Técnicas estadísticas aplicadas al análisis histórico

J. L. Pereira Iglesias
M. Rodríguez Cancho
F. Sánchez Marroyo
Universidad de Extremadura

I.-Los datos como punto de partida.

Es habitual que en los trabajos actuales de investigación histórica aparezcan numerosas técnicas estadísticas incorporadas al análisis demográfico, económico y social. No obstante, el tratamiento estadístico que se da al material histórico es, en la mayoría de las ocasiones, elemental. El uso de los diferentes estadísticos y parámetros no significa que se consiga una integración adecuada en la explicación histórica global.

En esta comunicación nuestro interés se centra en mostrar un conjunto de técnicas, a partir de las cuales consideradas como instrumentos y recursos metodológicos, el historiador dispone de un aparato científico al margen de las propias cronologías tradicionales y de los estrictos límites temporales.

El material documental que ha servido de base para este trabajo es representativo de un período temporal amplio. Se trata, por una parte, de fuentes que tienen como objetivo fundamental el registro de los efectivos demográficos, bien en su propia finalidad, o bien indirectamente como medio de llegar a un conocimiento fiscal de la población del Antiguo Régimen. Por otra, se manejan diversos materiales de naturaleza tributaria y hacendística, sin olvidar repertorios sobre patrimonios privados[1].

Este acervo documental proporciona datos numéricos que conforman los diferentes valores que toman las variables demográficas y económicas.¿Cómo analizar estos datos numéricos? Para poder realizar esta operación es obligado por parte del historiador el acudir a las técnicas estadísticas. Estas nos proporcionan los recursos metodológicos necesarios para tratar el material susceptible de cuantificación y seriación, ya sea el correspondiente al total de una población o sólo una muestra de la misma, ya sean las relaciones de propiedad de las dehesas en un espacio concreto.

El paso inicial consiste en realizar una descripción estadística del material

[1] Las fuentes utilizadas son:

Tomás González. **Censo de población de las provincias y partidos de la Corona de Castilla en el siglo XVI.** Madrid. 1829.(Reedición facsímil del I.N.E., 1982).

B.R.A.H. **Censo de población de Floridablanca de 1787.** Sig. 9/6.202.

A.D.E. **Censo de la poblacion de España de 1887.** Madrid. 1889.

A.H.P.C. **Amillaramiento de Cáceres.** 1903, s/c.

A.H.P.C., Sección Real Audiencia, **Libro de Yerbas de Cáceres.** 1731. Leg. 563.

cuantificado, para lo cual se ha recurrido a la herramienta informática, incorporada definitivamente al trabajo histórico[2]. Los distintos valores de la(s) variable(s) son objeto de un recuento para detectar las frecuencias de los mismos. La **ordenación** de los datos debe hacerse conforme a determinados criterios según la propia naturaleza del trabajo que se desarrolla. El hecho de que las distintas variables que maneja el historiador ofrezcan numerosos valores similares o diferentes hace necesario medir las **frecuencias** o repeticiones de tales valores y, a partir de aquí, construir las **distribuciones de frecuencias simples o agrupadas.**

Las distribuciones de frecuencias de la segunda clase implican el agrupamiento de los datos en **clases o intervalos** limitados por dos valores de cierre del intervalo: los **límites inferior y superior** del intervalo. Esta operación estadística la realizan en muchas ocasiones los historiadores; el mero hecho de agrupar los datos de una serie cronológica en períodos quinquenales o decenales constituye ya un agrupamiento de los datos de la variable en intervalos temporales.

Pero, ¿Existe algún criterio estadístico para agrupar los datos o es algo aleatorio y arbitrario? La respuesta a esta interrogante está condicionada por la enorme importancia que tiene la **marca de clase** en los cálculos estadísticos ulteriores. No olvidemos que uno de los grandes problemas en el análisis estadístico del pasado radica en la determinación de los promedios. Los agrupamientos de datos conllevan un coste, la pérdida de información. Por tal razón, el historiador debe cuidar de elegir adecuadamente el número idóneo de clases y la amplitud de las mismas. Los manuales de estadística abogan por la utilización de varios procedimientos: **criterio de Norcliffe, de Huntsberger**, etc.[3] Para los historiadores poco familiarizados con los cálculos matemáticos el criterio más sencillo es el de Norcliffe, que consiste en fijar un número igual o aproximado de clases al valor de la raíz cuadrada de N (número de datos). En cuanto al tamaño de los intervalos es aconsejable dividir el **recorrido o rango** de los datos a agrupar por el número de clases previamente estimado. El resultado que se obtenga será la longitud de clase[4].

