PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE LETRAS

Métodos y técnicas de investigación cuantitativa

César Antonio Aguilar

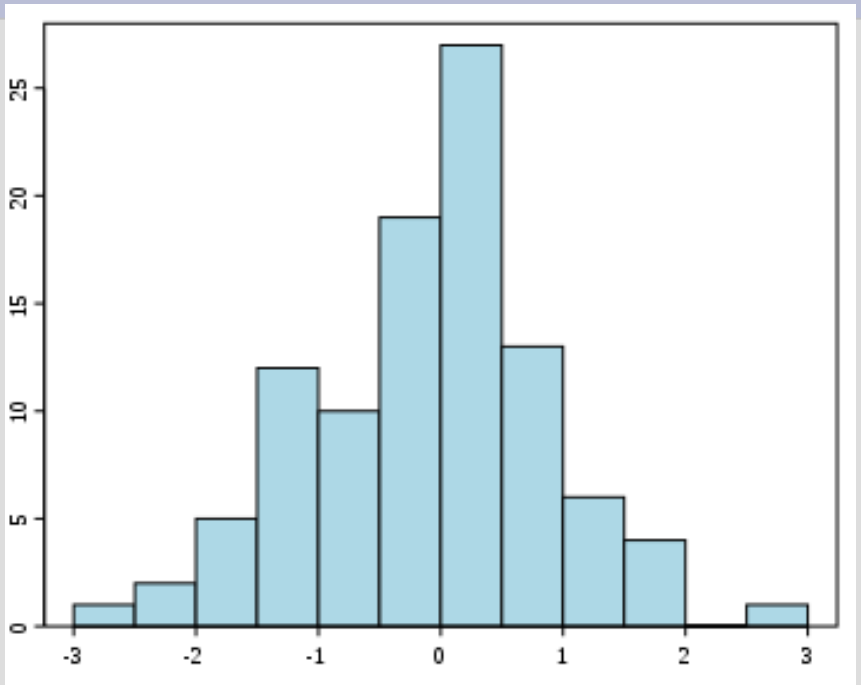
Facultad de Lenguas y Letras

22/04/2013

Cesar.Aguilar72@gmail.com

Distribución de frecuencias (1)

En esta sesión, retomaremos nuestro análisis de poblaciones, ahora enfocándonos en lo que se conoce como **distribución de frecuencias**, esto es, la agrupación de datos que son excluyentes entre sí, y que reportan una serie de observaciones realizadas conforme criterios (o categorías) específicas.



Esto justo nos sirve para comparar y describir el comportamiento de dos o más poblaciones ante un mismo fenómeno. Veamos en detalle cómo opera esto.

Distribución de frecuencias (2)

Podemos empezar con el siguiente supuesto: cuando analizamos un fenómeno en dos contextos diferentes, podemos pensar que éste, si bien es un hecho singular, deberá presentar alguna particular dada la diferencia entre dichos contextos.

Así, lo que nos interesa representar en esta clase de gráficas son justo las variaciones que detectamos en las poblaciones que estudiamos. No descartamos las coincidencias (de hecho, son las que nos sirven de guía a la hora de recopilar los datos), pero lo que esperamos reconocer es justo serie de particularidades.

Vamos a denominar al conjunto de coincidencias como la representación de una **tendencia central**, y a las diferencias las reconoceremos como **variables**.

Distribución de frecuencias (3)

Veamos un ejemplo. Los siguientes datos son un ejemplo de un test para medir sus habilidades como aprendices de inglés, aplicado dos grupos de estudiantes. El primero se aplica dentro del marco de un laboratorio de idiomas, y el segundo se aplicó en el contexto de un curso tradicional.

La calificación máxima es de 20 puntos, y de ahí la escala es descendente hasta 0.

Marks out of 20	
Group A (Language laboratory) (N = 30)	*Group B (Traditional) (N = 30)*
15	11
12	16
11	14
18	18
15	6
15	8
9	9
19	14
14	12
13	12
11	10
12	15
18	12
15	9
16	13
14	16
16	17
17	12
15	8
17	7
13	15
14	5
13	14
15	13
17	13
19	12
17	11
18	13
16	11
14	7

Distribución de frecuencias (4)

Score	Group A		Group B	
5			/	1
6			/	1
7			//	2
8			//	2
9	/	1	//	2
10			/	1
11	//	2	///	3
12	//	2	~~////~~	5
13	///	3	////	4
14	////	4	///	3
15	~~////~~ /	6	//	2
16	///	3	//	2
17	////	4	/	1
18	///	3	/	1
19	//	2		

Hagamos entonces nuestra primera comparación, contrastando los datos que nos arrojan ambos grupos, de manera que podamos determinar qué rasgos hay que podamos reconocer como similares, y qué rasgos son distintos:

Pregunta: ¿reconocen alguna diferencia o similitud entre los grupos A y B? ¿Pueden describirlas?

Variables dependientes/independientes (1)

Hagamos entonces nuestra primera comparación, considerando dos tipos de variables, las cuales vamos a reconocer como **dependientes** o **independientes**, lo que nos ayudará a apreciar mejor cuáles son los factores que influyen en las diferencias/similitudes en los puntajes:

Independent variable	*Dependent variable*
Type of teaching method	Score on language test
Sentence complexity	No. of sentences recalled correctly by subjects
Social class	Percentage of *-ing* endings pronounced [ɪn] by subjects
Topic of newspaper article	Sentence length

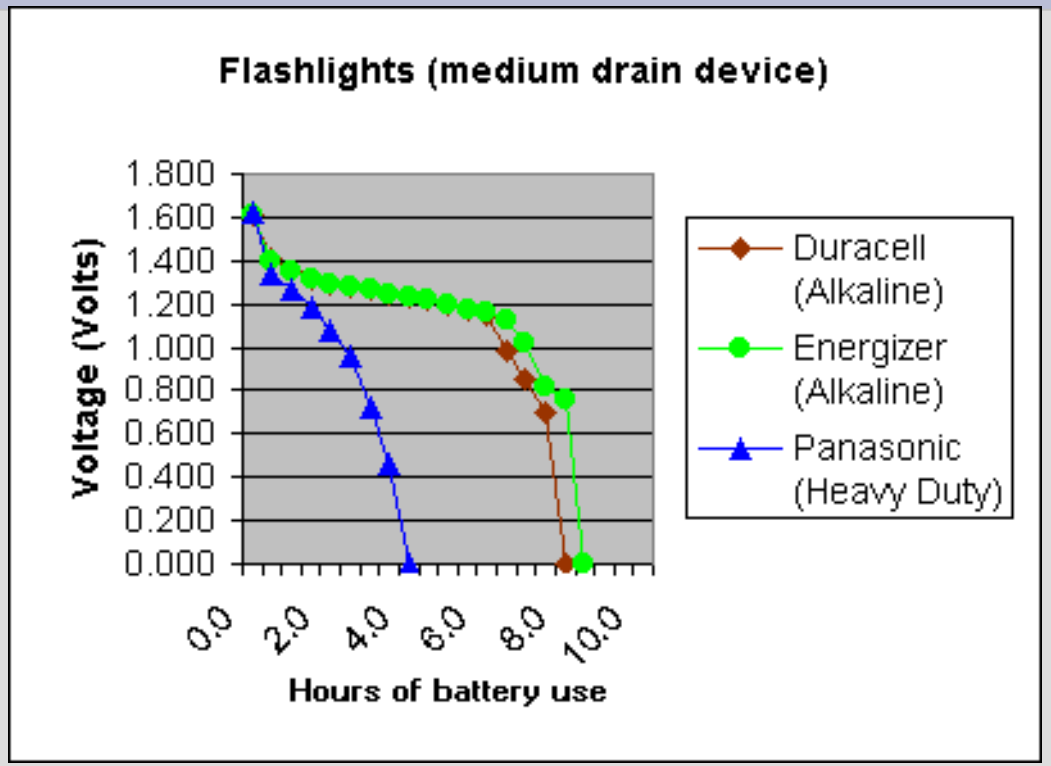
Pregunta: ¿qué otro factor considerarían importante, y que no esté considerado?

