena hacer hincapié en los diagramas de sectores. El objetivo es repartir los 360 grados de una circunferencia entre todos los datos, eso si, de acorde a los que haya de cada dato. Una vez más recurrimos al **"número total de datos"** antes citado. Primero hallaremos el ángulo que le corresponde a un sólo individuo. Después calcularemos la amplitud del ángulo que le corresponde a cada dato. Por último representaremos

el primer dato partiendo del ángulo 0 y continuaremos desde la última línea uniendo el ángulo que corresponde a cada dato.

Ejemplo: Si tengo la siguientes tabla de datos:

Color	Nº coches	Amplitud
Verde	7	7 * 14'4 = 100'8º
Azul	9	9 * 14'4 = 129'6º
Rojo	3	3 * 14'4 = 43'2º
Blanco	6	6 * 14'4 = 86'4º
Total:	25	

Tenemos que 360: 25= 14.4º



Vídeos: Diagrama de sectores y polígono de frecuencias, diagrama de barras e histogramas.

- Medidas de centralización:

· **Media:** Es un dato comprendido entre el más pequeño y el más grande. Hay que usar, una vez más, el número total de datos. Construimos una columna en la que multipliquemos el dato *x* por la frecuencia. Hacemos la suma de esa columna y la dividimos entre el número total de datos.

· **Moda:** Es el dato que más veces se repite, el que está de moda. Puede ocurrir con varios datos y entonces se dice que es multimodal (eso si, los que son iguales tienen que ser los que más se repiten)

· **Mediana:** El dato que está en el medio. Para calcular primero vamos a ver que lugar ocupa y luego quien hay en ese lugar. Para eso usamos la frecuencia absoluta acumulada, que va contando por nosotros los que llevamos colocados. Empecemos por calcular que lugar ocupa. Para eso usamos una fórmula que dice:

$$Lugar = \frac{N^o\ total\ de\ datos}{2} + 0'5$$

- Medidas de dispersión: Para eso tenemos que hacer otra columna en la que multiplicamos

cada dato al cuadrado por su frecuencia.

· **Varianza:** Hacemos la suma de esa columna y la dividimos por el número total de datos. Al resultado le restamos la media al cuadrado.

· **Desviación típica:** Es el más importante de este tipo. Es la raíz cuadrada de la varianza.

· **Coeficiente de variación:** Se obtiene dividiendo la desviación típica entre la media.

$$CV = \frac{\sigma}{\bar{x}}$$

Vídeos: Medidas de centralización, medidas de dispersión.

(Ya veremos una tabla a modo de ejemplo en la estadística bidimensional de como construirla para hacer los cálculos.)

- Medidas de posición: Las más importantes son los cuartiles (Q_1, Q_2 y Q_3) y los percentiles. Los cuartiles nos dividen el conjunto de datos en cuatro partes iguales. El objetivo es encontrar el valor de la variable que hace eso. Los percentiles dividen la variable en 100 partes. Los cuartiles son: $Q_1=P_{25}$, $Q_2=P_{50}$ y $Q_3=P_{75}$. Por tanto vamos a aprender a calcular los percentiles y así hacemos todo.

· **Datos no agrupados en intervalos:** Si los datos no vienen en una tabla los ordenamos de menor a mayor y para calcular el percentil correspondiente. Luego hay que calcular el lugar que buscamos y localizar el elemento que hay en ese lugar (Se redondea por arriba) (Ver ejemplo de la página 228 del libro)

Vídeo: Datos no agrupados.

· **Datos agrupados en intervalos:** En este caso construimos la tabla de frecuencias, llegando a poner el % acumulado (Basta con hacer las frecuencias acumuladas). Una vez hecho eso buscamos el intervalo que contenga el porcentaje que buscamos. Se trata de hacer una regla de tres en la que intervienen:

a) Tamaño del intervalo.

b) Porcentaje de datos que contiene ese intervalo (basta con restar al porcentaje de ese intervalo el del anterior)

c) Resta del porcentaje que buscamos menos el del intervalo anterior.