Conviene que los límites de los intervalos de clase sean números enteros y que tengamos presente los denominados **intervalos reales de clase** para poder clasificar aquellos datos situados a caballo entre un intervalo u otro. En igual sentido, conviene que los historiadores se acostumbren a emplear en sus cálculos las **frecuencias relativas** para evitar los resultados engañosos que a veces proporcionan los valores absolutos.

Esta descripción numérica puede expresarse mediante la siguiente tabla estadística:

---

[2] El paquete informático utilizado en este caso comprende el programa **Microsoft File**, banco de datos, el **Cricket Graph**, de gráficos, y el **Stat View**, de estadística.

[3] J.M.RASO, J. MARTIN VIDE, P. CLAVERO: **Estadística básica para Ciencias Sociales**. Barcelona, 1987, p. 25.

[4] El recorrido estadístico o rango es muy fácil de calcular. Basta con restar al valor más alto del conjunto de datos o de la serie el correspondiente valor más bajo. El resto será el rango buscado.

Distribución de frecuencias agrupadas.

1591

Intervalos	Frecuencias
73-646	103
646-1219	84
1219-1792	49
1792-2365	24
2365-2939	11
2939-3512	5
3512-4085	6
4085-4658	4
4658-5232	1
5232-5805	2
5805-6378	3
6378-6951	1
6951-7525	1
> 7525	1

1787

Intervalos	Frecuencias
44-693	145
693-1343	65
1343-1993	36
1993-2643	15
2643-3293	10
3293-3942	7
3942-4592	6
4592-5242	6
5242-5892	1
5892-6542	0
6542-7192	1
7192-7841	1
7841-8491	1
8491-9141	0
9141-9791	0
9791-10441	0
10441-11091	1

1887

INTERVALOS	FRECUENCIAS
77-1677	168
1677-3277	69
3277-4877	25
4877-6477	12
6477-8077	9
8077-9678	5

9678-11278	2
11278-12878	2
12878-14478	0
14478-16078	1
16078-17678	1
17678-19279	0
19279-20879	0
20879-22479	0
22479-24079	0
24079-25679	0
25679-27280	1

El paso siguiente es el procesamiento de estos datos[5] a fin de obtener los **parámetros de tendencia central y de dispersión** que constituyen un segundo nivel en el tratamiento estadístico de los mismos. La estimación de la(s) **moda(s)**, la **mediana**, la **media aritmética simple** o **ponderada**, la **media geométrica** y la **media armónica** completan las medidas de **tendencia central.** Llamamos la atención sobre el poco uso que de las medias geométrica y armónica han hecho los historiadores, desconociendo las ventajas de las mismas. La gran dispersión que suelen tener los datos cuantificados a partir de las fuentes del período preestadístico, aconseja que en la mayoría de las ocasiones se haya de corregir la influencia de los valores extremos sobre los correspondientes promedios. La media geométrica permite una corrección inicial en aquellos casos en los que los valores extremos de una serie o de un conjunto de datos agrupados desvían la medida de tendencia central hacia un límite por exceso. Las razones entre las diferentes medidas de tendencia central señalan que la media geométrica es un valor inferior a la media aritmética sencilla y superior por lo común a la media armónica.

Los **parámetros de dispersión** son sin duda alguna fundamentales en el tratamiento estadístico de la masa documental de datos históricos. Su medición permite aceptar o rechazar todos aquellos valores que se ajusten o no a las correcciones establecidas. El parámetro de dispersión más importante es la **desviación típica (standard).** Ahora bien, con el objeto de poder comparar series o matrices de datos de distinta naturaleza o magnitud se hace necesario proceder al cálculo del **coeficiente de variación** o, lo que es lo mismo, dividir la desviación típica por el promedio utilizado. Suele expresarse el coeficiente en porcentaje (%).

