

BLOQUE II

PROBABILIDAD Y MUESTREO

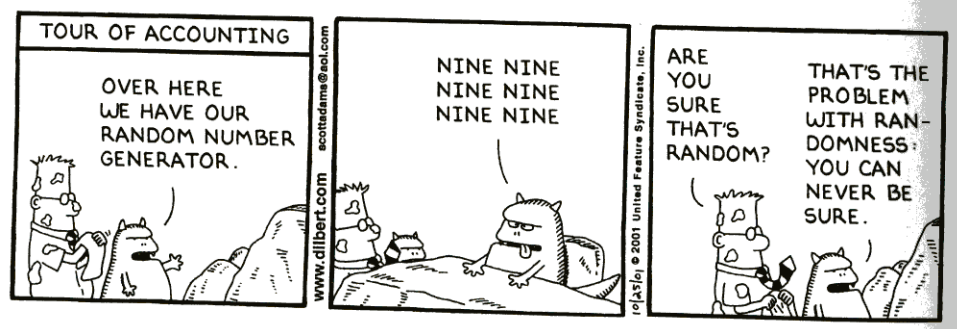
B, G Y H

CURSO 2009-2010

Parte 1

Probabilidad

1.1. Caos v. aleatorio



- Puede que parezcan lo mismo pero:
 - Lo aleatorio sigue unas reglas concretas y si hacemos un número suficiente de ensayos al final veremos un patrón consistente y claro
 - En lo caótico no existiría ese patrón. Nunca habría regularidad (o no somos capaces de detectarla).

1.2. Fenómenos aleatorios

- Hay muchas situaciones en la vida real que pueden ser descritas por medio de fenómenos aleatorios.
- Esas descripciones nos permiten conocer las probabilidades de que ocurran unas cosas u otras.
- Los resúmenes de esas probabilidades se denominan distribuciones o modelos de probabilidad
 - Los juegos de azar siguen modelos de probabilidad uniforme (por ejemplo, todos los números de la lotería tienen la misma probabilidad de salir)
 - Muchos fenómenos naturales siguen un modelo de probabilidad normal
 - Muchos fenómenos económicos siguen modelos de probabilidad lognormal o similares
 - Muchos fenómenos que se producen con poca frecuencia siguen el modelo de la distribución de Poisson
- Finalmente, hay modelos de probabilidad que no se suelen asociar con fenómenos en la realidad pero que se utilizan para los análisis estadísticos como una herramienta más

1.3. Vocabulario básico

- Fenómeno aleatorio: un fenómeno es aleatorio si sabemos los resultados que podrían ocurrir, pero no qué resultados concretos saldrán en cada ocasión
- Intento o ensayo: cada vez que intentamos un fenómeno aleatorio
- Resultado: El resultado de un ensayo es el valor medido, observado, o informado en después de ese intento. Los resultados pueden ser:
 - Discretos: Si producen valores diferenciados tal y como cara/cruz. A menudo luego contamos el número de veces que cada valor ha salido
 - Contínuos: Un valor numérico dentro de un rango de valores posibles
- Experimento aleatorio: Si repetimos muchos intentos con exactamente las mismas condiciones tenemos un experimento. Esto nos dará muchos resultados
- Variable aleatoria: Si recogemos todos los resultados de un experimento aleatorio, tenemos una columna de datos con valores aleatorios
- Probabilidad, la probabilidad es un valor entre 0 y 1 que nos dice como de probable es que ocurra un determinado resultado. Ese valor nos anticipa cuantas veces un resultado dado saldrá del total de veces que repitamos un intento de un fenómeno aleatorio
- Independencia, dos fenómenos son independientes si saber que un suceso ha ocurrido no altera la probabilidad de que ese fenómeno ocurra