Variables dependientes/independientes (2)

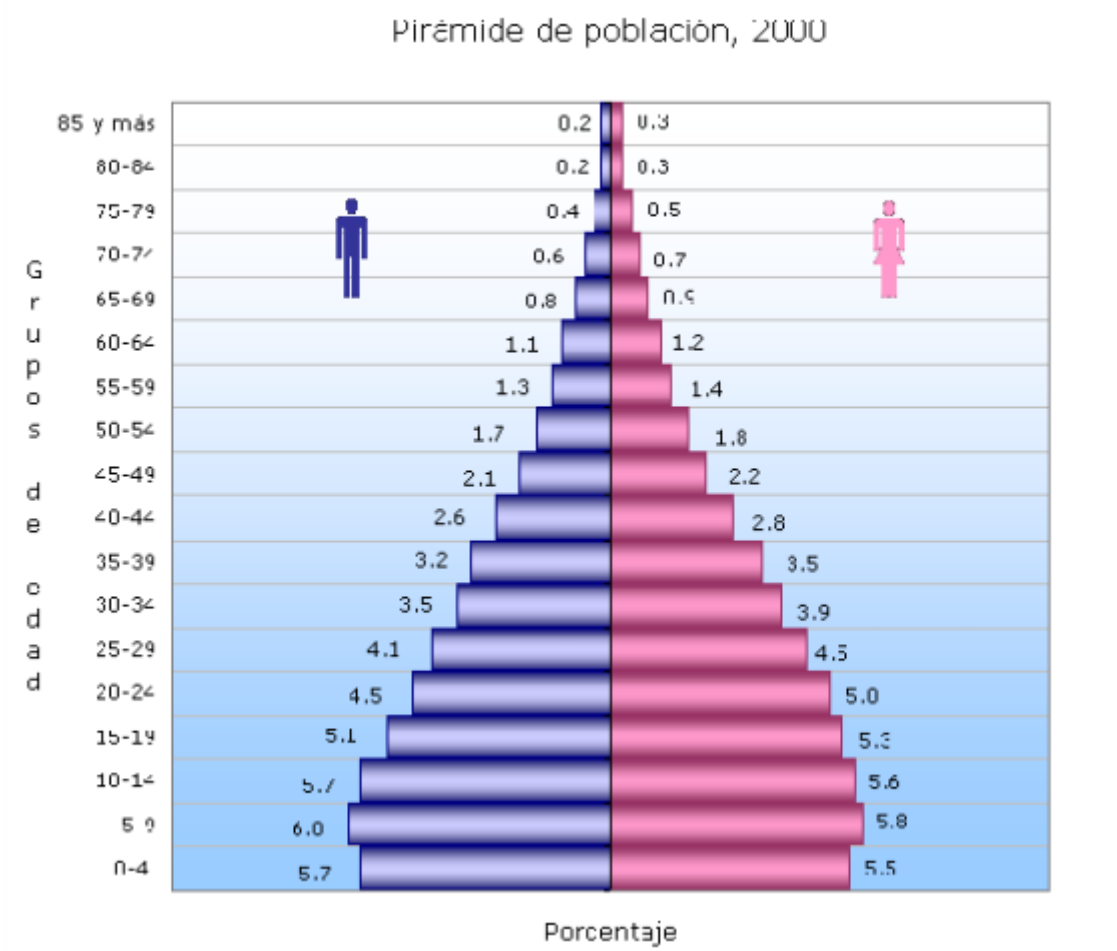
¿Qué vamos a entender por variables dependientes e independientes?

En el caso de las primeras, las consideramos así por la ausencia de relaciones que pueden tener entre sí, de modo que podemos tomarlas como nuestros objetos de investigación.

En el caso de las segundas, ocurre lo contrario: son variables que se influyen unas con otras,



Variables discretas/continuas (1)



Igualmente, otro tipo de rasgos que asumen los valores son los de ser **discretos** o **continuos**.

Decimos que son discretos cuando reconocemos intervalos o separaciones únicamente dentro de una escala determinada.

Variables discretas/continuas (2)

Por su parte, las variables continuas son aquellas que pueden asumir cualquier valor dentro de un intervalo específico. Por ejemplo, comparar pesos con tallas de cinturas:

Talla (cms)	160	165	168	170	171	175	175	180	180	182
Peso (kgs)	55	58	58	61	67	62	66	74	79	83

En este ejemplo, si bien establecemos una secuencia de mayor a menor, si notan los intervalos no son exactos, justo porque es difícil encontrar coincidencias simétricas entre los sujetos de nuestra población de estudio.

Ejercicio

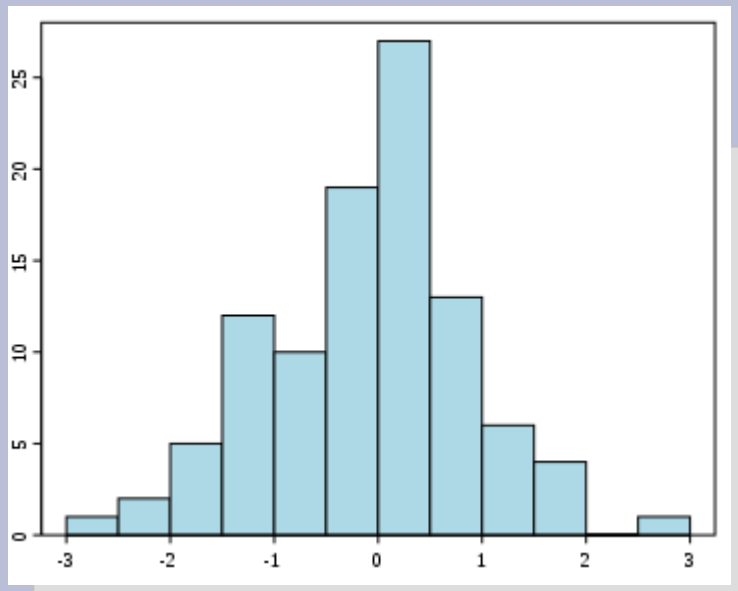
Traten de resolver las siguientes preguntas:

1. ¿Consideran pertinentes o no las variables dependientes/independientes que se han formulado para el análisis de los resultados del test?

2. ¿Qué tipo de valor asumen nuestras variables a analizar: son discretas o continuas?

3. Revisando los datos, ¿qué podríamos decir respecto a las habilidades que muestras los estudiantes tras aplicarles el test: hay alguna diferencia sustancial entre aplicarlo en un laboratorio, o en una sesión tradicional de clase?