Los **coeficientes de sesgo** y **curtosis** también son muy adecuados para averiguar la "forma" que presentan los datos de una distribución de frecuencias. Pensemos en el caso de los censos fiscales y en la forma que ofrece el correspondiente **histograma** o **polígono de frecuencias.**

En las tablas que se recogen a continuación se ofrecen estos valores.

[5] Con objeto de disponer de datos completos para los tres recuentos y censos de población se han seleccionado aquellos núcleos en los que constan tales valores, es decir, 295 municipios extremeños. El total de la población estudiada para 1591 (resultado de utilizar el coeficiente 3,5) es de 376.053 habitantes; para 1787 , 357.957 habitantes y , en 1887, 684.206.

Promedios y coeficientes de dispersión

	1591	1787	1887
Media a.	1274,7	1213,4	2319,3
Media g.	867,4	730,1	1451,3
Media h.	589,7	437,2	919,6
Desviación	1301,2	1404,1	2809,0
Curtosis	10,0	11,6	23,9
Sesgo	2,7	2,8	3,8
C.V.	102,0	115,7	121,1

El análisis estadístico, al margen de la validez real de las cifras, refleja que en 1787 la población no había recuperado aún los niveles registrados por el censo de 1591. La media de habitantes por municipio, dato de un valor estadístico innegable, pero demográficamente relativo, es claramente superior en los datos correspondientes a esta última fecha. Sin embargo, con respecto al censo del siglo XIX, mucho más riguroso en su confección y elaboración, se han visto intensificados los promedios, signo evidente del crecimiento demográfico experimentado a lo largo de aquella centuria. Como se aprecia, la media aritmética simple resulta menos adecuada a la realidad demográfica de los casos estudiados, que el empleo de la media geométrica y armónica. La primera se ve más influida por los valores extremos, mientras que los otros promedios corrigen mejor la incidencia de las desviaciones.

Los coeficientes de desviación, muy altos, ponen de relieve la heterogeneidad de las cifras. Este fenómeno estadístico de dispersión progresiva, unos municipios cada vez mayores, otros, por el contrario, cada vez más despoblados, se justifica desde el punto de vista del análisis de la población por un proceso de concentración.

Un problema complementario, que la estadística puede ayudar a resolver, es el que con frecuencia se presenta a los historiadores que nos movemos en la etapa preestadística, determinar el grado de ocultación del material documental que manejamos. En este caso, en el que se utilizan fuentes de amplia cobertura temporal, surgen, razonablemente, dudas acerca de la validez estadística de los valores demográficos. Los datos de los recuentos de 1591 y 1787 conllevan un margen de error debido a una voluntad de ocultación que resulta difícil de cuantificar. Hacemos frente al problema a partir de dos vías.

Por una parte, intentamos comprobar el crecimiento de la población en períodos temporales sucesivos y en espacios diferentes. En cuanto al tiempo, se mantienen las fechas de 1591, 1787 y 1887 como puntos de referencia documental para nuestro estudio. Desde el marco espacial se establece el análisis estadístico e histórico a partir de la idea de **Partido** como unidad territorial y jurisdiccional, tal y como estaba vigente a finales del Antiguo Régimen[6]. La **tasa de crecimiento** es un instrumento que posibilita la constatación del rigor de las fuentes.

[6] Extremadura se concebía entonces estructurada en ocho partidos: Alcántara, Badajoz, Cáceres, Llerena, Mérida, Plasencia, La Serena y Trujillo.

Mediante este proceso se busca detectar unos niveles de homogeneidad en el comportamiento que dé credibilidad al material documental[7].

Tasas de crecimiento medio anual en Extremadura (%). (1591-1887)

PARTIDO	1591-1787	1787-1887	1591-1887
Alcántara	-0,02	0,46	0,14
Badajoz	-0,01	0,65	0,20
Cáceres	0,06	0,52	0,21
Llerena	-0,06	0,70	0,19
Mérida	0,06	0,81	0,32
Plasencia	-0,08	0,49	0,11
La Serena	0,10	0,79	0,33
Trujillo	-0,09	0,71	0,17

De la observación de las cifras presentadas no se aprecia una disparidad en los valores, sino que resulta una coherencia interna en la distribución de las tasas. Es un indicador, bien de que los niveles de la posible ocultación informativa mantienen un comportamiento homogéneo, bien de que los datos registran un grado de veracidad similar.