1.4. Modelos de probabilidad uniforme

- En estos modelos la probabilidad de un suceso es la misma para todos los resultados.
 - La probabilidad de que salga un uno en un dado es $1/6$.
 - La probabilidad de que salga un número de lotería es 1 dividido por la cantidad de números de lotería que se hagan, etc.
- Cuando trabajamos con probabilidades homogéneas nos podemos hacer una serie de preguntas que se resuelven matemáticamente.
 - Probabilidad de que salga un valor en un experimento en el que hay dos posibles resultados: Aplicamos la distribución de Bernoulli (ejemplo, probabilidad de sacar cara al lanzar una moneda)
 - Probabilidad de tener al menos un acierto en una serie de intentos: Distribución geométrica (cual es la probabilidad de sacar una cara al menos en cinco lanzamientos)
 - Probabilidad de obtener un número concreto de aciertos en una serie de intentos: Distribución binomial (por ejemplo cual es la prob. de 3 caras en cinco lanzamientos). Este modelo lo veremos con más detalle en la sección Section 1.5 ., "Modelo de distribución binomial"
- Estos modelos de probabilidad son sobre todo importantes en los casos de juegos de azar o cosas parecidas. Algunos ejemplos en que podríamos utilizarlas en Psicología son:
 - Queremos evaluar si alguien tiene poderes mentales y comparamos el número de veces que adivina una carta un sujeto con supuesta telepatía
 - Cuantas veces acierta un ratón la puerta correcta para salir de un laberinto
 - Cuántos aciertos podemos esperar al azar al contestar preguntas de un test de alternativas

1.5. Modelo de distribución binomial

- En el examen de la asignatura hay tres alternativas y sólo una de ellas es correcta. Vamos a pensar que ponemos diez preguntas. Si alguien contesta todas las preguntas sin haber estudiado nada tenemos que:
 - La probabilidad de acertar cada pregunta es $1/3$. A esto lo llamamos p =probabilidad de éxito
 - La probabilidad de fallar cada pregunta es $2/3$. A esto lo llamamos q =probabilidad de fracaso
- ¿Cuántos aciertos podemos esperar en el test?
 - Esto se calcula multiplicando el número de ensayos por la probabilidad. En este caso contestamos diez preguntas así que: $np = 10 \times \frac{1}{3} = \frac{10}{3}$. Este valor se llama valor esperado y es equivalente a una media
- Ahora bien, no todo el mundo sacaría un 3. Como es un experimento aleatorio habrá quien sacará más y quien sacará menos. ¿Cuál es la varianza?
 - La varianza se obtiene multiplicando la probabilidad de éxito por la probabilidad de fracaso $pq = \frac{1}{3} \times \frac{2}{3} = \frac{2}{9} = 0.222$
 - La desviación típica es la raíz y produce 0.47
- ¿Qué probabilidad hay de sacar un 5 en el examen sin haber estudiado? La fórmula que calcula las probabilidades binomiales es .

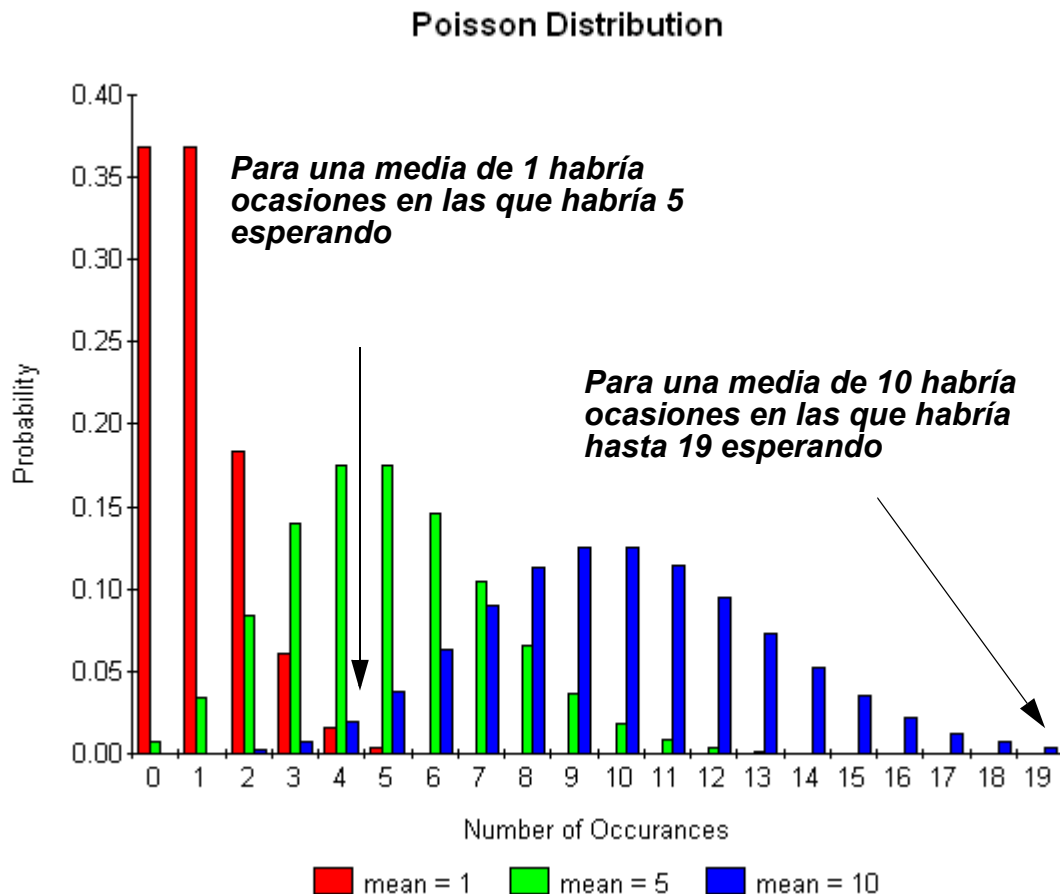
$$P(X = x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