Representación gráfica (1)



Si buen uno puede interpretar datos números y hacerse una idea respecto a los fenómenos que son detectables en un análisis estadístico, la ventaja de esta clase de análisis es que podemos graficar nuestros resultados de distintas maneras.

Así, tenemos un modelo gráfico llamado **histograma**, el cual es un una representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados. En el eje vertical se representan las frecuencias, y en el eje horizontal los valores de las variables, normalmente señalando las marcas de clase, es decir, la mitad del intervalo en el que están agrupados los datos.

Representación gráfica (2)

¿Cómo se elabora un histograma? Veamos la receta que nos da Wikipedia al respecto:

Paso 1: Determinar el rango de los datos. Rango es igual al dato mayor menos el dato menor.

Paso 2: Obtener los números de clases, existen varios criterios para determinar el número de clases (o barras). Sin embargo ninguno de ellos es exacto. Algunos autores recomiendan de cinco a quince clases, dependiendo de cómo estén los datos y cuántos sean. Un criterio usado frecuentemente es que el número de clases debe ser aproximadamente a la raíz cuadrada del número de datos. Por ejemplo, la raíz cuadrada de 30 (número de artículos) es mayor que cinco, por lo que se seleccionan seis clases.

Representación gráfica (3)

Paso 3: Establecer la longitud de clase: es igual al rango dividido por el número de clases.

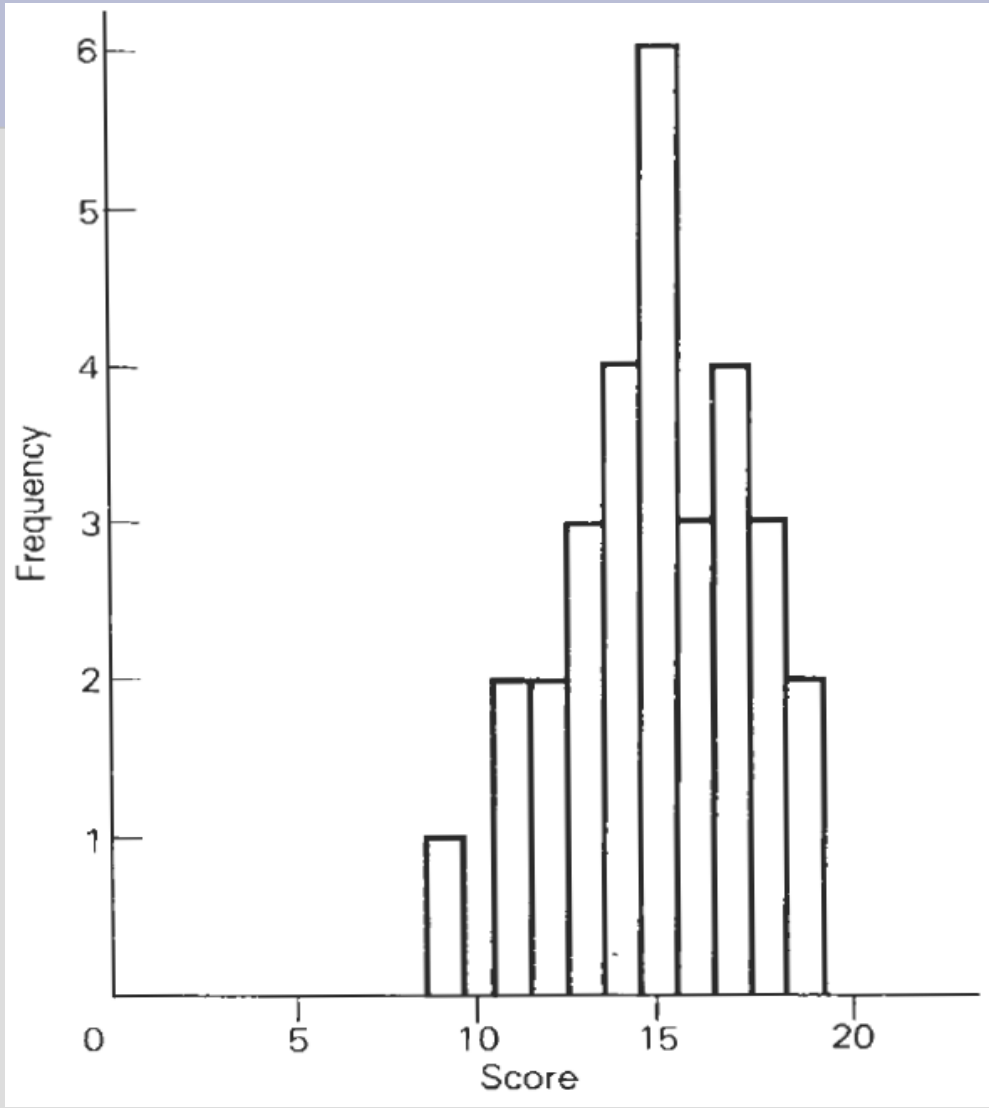
Paso 4: Construir los intervalos de clases: Los intervalos resultan de dividir el rango de los datos en relación al resultado del PASO 2 en intervalos iguales.

Paso 5: Graficar el histograma. En caso de que las clases sean todas de la misma amplitud, se hace un gráfico de barras, las bases de las barras son los intervalos de clases y altura son la frecuencia de las clases. Si se unen los puntos medios de la base superior de los rectángulos se obtiene el polígono de frecuencias.

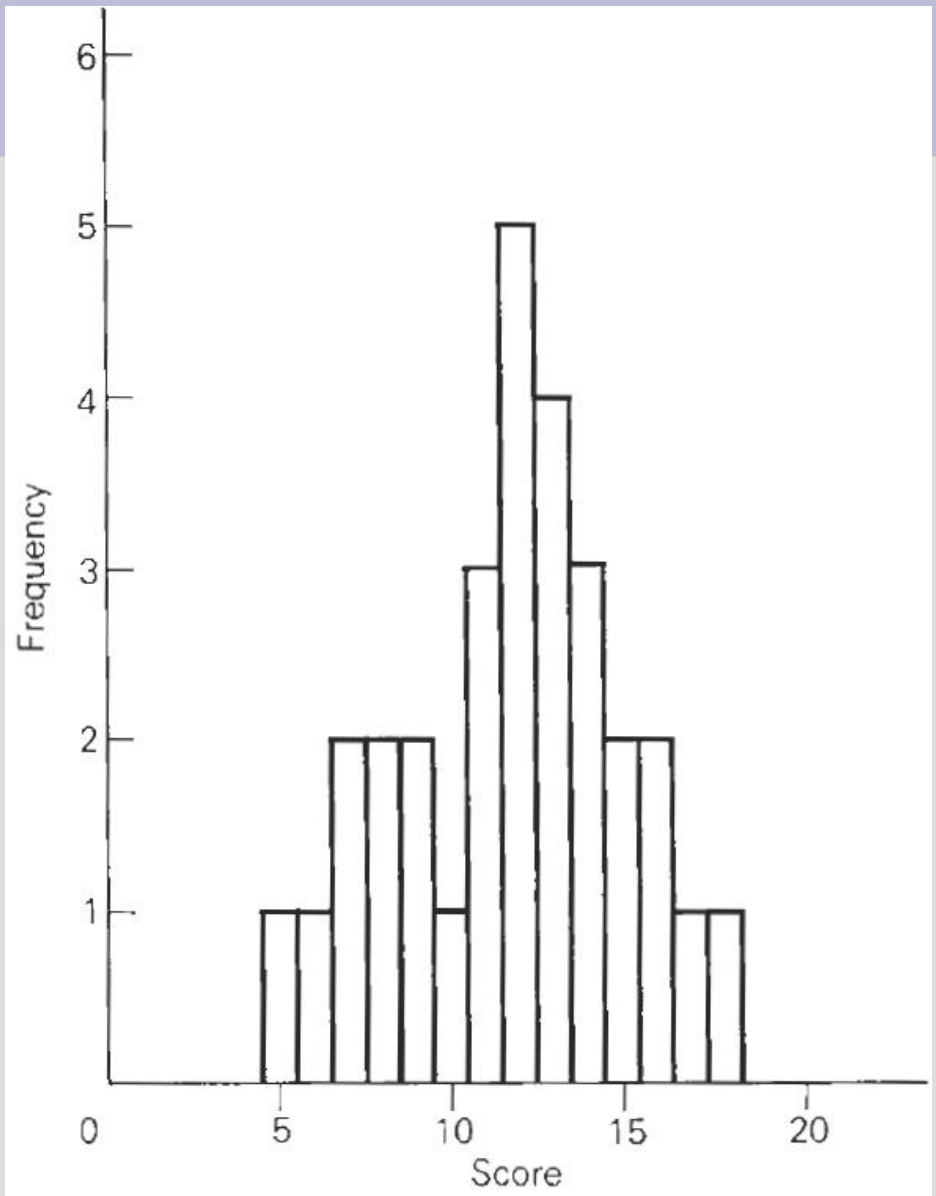
Representación gráfica (4)