Una vez calculado por la regla de tres el valor de la x se lo sumamos al menor valor de nuestro intervalo. (Ver ejemplo segundo de la página 229)

Vídeo: Datos agrupados.

2.- Estadística bidimensional

- Nube de puntos o diagrama de dispersión: Consiste en representar en unos ejes coordenados los pares de puntos (x,y) que nos dan en la tabla (sin la frecuencia) Esto nos dará una idea de de la relación que hay entre ellos, aunque, a través de los parámetros que veremos luego nos haremos mejor a la idea.

Los datos pueden dárnolos en tablas simples o de doble entrada. Nosotros trabajaremos con tablas simples, por lo que si nos dan una de doble entrada la pasaremos a simple.

Vídeo: Paso de tabla doble a simple

- Cálculo de parámetros: La mayoría son los mismos que en la estadística unidimensional, pues se limitan a calcularlos para cada variable.

· **Medias marginales:** Consiste en calcular las medias de cada una de las variables. (*En la tabla aparecen dos columnas con cada dato por su frecuencia*)

· **Desviaciones típicas marginales:** Es la desviación típica de cada una de las variables. Tendremos que calcular primero la varianza de cada una y luego haremos la raíz cuadrada. (*Para calcular la varianza en la tabla ponemos dos columnas con cada dato elevado al cuadrado por su frecuencia*)

· **Covarianza:** Este parámetro es nuevo. Para calcularlo haremos en la tabla una nueva columna en la que multiplicaremos los dos datos y la frecuencia. Dividiremos la suma de estos productos entre el número total de datos y a ese resultado le restaremos el producto de las medias marginales. (*En la tabla aparecen una columna con el producto de las dos variables por la frecuencia*)

Si la covarianza es **positiva** quiere decir que c*uando aumenta la variable X lo hace también la Y*, mientras que cuando es **negativa**, *al aumentar la X disminuye la Y.*

Como ejemplo puedes ver el que viene en la página 247 del libro. Las columnas 4ª y 6ª se usan para calcular las medias marginales, las 5ª y 7ª para calcular las desviaciones típicas marginales y la 8ª para calcular la covarianza.

Vídeo: Parámetro sin frecuencias, parámetros con frecuencias.

- Correlación: La covarianza nos indica como es la relación entre las dos variables, pero no la fortaleza de dicha relación. Eso va a estudiarlo el coeficiente de correlación.

Comenzaremos por hablar de la correlación:

· **Correlación funcional:** Todos los puntos están alineados. Se sabe que va a ocurrir exactamente (El listado de los precios de la churrería)

· **Correlación directa:** Al aumentar una variable aumenta la otra (Si hubiéramos calculado la covarianza saldría positiva) (La nube de puntos está inclinada hacia arriba)

· **Correlación inversa:** Al aumentar una variable disminuye la otra (Si hubiéramos

calculado la covarianza saldría negativa) (La nube de puntos está inclinada hacia abajo)

· **Correlación nula:** No existe relación entre las variables (a nube de puntos aparece muy dispersa)

Pero toda la información nos la va a dar el coeficiente de correlación de Pearson. Este se calcula con la fórmula:

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

Su valor está comprendido entre -1 y 1.

Este coeficiente nos aporta una doble información:

· **Según su signo:** Como podemos observar, y teniendo en cuenta que las desviaciones típicas son siempre positivas, el signo de *r* depende del signo de la covarianza. Luego si $r > 0$ la correlación será directa, y si $r < 0$ la correlación será inversa.

· **Según su valor absoluto:**

- Si $r = -1$ o $r = 1$, la relación es funcional.
- Si r está próximo a 0 la correlación es muy débil.
- Si r está entorno a 0'5 ó -0'5 la correlación es débil.
- Si r está próximo a 1 ó -1 la correlación es fuerte.