- Esta fórmula es un poco complicada así que es conveniente hacerla con calculadora o por ordenador. Una página (hay muchas) que ofrece este cálculo es esta: <http://stattrek.com/Tables/Binomial.aspx>
- En nuestro caso $P(X = 5)$ por lo que el resultado es 0.13

- ¿Qué probabilidad hay de sacar un cinco o más de un 5 sin haber estudiado nada?
 - Esto se calcula sumando la probabilidad de sacar un 5, 6, 7, 8, 9, o 10
 - En la calculadora online que he puesto antes esto se calcula automáticamente. La probabilidad de aprobar sin haber estudiado y contestando al azar es de 0.21 (antes de poneros contentos recordar que en el examen, las preguntas falladas descuentan por lo que esos cálculos no se aplican).

1.6. El modelo de Poisson

- El modelo de Poisson es un modelo de distribución de probabilidad para datos discretos
 - Este modelo surge como el número de veces que ocurre un fenómeno que tiene una probabilidad muy baja de ocurrir
 - Lo que controla esa distribución es la media del número de veces que ese fenómeno ocurre, pero lo interesante es ver para esa media cual es la probabilidad que se den ciertos valores extremos.
- Supongamos que la media de gente esperando en una cola para que te atiendan es de 1, 5 o 10 personas



- En estos modelos, los resultados son generalmente recuentos de sucesos que ocurren en una serie de grupos
 - En la II guerra mundial, dividiendo Londres en zonas, el modelo de Poisson describía el número de cohetes que cayeron en cada zona (supuesto de independencia)
 - El número de accidentes en un periodo de tiempo (por ejemplo por día)
 - El número de personas que están en una cola por periodo de tiempo
 - El número de casos de leucemia por ciudad puede ser modelado con Poisson. Veremos este ejemplo con más detalle.

- En la película ***Una acción civil***, John Travolta es un abogado que intenta ganar un caso de contaminación en una ciudad donde ha habido al parecer muchos casos de leucemia
 - En la ciudad de Woburn ocurrieron 7 casos de leucemia
 - En Estados Unidos hay 30.800 casos de leucemia al año y una población de 280 millones de personas. En Woburn había 35000 habitantes así que tocarían 3.85 casos de promedio
 - Para ver si 7 casos es mucho o poco podríamos usar varios métodos (la binomial por ejemplo) pero el cálculo tiene partes que ni siquiera con ordenador se harían fácilmente. Es mejor usar Poisson.
 - La fórmula para Poisson es: (una calculadora de probabilidades se puede encontrar [aquí](#))