Vamos a ver cómo quedan graficados los datos que tenemos.

En el caso del grupo A vemos:



Representación gráfica (5)

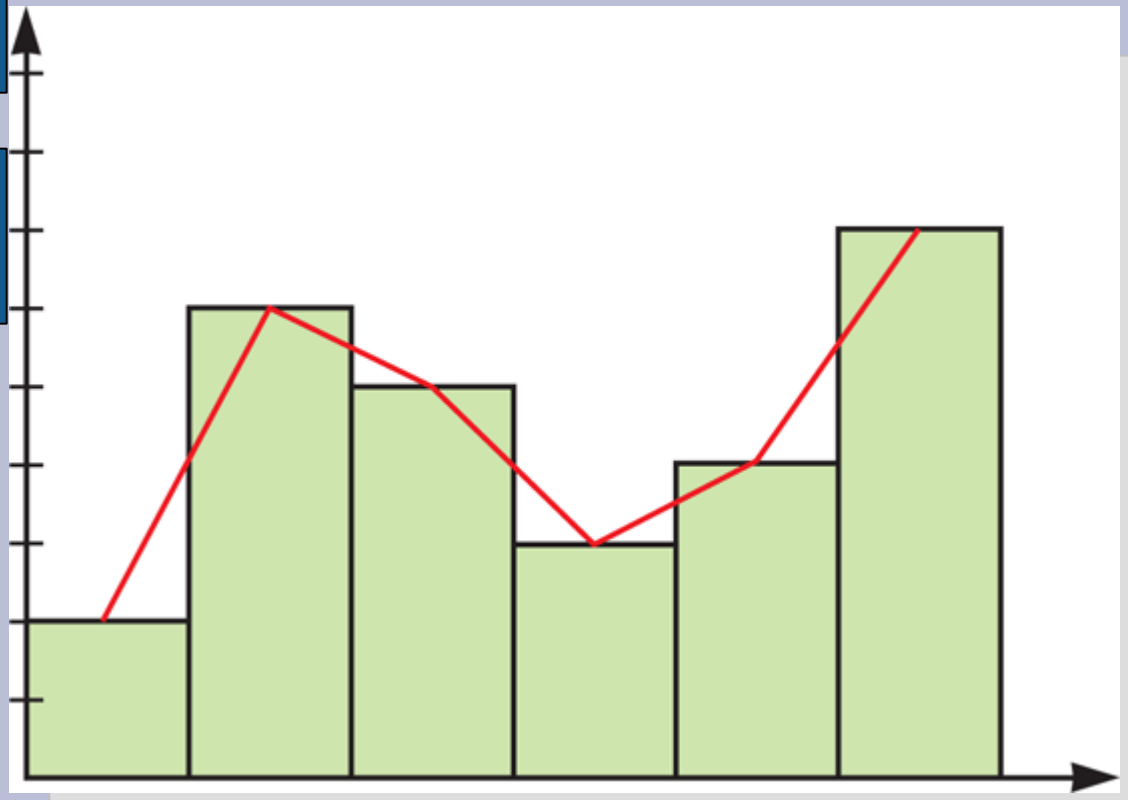


Y en el caso del grupo B observamos:

Ejercicio

Observando las gráficas: ¿qué similitudes/diferencias reconocen entre los dos grupos estudiados?

Representación gráfica (6)



Ahora bien, supongamos que queremos comparar estos datos en una misma gráfica. Podemos entonces por usar un **polígono de frecuencias**, que es es un gráfico que se realiza a través de la unión de los puntos más altos de las columnas en un histograma de frecuencia (el cual utiliza columnas verticales para mostrar tales frecuencias).

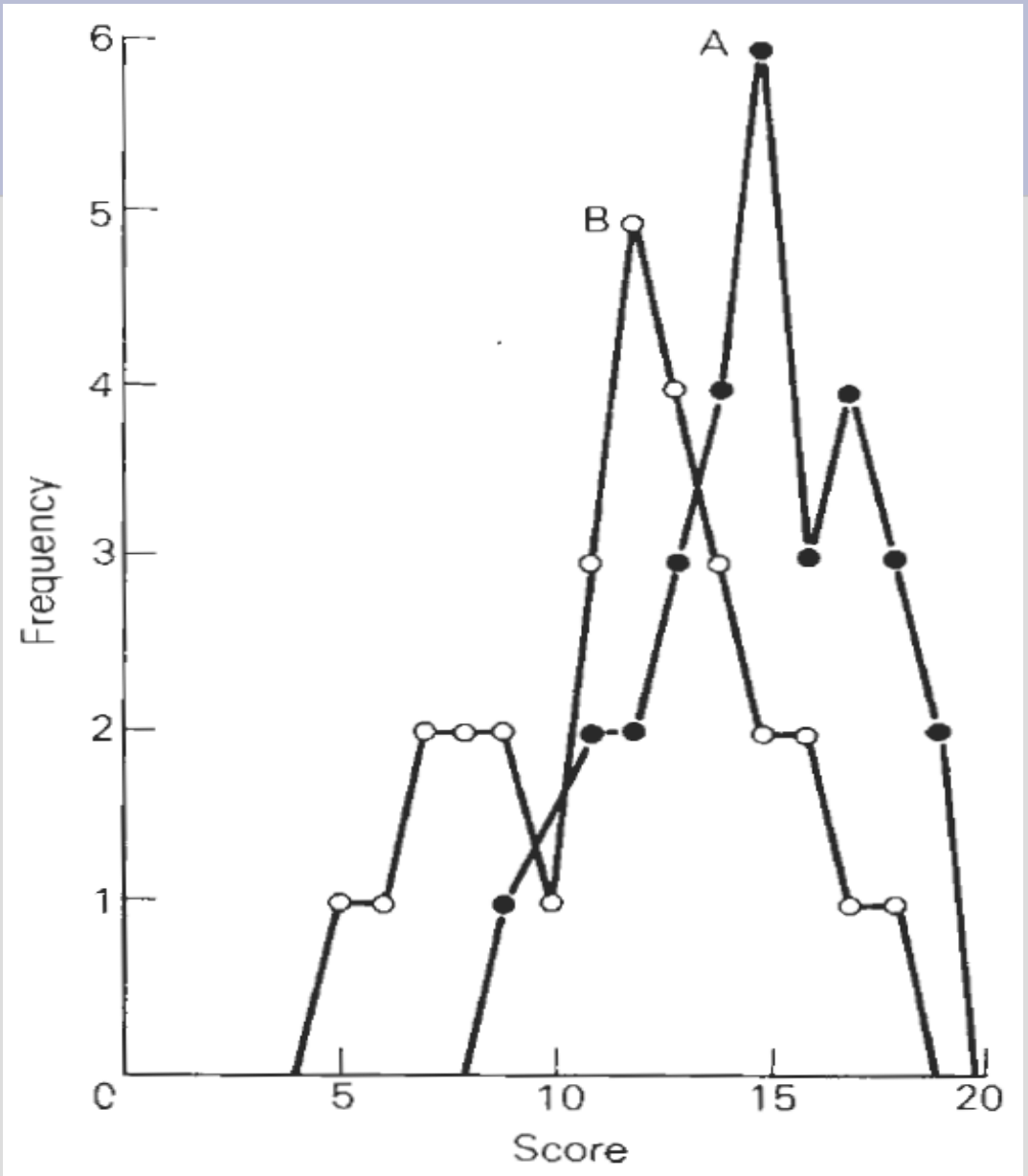
Un ejemplo simple:

Representación gráfica (7)

¿Cómo se ven nuestros datos cuando los contrastamos unos con otros?

Ésta es la gráfica:

Pregunta: ahora sí, ¿qué rasgos podemos reconocer entre los dos grupos?

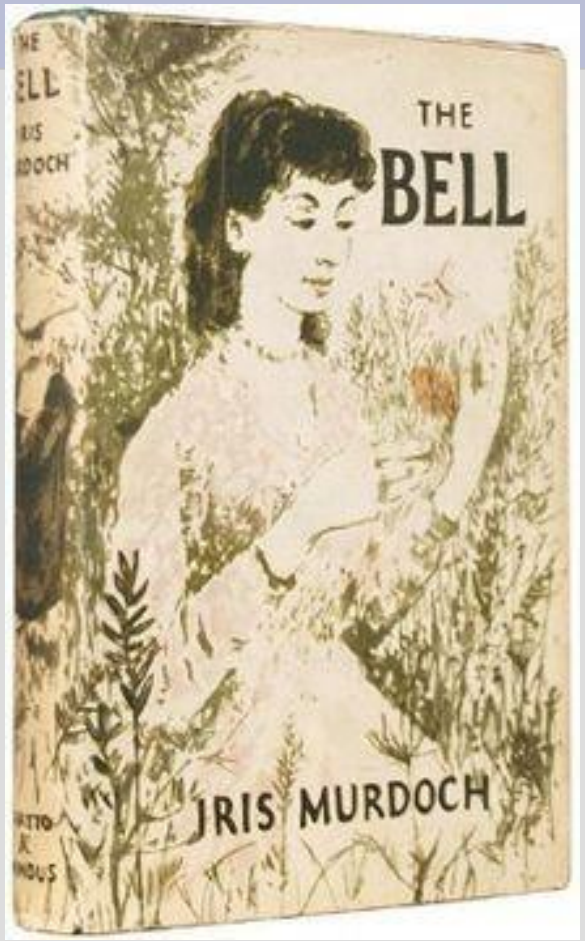


Contando palabras (1)

Veamos otro ejemplo que podemos aplicar al estudio de corpus lingüísticos. En este caso, vamos a medir la extensión de palabras que tienen las oraciones de una novela en inglés (*The Bell*, de Iris Murdoch, 1958).

La idea aquí es hacer este cálculo a partir de las primeras 100 oraciones de dicha novela.

Para mayores detalles, revisen el capítulo 2 del libro de Butler (al cual pueden acceder a través del blog del curso).



Contando palabras (2)

Hagamos un listado de nuestras oraciones:

Sentence length (no. words)	Frequency	Sentence length (no. words)	Frequency
3	1	23	4
4		24	2
5	1	25	4
6	2	26	
7	2	27	1
8	8	28	3
9	3	29	3
10	3	30	2
11	5	31	1
12	3	32	2
13	3	33	2
14	8	34	1
15	7	35	
16	3	36	2
17	4	37	1
18	1	38	1
19	6	39	
20	4	40	1
21	2	41	
22	3	42	1

Contando palabras (3)

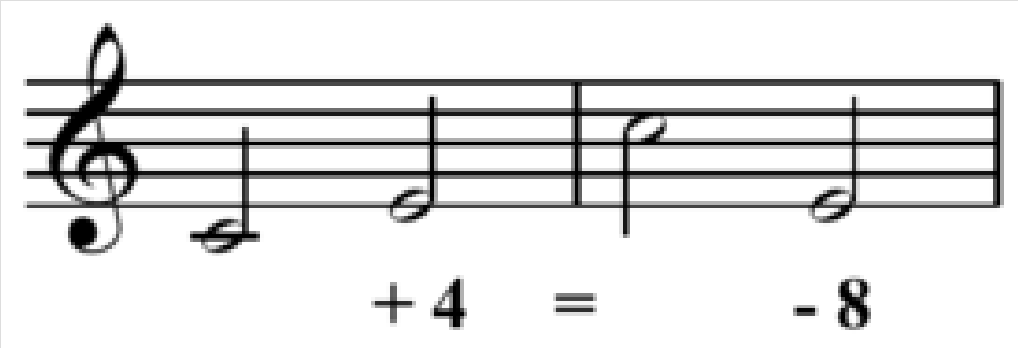
Al igual que el caso anterior, no nos queda del todo claro determinar cuál es la oración más corta o más extensa. Para poder reconocer esto, podemos agrupar tales oraciones en **intervalos de clase**, esto es, dividirlos en conjuntos específicos valores numéricos, de modo que podamos establecer una escala continua. Por ejemplo:

Sentence length (no. words)	*Frequency*
1- 5	2
6-10	18
11-15	26
16-20	18
21-25	15
26-30	9
31-35	6
36-40	5
41-45	1

Contando palabras (4)