Por otra, se trata de medir la **relación** existente entre las cifras de población de cada núcleo, en los tres recuentos estudiados. Con ello pretendemos evaluar el grado de concordancia de los valores demográficos, en los distintos núcleos, tanto en la variable temporal como en la espacial.

Coeficientes de Correlación de Pearson

Período	Coeficiente
1591-1787	0,88
1591-1887	0,81
1787-1887	0,92

El valor de los **coeficientes de correlación** es muy alto, lo que muestra la intensa relación que existe, a lo largo del período estudiado, en la evolución demográfica de los diferentes núcleos. Se confirma lo que ya las tasas de crecimiento habían señalado.

[7] En este sentido, el profesor Eiras Roel ya aplicó semejantes criterios estadísticos a recuentos y censos de población de Galicia. "Test de concordancia aplicado a la crítica de vecindarios fiscales de la época preestadística" en **Actas I Jornadas de Metodología Aplicada de las Ciencias Históricas**, vol.II. Santiago. 1975.

II.- La representación gráfica.

La representación gráfica de las distribuciones de frecuencias se realiza generalmente mediante los denominados **histogramas**, siendo muy útil la construcción de los **polígonos de frecuencias**. La dispersión de los datos también tiene su representación gráfica: **diagramas de dispersión.** A partir de ellos podríamos observar como determinados núcleos -por exceso o por defecto- quedan fuera de los intervalos de confianza típicos de la **distribución normal.** Cuando la dispersión de los valores es elevada sólo la utilización de unos límites de confianza flexibles permitirá englobar una muestra significativa de pueblos.

Por último, nuestro análisis descriptivo finaliza con las **medidas de concentración**, también llamadas de desigualdad, **Indice de Gini** y **Curva de Lorenz.** Dadas sus características, utilizamos para su presentación los informes referentes al reparto de los medios de producción, en este caso, el factor tierra. Se toman dos instantes temporales, uno, 1731, en pleno Antiguo Régimen, y otro, 1903, en la Restauración, cuando la sociedad agraria tradicional alcanzó su cénit. Estos indicadores miden, aun cuando los datos conllevan una notable ocultación de la riqueza, el grado de concentración de la propiedad de las dehesas o, dicho de otra forma, el nivel de desigualdad en la distribución del suelo. La presentación de los datos en una tabla estadística es el primer paso para el cálculo del Indice y de la curva.

DISTRIBUCION DE FRECUENCIAS ACUMULADAS
1731

Intervalos	Frecuencia	F.r.a.(p)	F.r.a.(q)
1-10	14	5,0	0,03
11-25	9	8,3	0,11
26-50	32	19,8	0,67
51-100	48	37,0	2,36
101-200	46	53,6	5,57
201-400	45	69,8	11,84
401-650	24	78,4	17,69
651-1000	12	82,7	22,29
1001-1300	5	84,5	24,96
1301-1750	9	87,8	31,33
1751-2500	8	90,6	39,22
2501-5000	16	96,4	67,06
5001-7500	8	99,3	90,26
7501-13500	2	100,0	100,0

DISTRIBUCION DE FRECUENCIAS ACUMULADAS
1903

Intervalos	Frecuencias	F.r.a.(p)	F.r.a.(q)
1-5	117	21,9	0,13
6-15	93	39,4	0,50
16-30	49	48,6	0,92
31-50	29	54,0	1,36
51-75	36	60,8	2,21
76-100	16	63,8	2,74
101-150	32	69,8	4,24
151-200	24	74,3	5,82
201-300	22	78,4	7,89
301-500	34	84,8	13,00
501-750	19	88,4	17,44
751-1000	12	90,6	21,38
1001-1500	13	93,0	27,47
1501-2000	10	94,9	34,03
2001-3000	5	95,9	38,72
3001-4000	7	97,2	47,90
4001-6000	10	99,1	66,64
6001-8000	2	99,4	71,89
15001-25000	2	99,8	86,88
25001-45000	1	100,0	100,00