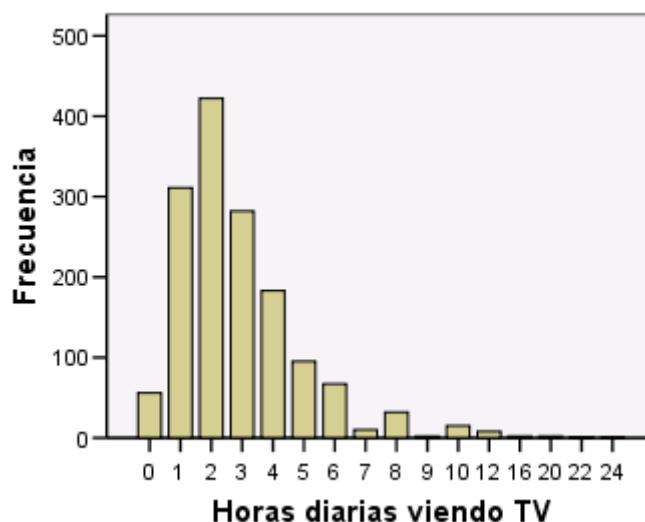
$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

- λ es 3.85 en este caso (es la media de sucesos del fenómeno a estudiar). Con esta fórmula podemos calcular la probabilidad de que haya cero enfermos, un enfermo, dos enfermos, etc. x es el número de casos
- Puesto que en Woburn hubo 7 casos, podemos calcular la probabilidad de exactamente 7 casos. Esto da 0.05 (5%).
- Si dividimos Estados Unidos en agrupaciones de 35000 habitantes tendríamos 8000. El 5% de 8000 es 400. Por tanto, si el modelo de Poisson es correcto para el reparto de casos de leucemia podemos esperar hasta 400 ciudades de ese tamaño con exactamente 7 casos de leucemia por año. 7 ya no parece tan exagerado dado este resultado (también podríamos calcular en cuantos casos hay 7 o más de 7: sería 0.09).
- ¿Qué hubiera sido un valor realmente extremo? Uno que sólo hubiera 1 caso en todo Estados Unidos. Para obtener este valor necesitaríamos que la probabilidad fuera 13 o 14 casos de leucemia (casi el doble de los ocurridos en Woburn)

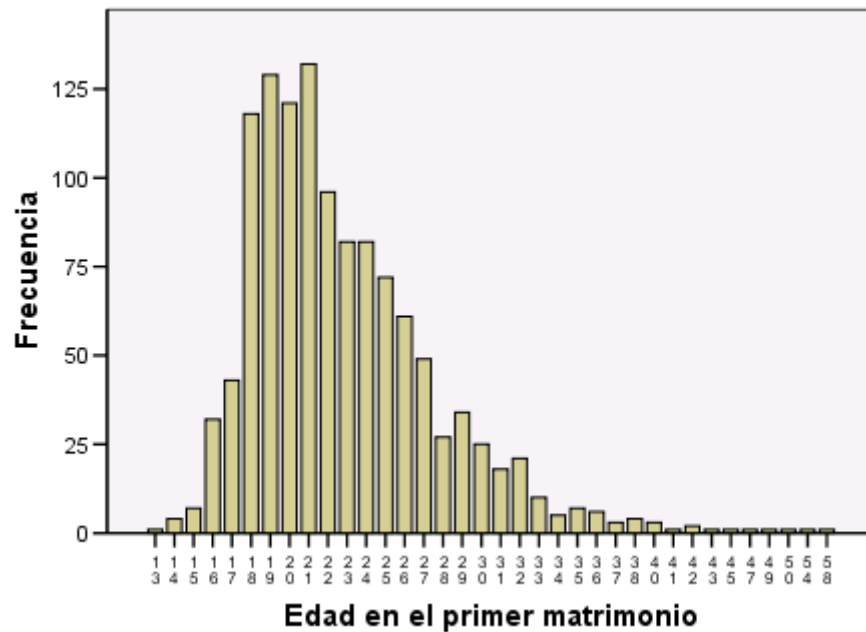
- En resumen, aunque 7 casos parece mucho, usando el razonamiento anterior este resultado no es tan extraordinario como pudiera parecer al principio. Se puede esperar que unas 700 de esas supuestas ciudades del mismo tamaño que Woburn en Estados Unidos tuvieran al menos ese número o más de enfermos
- Conclusiones sobre Poisson
 - Este modelo es apropiado para cosas que no ocurren muy a menudo pero hay muchos intentos en que pueden ocurrir
 - A menudo en este tipo de situaciones las impresiones subjetivas que tenemos las personas no son adecuadas (es difícil valorar si algo es mucho o poco a ojo). Pongamos el caso de 7 o 13 casos de leucemia en Woburn: nuestra impresión subjetiva puede ser muy diferente antes que después de hacer los cálculos