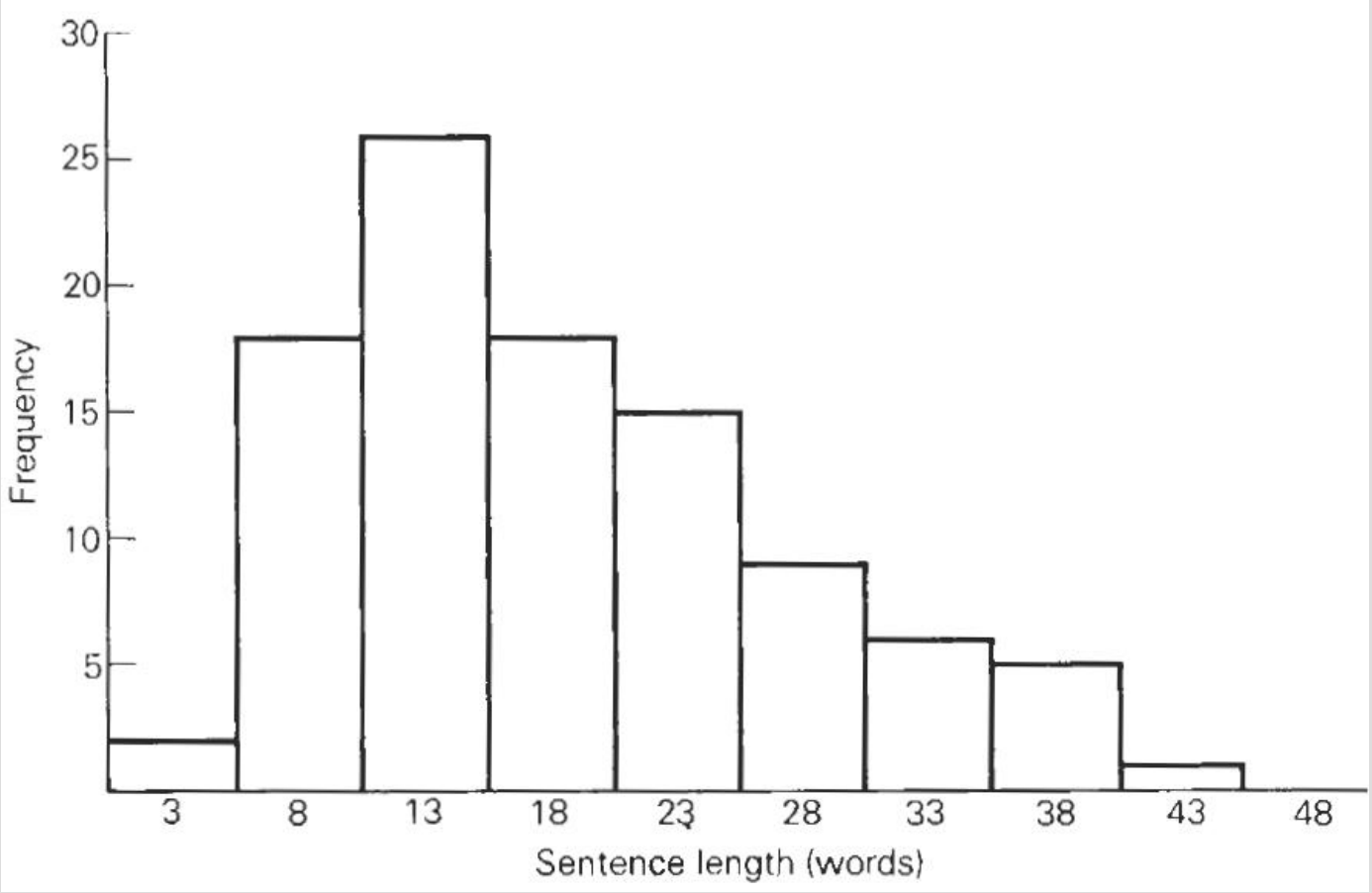
Los intervalos de son los límites a los extremos a los que llega una función. Se usan como un resumen respecto a la cantidad de datos, en el caso de que esto sean muy grandes. Los límites extremos de cada clase se les llaman **límite inferior** y **superior** de clase, respectivamente.

Para entenderlo de forma metafórica, operan igual que los intervalos en una partitura:



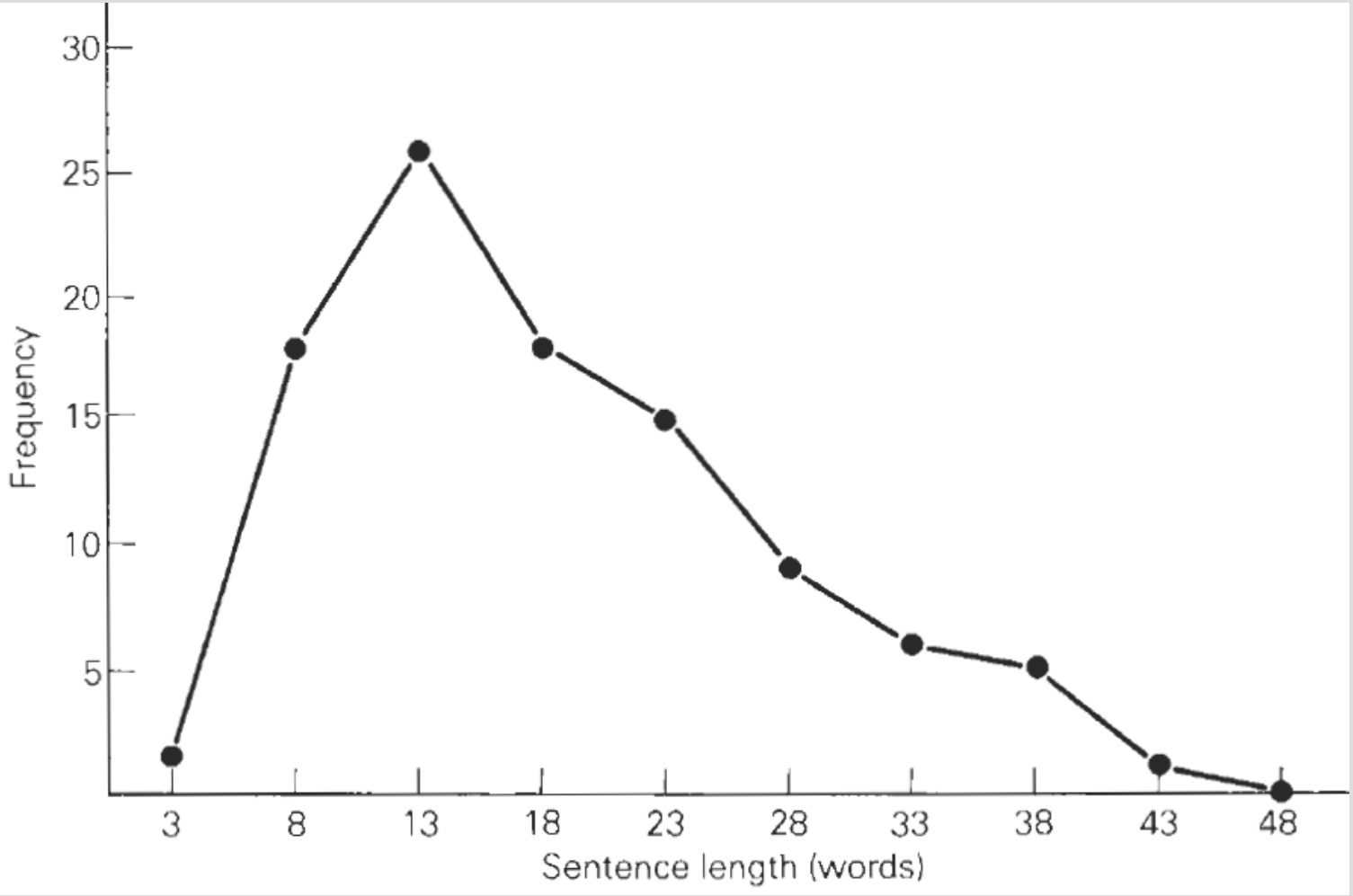
Contando palabras (5)

Una vez que hemos organizado nuestros datos en intervalos, entonces pasamos a representarlos en un histograma como el siguiente:



Contando palabras (6)

Si en vez de usar un histograma empleamos un polígono de frecuencias, entonces tenemos:



Ejercicio (1)

Vamos a hacer algo similar al conteo de palabras anterior, pero esta vez nos interesa detectar secuencias constitutivas de términos dentro de un corpus especializado. En concreto, queremos saber:

1. En informática, el término *sistema*, ¿con qué adjetivos aparece asociado de manera recurrente?
2. ¿Cuál de los dos términos es más frecuente: *base de datos* o *base de conocimientos*?
3. ¿Cuántos tipos de *bases de datos relacionales* hay?
4. El adjetivo *genético*, ¿con qué términos aparece asociado en informática?
5. Los términos *secuencia* y *cadena,* ¿son sinónimos?