Vídeo: Coeficiente de correlación

- Recta de regresión: En aquellos casos en los que la correlación sea fuerte podemos realizar estimaciones de lo que debería ocurrir en aquellos casos en los que no tenemos datos. La fórmula para calcular la **recta de regresión de y sobre x** es:

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

Mira los ejemplos de las páginas 250 y 251.

Vídeo: Recta de regresión

3.- Probabilidad. Distribución binomial y normal

- Probabilidad condicionada: Vamos a escribirla $P(B/A)$. Tiene sentido hablar de ella cuando el resultado del primer suceso altera el del segundo. En este caso los sucesos son *dependientes* (por ejemplo, si sacamos una bola de una urna y no la devolvemos). Si el resultado del primer experimento no condiciona el del segundo decimos que los sucesos son *independientes* (por ejemplo, *s*i sacamos una bola de la urna y la devolvemos).

Si realizamos un diagrama de árbol, la *probabilidad condicionada* son las probabilidades que colocamos en las segundas ramas y sucesivas.

- Teoremas de probabilidad:

· **Regla del producto:** Esta regla se aplica cuando vamos a calcular la probabilidad de que *ocurra un camino del diagrama de árbol.* Estamos calculando la probabilidad de una intersección (probabilidad de que la primera sea A *y* la segunda sea B *y* …) Se obtiene multiplicando todas las probabilidades que nos encontramos en el camino.

· **Regla de la suma:** En muchos casos lo que nos piden no contiene sólo un camino, sino que son varios. En ese caso calcularíamos la probabilidad de cada camino por el método anterior y sumaríamos todas las que nos sean favorables.

· **Teorema de Bayes:** Para distinguirlo yo siempre miro en que sentido va la condición, es decir, que dan por seguro y que probabilidad me piden. Si dan por seguro que ha ocurrido un suceso del segundo experimento y me piden calcular la probabilidad de que ocurrido eso salga algo del primero, entonces estoy ante el teorema de Bayes. Para calcular la probabilidad que nos piden tenemos que dividir la probabilidad del camino que nos es favorable entre la suma de las probabilidades de todos los caminos que contienen lo que es seguro.

Vídeos: Reglas del producto y la suma, Teorema de Bayes.

- Distribuciones de probabilidad discreta: Es una forma de teorizar modelos prácticos en los que la variable sólo toma determinados valores (Nº de hijos). Vamos a ver como calcular los tres parámetros que hemos dado. Para eso vamos a hacer una tabla en la que tendremos cuatro columnas: El valor de la variable (x), la probabilidad (p), la probabilidad por el valor de la variable (p . x) y la probabilidad por el cuadrado de la variable (p . x^2).

· **Media o esperanza matemática:** Este parámetro se calcula sumando la tercera columna.

· **Varianza:** Este parámetro se calcula sumando la cuarta columna y restando al resultado el cuadrado de la media antes calculada.

· **Desviación típica:** Es la raíz cuadrada de la varianza.

Ver el ejemplo segundo de la página 267.

Vídeo: Calculo de los parámetros.

- Binomial: La binomial es un caso concreto de distribución de probabilidad discreta. Tiene las siguientes características:

· El resultado de cada prueba tiene sólo dos opciones.

· La probabilidad de éxito es siempre fija y la denominamos *p*. Por tanto la de fracaso será $q=1-p$

· El resultado de cada prueba es independiente de los anteriores.

La binomial viene caracterizada por el número de experimentos, *n*, y la probabilidad de éxito, *p*, es decir, B(n,p).

La fórmula para calcular la probabilidad de obtener *k* éxitos es:

$$P(x=k)=\binom{n}{k}p^k\cdot q^{n-k}$$

En esta fórmula el número combinatorio cuanta por nosotros las distintas posibilidades de ordenar los éxitos en el total de experimentos. Luego multiplicamos por la probabilidad de acertar elevado a los aciertos obtenidos por la probabilidad de fallar elevado a los fallos obtenidos.