ACTIVIDADES

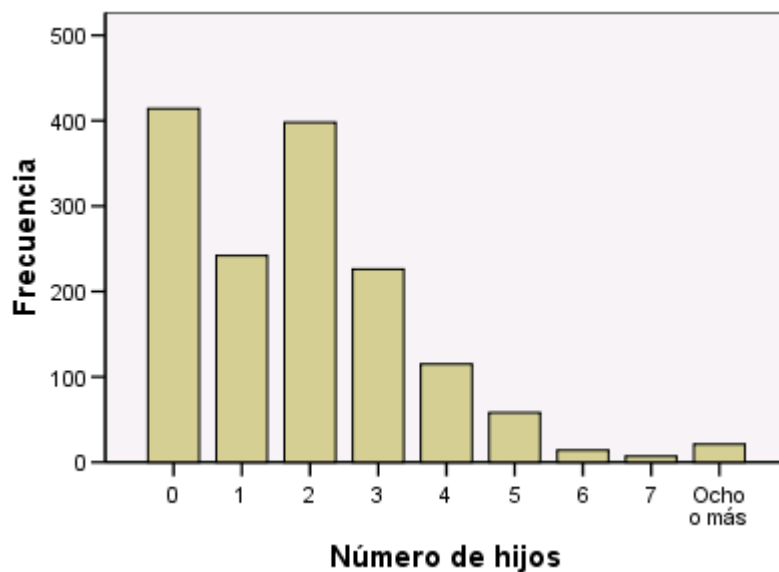
- Ejercicio 1.6.1 ¿Qué distribución de probabilidad teórica crees que se parecerá más a las contestaciones a la pregunta: ¿Cuántas horas pasas al día viendo la televisión?
- Ejercicio 1.6.2 Abajo se muestra el diagrama de barras de las contestaciones de 1500 sujetos. ¿A qué distribución de probabilidad teóricas dirías que se parece más este diagrama de barras?



Ejercicio 1.6.3 ¿A qué distribución teórica dirías que se parece la edad en años del primer matrimonio de ese grupo de 1500 personas?



Ejercicio 1.6.4 ¿Y la distribución del número de hijos? ¿Dirías que hay irregularidades con respecto a un modelo de Poisson? ¿Por qué?



1.7. Modelo de probabilidad normal

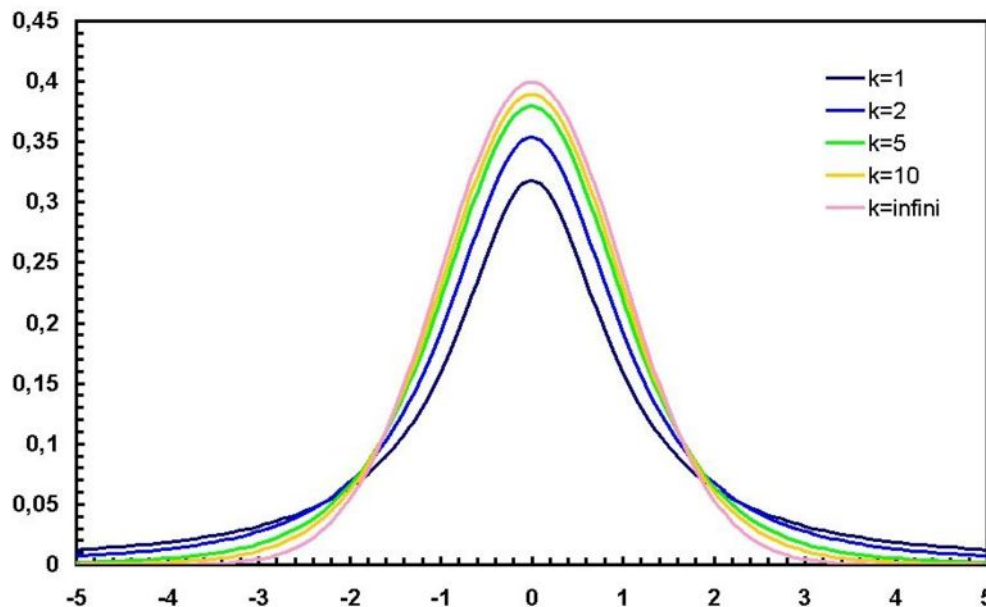
- Ya hemos visto el tema de la distribución normal en Section 3.32. Aquí sólo comentar que la distribución normal es la apropiada para fenómenos que tienen un punto medio que es más probable y que los resultados alejados de la media son progresivamente menos probables.
- El modelo de probabilidad normal es apropiado para muchos fenómenos. Cualquier situación en la que se produzcan una serie de cambios pequeños y aditivos (que suman o restan) producirán datos que siguen aproximadamente la distribución normal. Si los efectos no son aditivos, sino multiplicativos entonces las distribuciones serán del tipo de la lognormal (Section)
 - Ya vimos varios ejemplos en la Section de fenómenos que siguen la distribución normal
 - En general, se piensa que muchos fenómenos siguen un modelo de distribución normal lo suficientemente bien como para que la usemos como aproximación