Ejercicio (2)

Para resolver estas dudas, vamos a usar un corpus electrónico llamado **Bwananet**, desarrollado por la Universidad Pompeu Fabra (Barcelona, España):

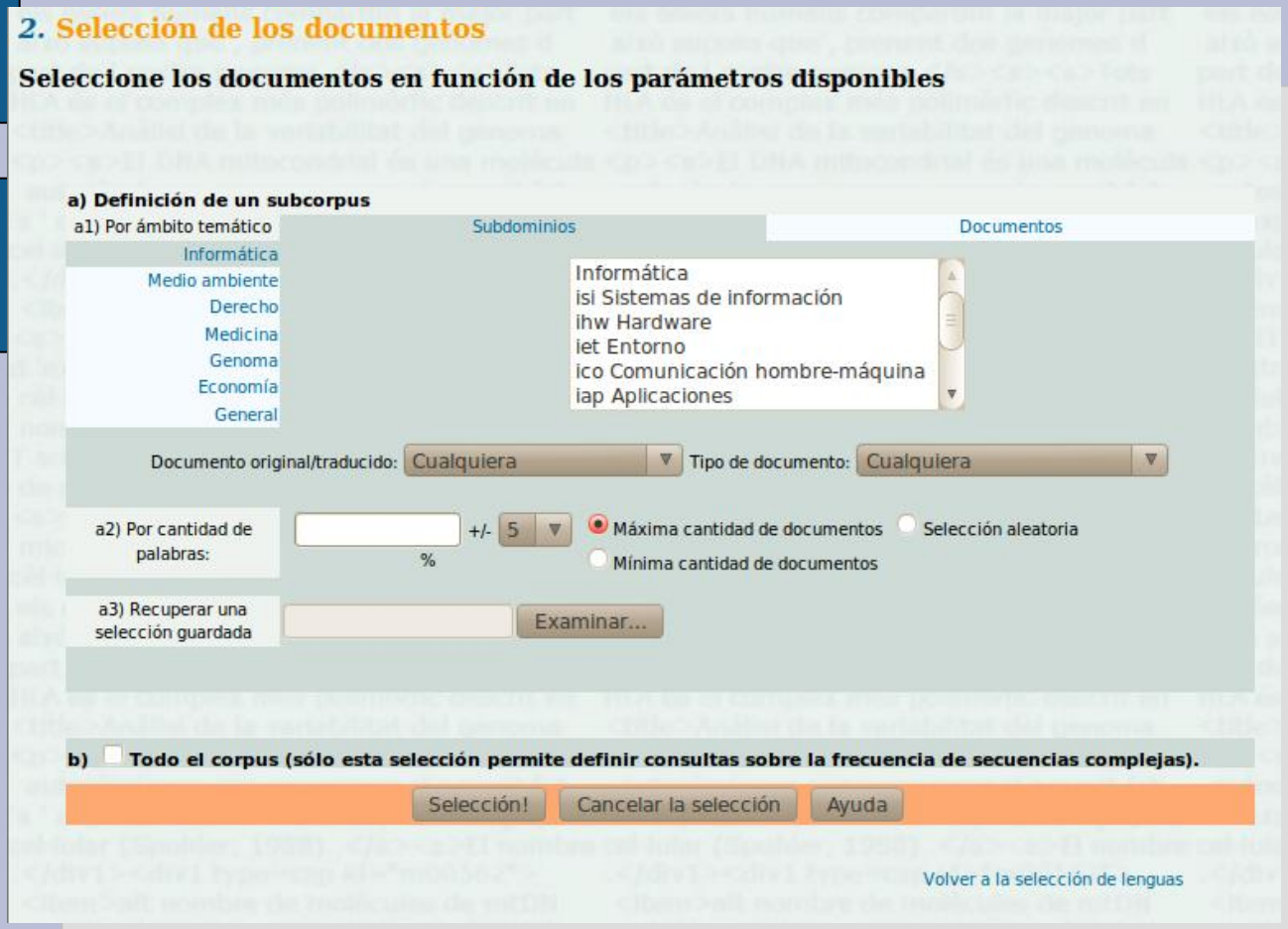
http://bwananet.iula.upf.edu/

Este corpus consiste en una colección de documentos especializados en varias áreas (informática, medio ambiente, derecho, medicina, ciencias genómicas, etc.), desarrollado por el Instituto Universitario de Lingüística Aplicada (IULA).

iula

UNIVERSITAT POMPEU FABRA

Ejercicio (3)