Si nos piden calcular la $P(x\geqslant k)$ (*o mayor que, o menor igual, o menor que*), hay que contemplar todos los casos posibles. En ese caso puede ser interesante ver cuantos casos componen el suceso contrario, pues la probabilidad de un suceso es igual a 1 menos la de su contrario.

Ejemplo:

Si tengo una binomial B(6,0'3) y me piden $P(x\geqslant 2)$, tengo que

$$P(x\geqslant 2)=P(x=2)+P(x=3)+P(x=4)+P(x=5)+P(x=6)$$

pero su contrario necesita menos cálculos.

$$P(x\geqslant 2)=1-P(x<2)=1-\left[P(x=0)+P(x=1)\right]$$

Muy importante respetar el paréntesis o calcular primero la probabilidad del contrario antes de restárselo a 1.

Los parámetros de la binomial son fáciles de calcular y sus fórmulas son:

· Media: $\mu=n\cdot p$

· Varianza: $V=n\cdot p\cdot q$

· Desviación típica: $\sigma=\sqrt{V}$

Ver los ejemplos de la página 269.

Vídeo: Binomial.

- Distribuciones de probabilidad continuas: Es una forma de teorizar modelos prácticos en los que la variable puede tomar todos los valores comprendidos en un determinado intervalo. Es

importante darse cuenta que en estas distribuciones *no tiene sentido hablar de que la probabilidad tome un valor concreto, pues esa probabilidad va a ser cero*. Por tanto dará lo mismo hablar de que calcules la probabilidad de que sea mayor o igual que un número o de que sea mayor que dicho número, pues la de que sea igual es cero.

· **Función de densidad:** Una función de densidad tiene las siguientes características:

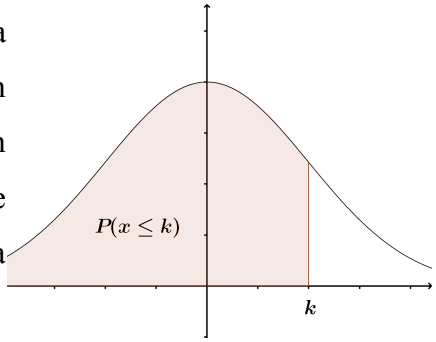
- $f(x) \geqslant 0$ para todo valor de x.
- El área comprendida entre el eje X y la función tiene que ser 1.

Para calcular la probabilidad de que la variable esté comprendida entre dos valores basta con calcular el área que encierra la función entre esos dos valores.

Ver el ejemplo segundo de la página 270 que continua en la 271.

Vídeo: Distribuciones de probabilidad continuas.

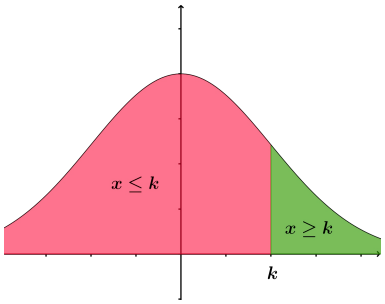
- Distribución normal: En principio vamos a trabajar con la distribución N(0,1). Para eso tenemos unas tablas que vienen al final del libro. La función de densidad de una distribución normal se conoce con el nombre de campana de Gauss. De hecho, las tablas nos dan el valor de la probabilidad de que la variable x sea menor o igual que k (con k>0).



Todo tendremos que reducirlo a poder usar las tablas, es decir, que cada caso que nos pregunten tenemos que llevarlo a una probabilidad como la que aparece en el gráfico y para un valor de k positivo.