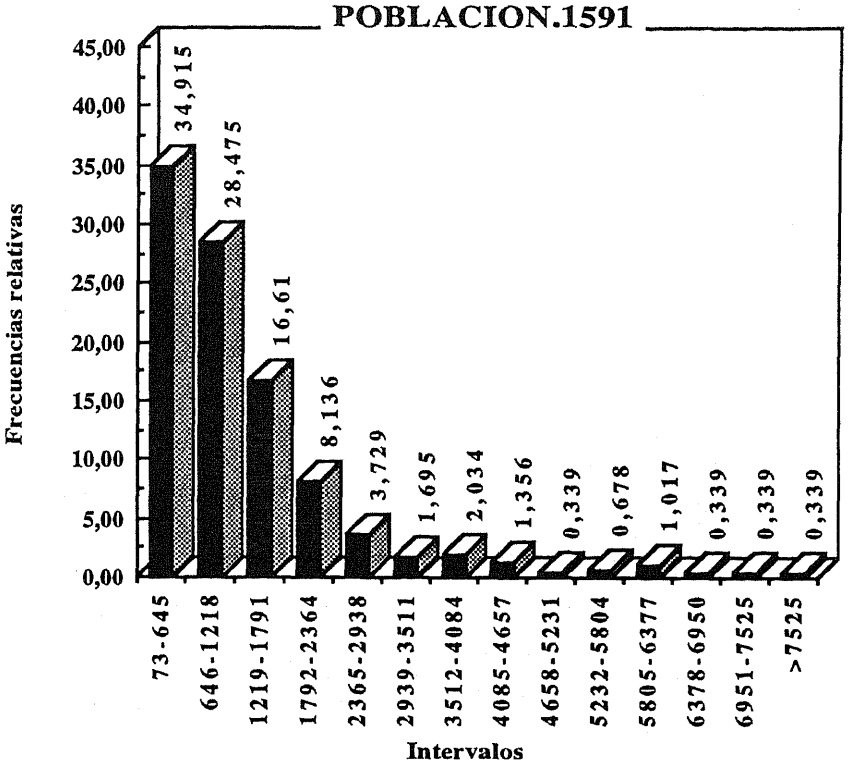
En 1731 el valor del **Indice de Gini** se sitúa en **0,61**. Indicador relativamente alto que refleja un cierto grado de concentración de la propiedad adehasada en la **Tierra de Cáceres**. En 1903 se ha intensificado este fenómeno, alcanzando el Indice un **0,69**. Las **curvas de Lorenz** (Cfr. Apéndice), complemento gráfico, reflejan con precisión este comportamiento.

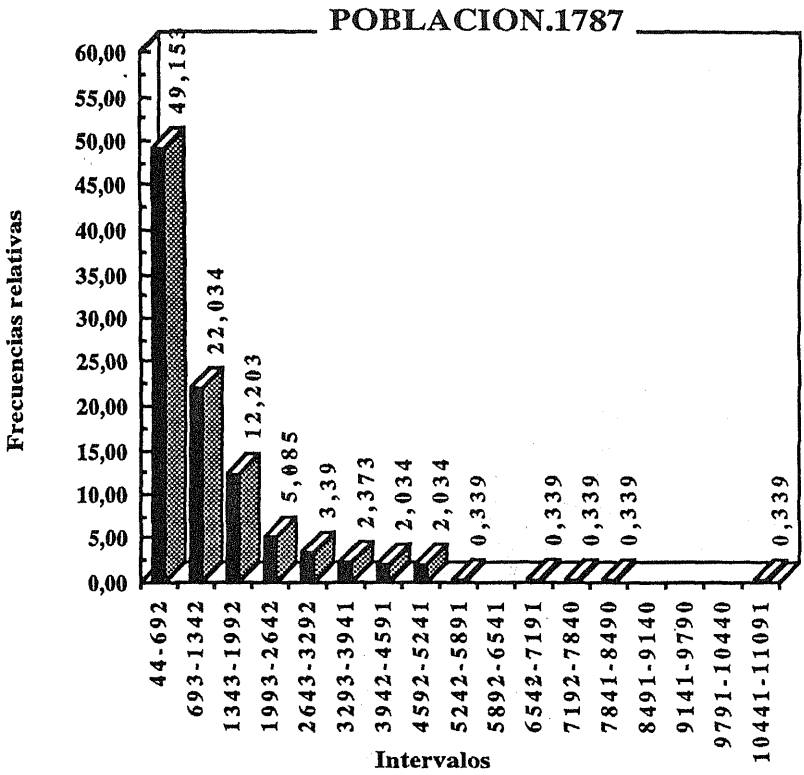
Ahora bien, es necesario tener en cuenta un hecho que exige por su significado una interpretación estadística e histórica, si no se quieren obtener conclusiones equívocas, incompatibles con el rigor del análisis científico. En efecto, entre 1731 y 1903, los índices señalan que la concentración de la riqueza rústica se ha intensificado, pero, al mismo tiempo, la evidencia histórica muestra que el número de propieterios aumentó de forma considerable. Estaríamos en presencia entonces de un fenómeno de dispersión de la propiedad; sin embargo, conviene apuntar que la simultaneidad de ambos conceptos no es incompatible. Se ha producido una dispersión por el incremento del número de pequeños propietarios, que accedieron a tal condición a partir de las transformaciones agrarias del Siglo XIX. Pero, al mismo tiempo, de forma paralela los grandes poseedores siguen acumulando la parte más substancial del terrazgo.