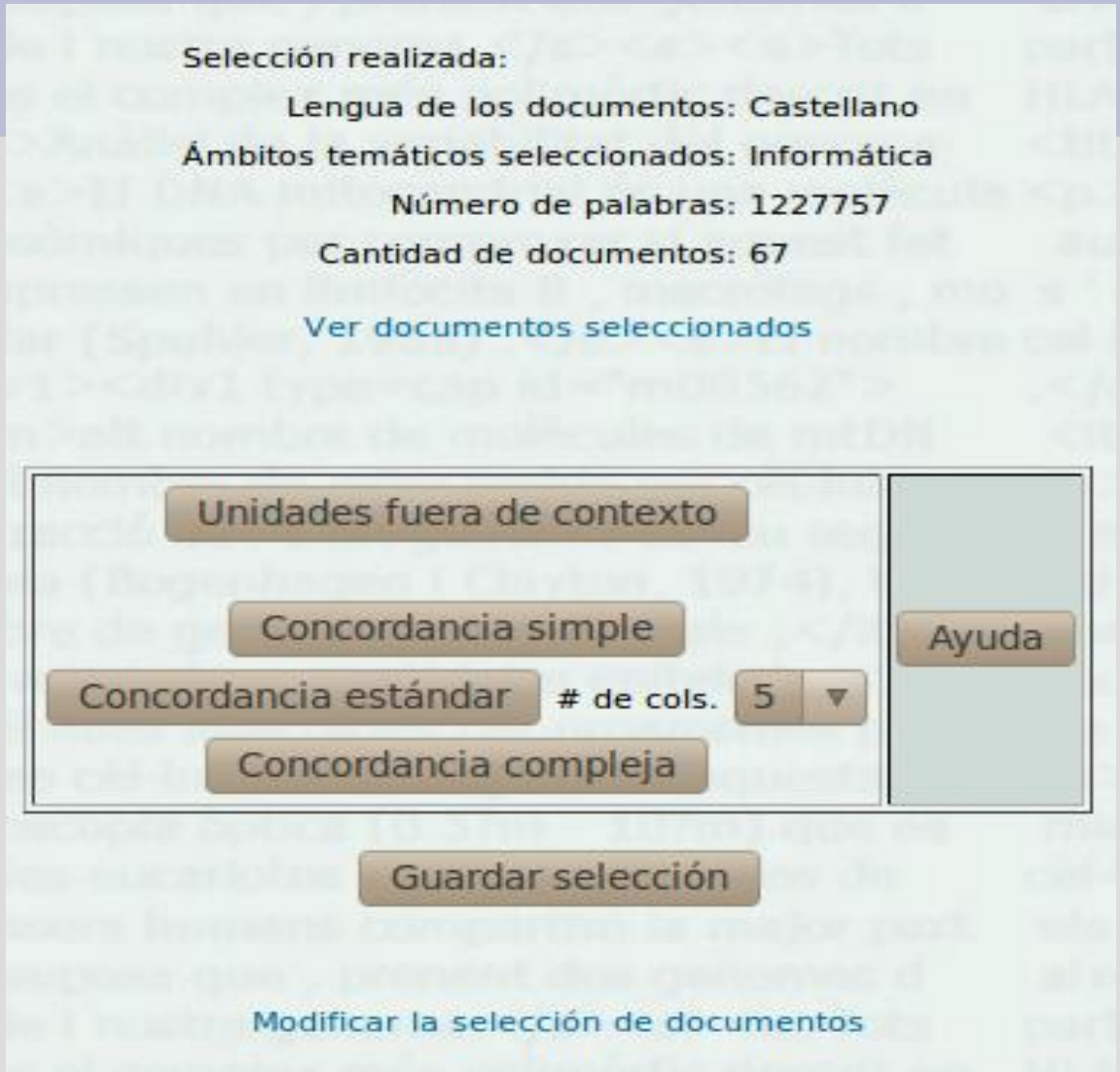


Bwananet cuenta con una ventana que nos describe primero cuáles son los temas que pueden ser buscados dentro de él, y en seguida nos permite establecer restricciones respecto a nuestra búsqueda. Eso nos ayuda refinar qué clase de información queremos obtener, ya que nos brinda un control sobre ésta.

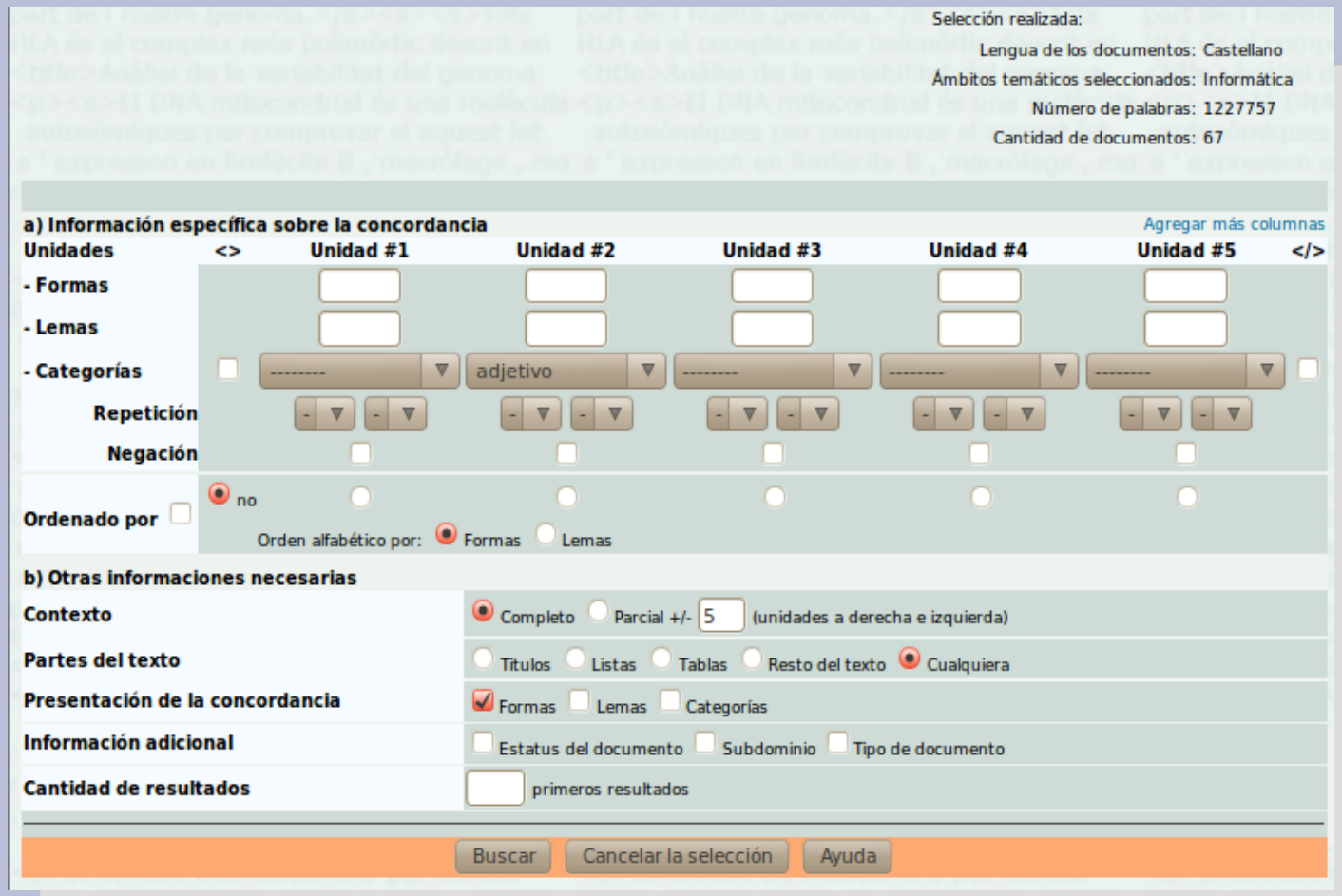
Ejercicio (4)

Posteriormente, podemos establecer parámetros para determinar qué cosas vamos a buscar.

Bwananet es un corpus que prioriza la búsqueda de unidades léxicas, dado que fue construido para resolver tareas en terminología (en concreto, identificación de términos en textos científicos y técnicos).



Ejercicio (5)



Las opciones que tenemos para hacer búsquedas son varias: en este caso, seleccionamos la búsqueda de concordancia estándar, esto es, consideramos secuencias formadas por máximo 5 palabras. ¿Qué podemos hacer con esto? Lo veremos a continuación.

Ejercicio (6)

1. En el caso particular de los términos **base de datos/base de conocimientos**, desarrollen las fórmulas de probabilidades condicionales que les corresponde a cada uno de los dos.

2. ¿Cuántas **bases de datos**/bases de conocimiento han localizado en Bwananet?

3. ¿Cuáles serían las variables dependientes/independientes que guiarían su análisis?

4. Respecto al rango de posibles combinaciones que asume el término **base**, ¿podemos fijar variables discretas o continuas?

5. Traten de desarrollar un histograma con los rangos de variación detectados entre las combinaciones **base de datos/base de conocimientos**.

Gracias por su atención

Blog del curso:
http://cesaraguilar.weebly.com/meacutetodos-y-teacutecnicas-de-investigacioacuten-cuantitativa.html