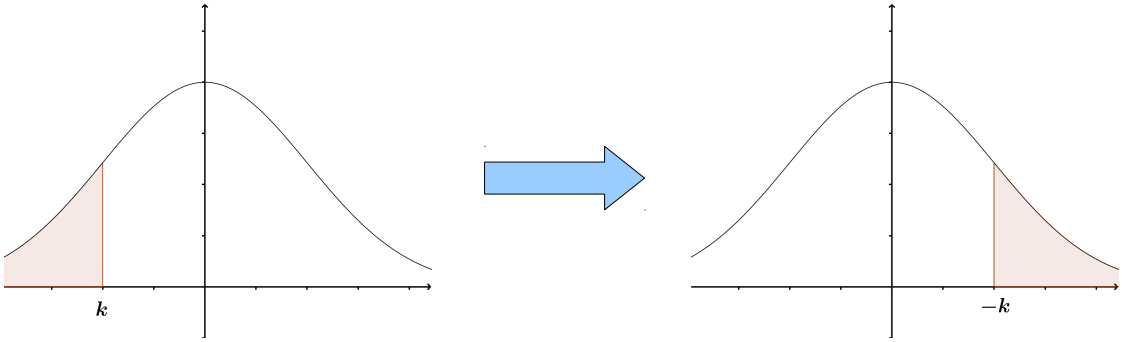
Veamos algunas cosas a tener en cuenta:

· Cada valor de k nos deja dos zonas una a su derecha ($x \geqslant k$) y otra a su izquierda ($x \leqslant k$)



· Siempre es aconsejable pintar, de forma aproximada, la región que nos piden, para ver como llegar a algo parecido a lo del primer gráfico, pues es el que nos permite usar la tabla.

· Si k<0 tenemos que coger su valor opuesto, que es positivo, y trabajar con él. Eso si, tenemos que coger la zona que es igual a la nuestra para el valor opuesto de k.



En el caso que tenemos en las gráficas habrá que calcular la probabilidad del contrario. Por ejemplo:

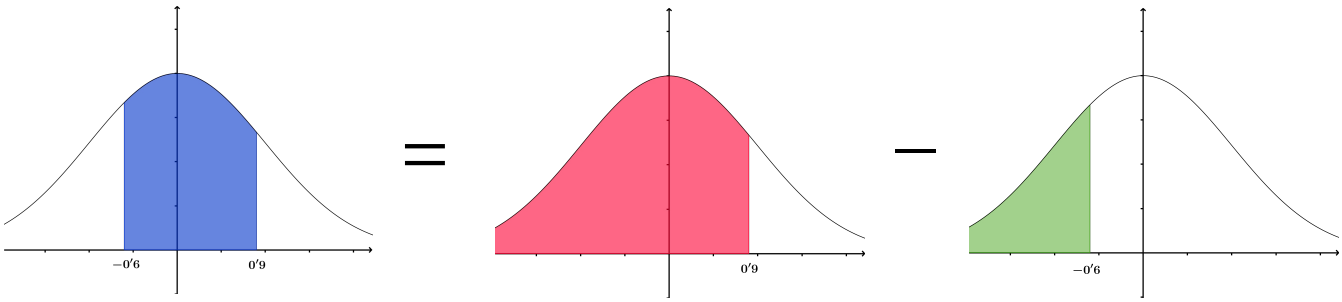
$$P(x\leqslant -1\,'31)=P(x\geqslant 1\,'31)=1-P(x\leqslant 1\,'31)=1-0\,'9049=0\,'0951$$

· Si nos piden calcular la probabilidad de que x esté comprendido entre dos valores, *siempre* se calcula la probabilidad como en el siguiente ejemplo:

$$P(-0\,'6\leqslant x\leqslant 0\,'9)=P(x\leqslant 0\,'9)-P(x\leqslant -0\,'6)$$

Luego en cada caso me tendré que plantear como calcular las dos probabilidades que tengo en dicha recta. Sería bueno hacerlas aparte y luego restar.

Veámoslo gráficamente:



Para ilustrar más los distintos tipos podéis ver los ejemplos que vienen en las páginas 272 y 273.

Si no estuviéramos trabajando con la N(0,1), sino que lo estuviéramos haciendo con otra normal, por ejemplo la N(80,3'5), se hace prácticamente igual, sólo que hay que tipificar la variable, es decir, hay que pasar el valor que nos den por una fórmula que nos lo transforma en un valor de la N(0,1) y entonces actúo como antes dijimos.

Ver el último ejemplo de la página 273.

Vídeo: Distribución normal.