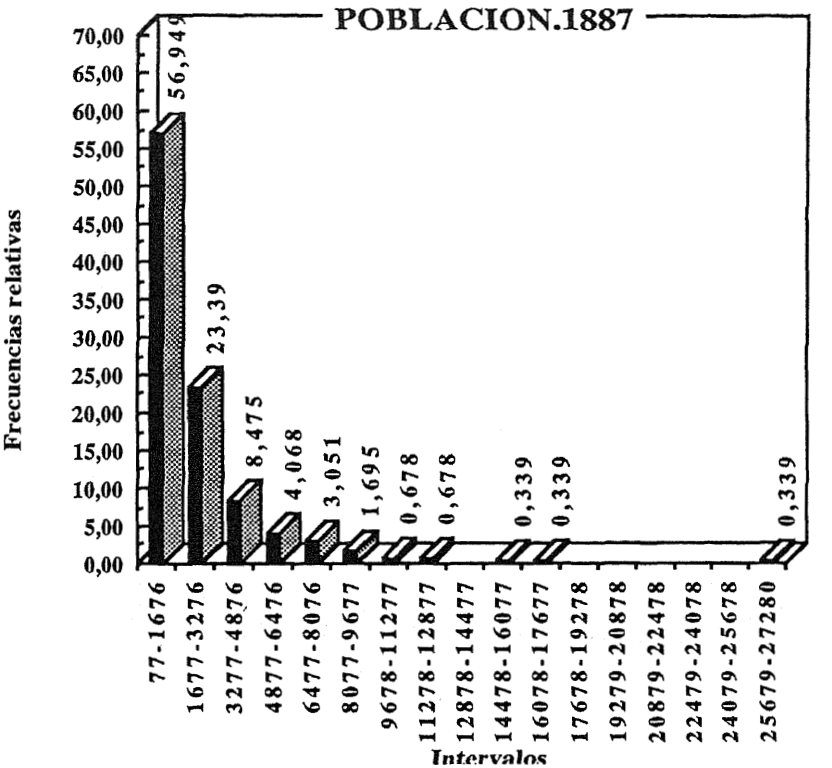
Es en definitiva un ejemplo que muestra la necesidad de interrelacionar el análisis estadístico con el trabajo empírico del historiador.