1.8. La distribución t , χ^2 , y la F

- Estas distribuciones no corresponden con fenómenos de la naturaleza per se tal y como las anteriores
 - Su importancia radica en su utilización para fines estadísticos
 - Son, como la distribución normal, distribuciones continuas. Eso significa que se puede calcular probabilidades por encima o por debajo de número con un número de decimales cualquiera (a diferencia de las distribuciones discretas en que se calcula la probabilidad de valores discretos)

1.9. La distribución t

- La t está relacionada con la distribución de las medias de muestras pequeñas extraídas sucesivamente de una población
- Su forma exacta depende del número de elementos que hay en cada caso
 - El número de elementos se denomina los grados de libertad
 - Existe una distribución t para 1 grado de libertad, para 2 grados de libertad, etc.
 - Cuando el número de grados de libertad es cercano a 30 la distribución t es casi igual a la normal
- La forma de la distribución t se puede ver en el siguiente gráfico



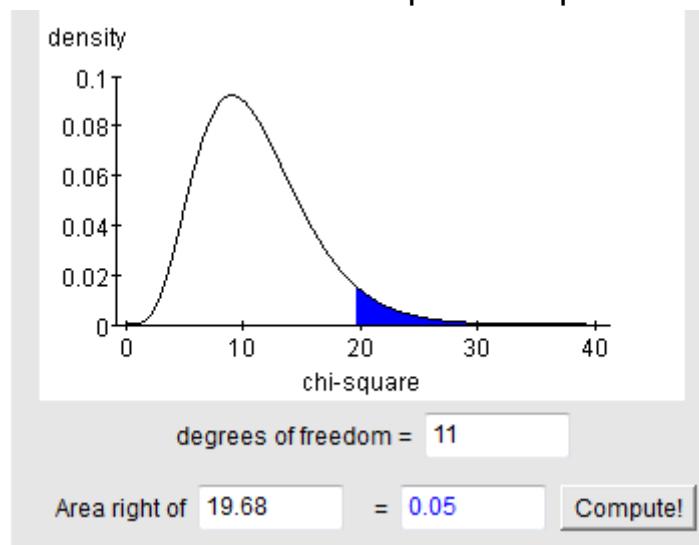
(tomado de la wikipedia)

- La línea con $k=\infty$ es la de la distribución normal (es la más alta)
- La de $k=1$ es la más baja y con brazos más anchos

- ¿Es importante la diferencia entre la normal y la distribución t?
 - Supongamos que le vamos a dar un ordenador gratis al 5% superior de los estudiantes de un curso de estadística
 - No obstante, sólo hay 6 estudiantes así que simplemente se lo damos al primer estudiante
 - No obstante, en los resultados tenemos dos estudiantes que han sacado una nota muy buena. En puntuaciones típicas uno de ellos ha sacado un +2 y el otro un +2.5 (al que le hemos dado el ordenador).
 - ¿Se merecería el sujeto de +2 también el ordenador?
 - El estudiante razona que si asumimos que las notas de la asignatura siguen la distribución normal en la población un +2 supone dejar por debajo al 97.7% de la población
 - Sin embargo, puesto que la muestra es pequeña sería más apropiado utilizar la distribución t con 5 grados de libertad (se utiliza $n-1$ para calcular los grados de libertad en este caso). En este caso, +2 dejaría por debajo de sí al 94.9% de la población así que no le correspondería ordenador (aunque por tan poco que en fin...)