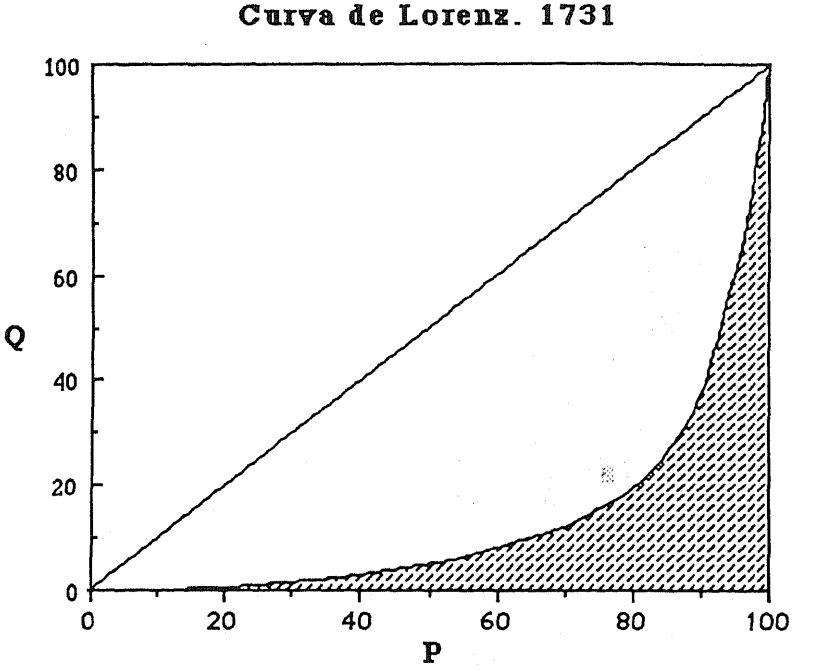
* * *

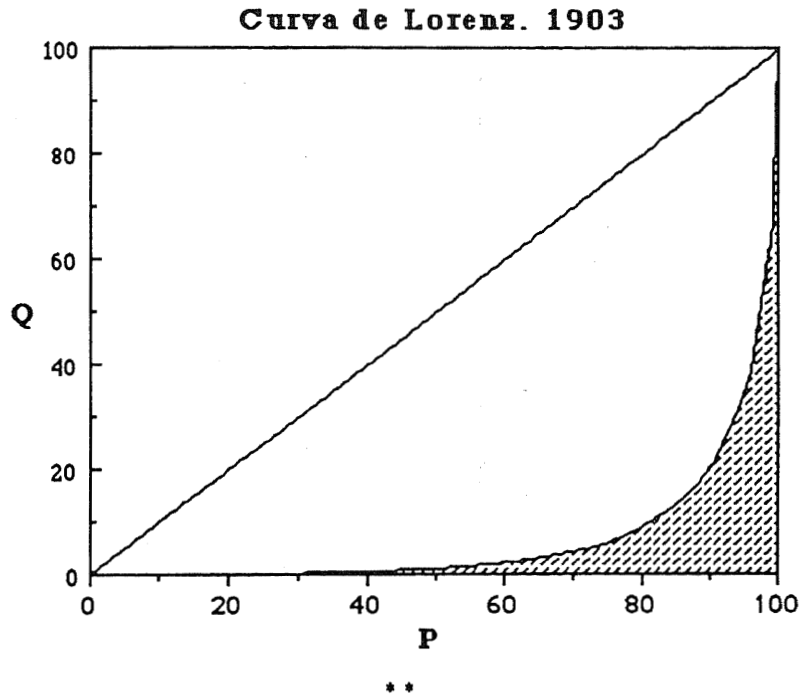
APENDICE GRAFICO




POBLACION.1787
Frecuencias relativas
60,00
55,00
50,00
45,00
40,00
35,00
30,00
25,00
20,00
15,00
10,00
5,00
0,00
49,153
22,034
12,203
5,085
3,39
2,373
2,034
2,034
0,339
0,339
0,339
0,339
44-692
693-1342
1343-1992
1993-2642
2643-3292
3293-3941
3942-4591
4592-5241
5242-5891
5892-6541
6542-7191
7192-7840
7841-8490
8491-9140
9141-9790
9791-10440
10441-11091
Intervalos


POBLACION.1887
Frecuencias relativas
70,00
65,00
60,00
55,00
50,00
45,00
40,00
35,00
30,00
25,00
20,00
15,00
10,00
5,00
0,00
56,949
23,39
8,475
4,068
3,051
1,695
0,678
0,678
0,339
0,339
0,339
77-1676
1677-3276
3277-4876
4877-6476
6477-8076
8077-9677
9678-11277
11278-12877
12878-14477
14478-16077
16078-17677
17678-19278
19279-20878
20879-22478
22479-24078
24079-25678
25679-27280
Intervalos


Curva de Lorenz. 1731
100
80
60
40
20
0
Q
0
20
40
60
80
100
P


Curva de Lorenz. 1903
100
80
60
40
20
0
Q
0
20
40
60
80
100
P

* *