1.10. La distribución χ^2

- La distribución χ^2 surge en muchas situaciones pero quizás la más importante es cuando se tiene que evaluar la diferencia entre unos valores esperados y unos valores observados
 - Esto lleva a un test que permite valorar hasta qué punto estas diferencias son grandes o no teniendo en cuenta los grados de libertad (los cuales a su vez dependen del número de elementos en los datos)
- En la Sección 2.16. vimos un ejemplo sobre si los signos del zodiaco estaban repartidos de una manera homogénea entre personas de éxito
 - Al final de ese ejercicio calculamos un valor de χ^2 que permitía decidir si en general los signos se repartían de manera homogénea o si por el contrario había más abundancia de algunos signos. El valor que obtuvimos fue $\chi^2 = 5,904$
 - No obstante, en aquel momento no podíamos decir si eso era mucho o era poco. Para hacer esa valoración podemos utilizar la distribución χ^2
 - Los grados de libertad en este caso es el número de signos del zodiaco menos 1 (en este caso 11)
 - Un valor alto de χ^2 con 11 grados de libertad es el que deja por debajo de sí el 95% y el 5% por encima. Este valor lo podemos calcular con esta calculadora de χ^2 <http://www.stat.tamu.edu/~west/applets/chisqdemo.html>
 - En esa calculadora ponemos 11 en la casilla degrees of freedom y 0.05 en la casilla azul. El valor que nos aparece es 19.68. La parte



blanca corresponde con valores que si salen indican que las

diferencias entre signos del zodiaco no son importantes. Si en cambio salen los que están en azul, las diferencias son importantes. En nuestro caso salió 5.904 luego las diferencias no son importantes

- En la Sección 2.19. vimos si hubo mayor supervivencia según la clase en la que viajaban en el Titanic.
 - El resultado que obtuvimos fue $\chi^2 = 188.4$
 - Los grados de libertad en una tabla de contingencias como estas se calculan multiplicando el número de filas menos 1, por el número de columnas menos 1. En nuestro caso teníamos cuatro clases de pasajeros y dos categorías de supervivencia $(4 - 1) \times (2 - 1) = 3$. Usando la calculadora de antes con 3 en degrees of freedom and 0.05 en la casilla azul nos da un valor de 7.815. Si el valor que nos sale es mayor que ese valor entonces sí que hay diferencias entre los valores esperados y los observados (el valor es 188.4 así que las diferencias están claras)

1.11. La distribución F

- La F aparece como una forma de comprobar si la cantidad de varianza explicada en, por ejemplo, una regresión es mucha o poca.
 - En la sección Sección 4.15. vimos una fórmula que nos permitía valorar el ajuste de la recta. Esa fórmula era la siguiente:

$$R^2 = \frac{SCR}{SCT} \quad .$$

- Recordar también que usábamos esta fórmula para calcular SCR

$$SCR = SCT - SCE$$

- Una forma alternativa y que es muy usada en la práctica para calcular si una recta de regresión u otra técnica estadística explica mucha o poca varianza es hacer lo siguiente:

$$\frac{SCR}{SCE}$$

- El valor obtenido se puede comparar con una distribución F con n grados de libertad en el numerador y n grados de libertad en el denominador

Ejemplo

En el ejemplo de la relación entre la edad y el CPK. El valor de SCR es 315142 y el del SCE es 1112685. El cociente es 4.53. En este caso el numerador tiene 1 grado de libertad y el denominador tiene 16 grados de libertad. Utilizando la calculadora que está aquí: <http://www.stat.tamu.edu/~west/applets/fdemo.html> hacemos lo siguiente

Este valor es ligeramente mayor que el de 4.53 por lo que la edad explica varianza del CPK

