Introducción a la Estadística

Sesión 1

Historia
Utilidad y aplicaciones
Definiciones y términos
Tipos de variables
Niveles de medición y requisitos

Dr. Carlos J. Vilalta (DAP)
Estadística - Curso propedéutico MAPP 2016
Centro de Investigación y Docencia económicas (CIDE)
Contacto: carlos.vilalta@cide.edu

¿Qué es la Estadística?

- Fisher (1956):

"The purpose of statistics is to develop and apply methodology for extracting useful knowledge from both experiments and data. In addition to its fundamental role in data analysis, statistical reasoning is also extremely useful in data collection (design of experiments and surveys) and also in guiding proper scientific inference"

- Rama de las Matemáticas que transforma los datos en información útil para la toma de decisiones

- Su estudio se divide tradicionalmente en 2 áreas:
  - **Descriptiva**: pasar de los datos a una descripción sintética o a la información
  - **Inferencial**: muestreo, pruebas de hipótesis y toma de decisiones sobre la base de información muestral

Utilidad en Políticas Públicas

- La Estadística es indispensable en la toma de decisiones científica o basada en evidencias:
  - Porque describe ordenada y metódicamente conjuntos de datos
  - No hay otra herramienta de análisis que lo haga con la misma precisión
    - Precisión: exactitud o fidelidad de un dato
  - Ayuda en la realización de pruebas empíricas de hipótesis o experimentos
    - Especulación/opiniones vs. hechos
    - Las pruebas son directas, los datos observables y mensurables y las hipótesis falseables
  - Es la mejor herramienta para inferir de una muestra a una población
    - Rara vez se hacen análisis sobre información de poblaciones completas
  - El análisis de PP requiere probar y conocer relaciones causales a efectos de proponer y evaluar diferentes cursos de acción
    - Nos permite asociar variables (análisis correlacional) y comparar grupos (análisis de diferencias)
  - La evidencia estadística y probabilística legitima la toma decisiones
  - Pero… en ocasiones la estadística se usa de forma equivocada, intencional o no intencionalmente

Historia

- 3,000 A.C.: tabletas con registros agrícolas y comercio en Sumeria.
- Siglo XVIII: desarrollo de la probabilidad en juegos de azar
- Siglo XIX: nació en como Ciencia Social vs. Matemática
- Etimología: Statista (*italiano*) = “status” (Estado) y sufijo “-ica” (relativo a)
- Comte: “Ciencia del Estado”
- Quételet: “Ciencia experimental de la Legislación”

Quételet (1846): sobre la utilidad de la distribución normal

MESURES de la POITRINE.	NOMBRE d'hommes.	NOMBRE PROPORTIONNEL.	PROBABILITÉ d'après L'OBSERVATION.	RANG dans LA TABLE.	RANG d'après le CALCUL.	PROBABILITÉ d'après LA TABLE.	NOMBRE d'OBSERVATIONS calculé.
Pouces.							
35	3	5	0,5000			0,5000	7
34	18	31	0,4995	52	50	0,4993	29
33	81	141	0,4964	42,5	42,5	0,4964	110
36	185	322	0,4823	33,5	34,5	0,4854	525
37	420	732	0,4501	26,0	26,5	0,4531	732
38	740	1303	0,3769	18,0	18,5	0,3799	1333
39	1073	1867	0,2464	10,5	10,5	0,2466	1858
			0,0597	2,5	2,5	0,0628	
40	1079	1882	0,1285	5,5	5,5	0,1359	1987
41	934	1628	0,2913	15	13,5	0,3034	1675
42	658	1148	0,4061	21	21,5	0,4130	1096
43	370	645	0,4706	30	29,5	0,4690	560
44	92	160	0,4866	35	37,5	0,4911	221
45	50	87	0,4953	41	45,5	0,4980	69
46	21	38	0,4991	49,5	53,5	0,4996	16
47	4	7	0,4998	56	61,8	0,4999	3
48	1	2	0,5000			0,5000	1
	5738	1,0000					1,0000

Figure 5.3. Quetelet's analysis fitting a normal distribution to data on the chest circumferences of Scottish soldiers. Column 1 gives the chest circumference in inches; columns 2 and 3 give the frequency and relative frequency distributions for 5738 individuals, columns 4–7 give the details of Quetelet's calculations (see text), and column 8 gives the fitted relative frequency distribution. (From Quetelet, 1846, p. 400.)

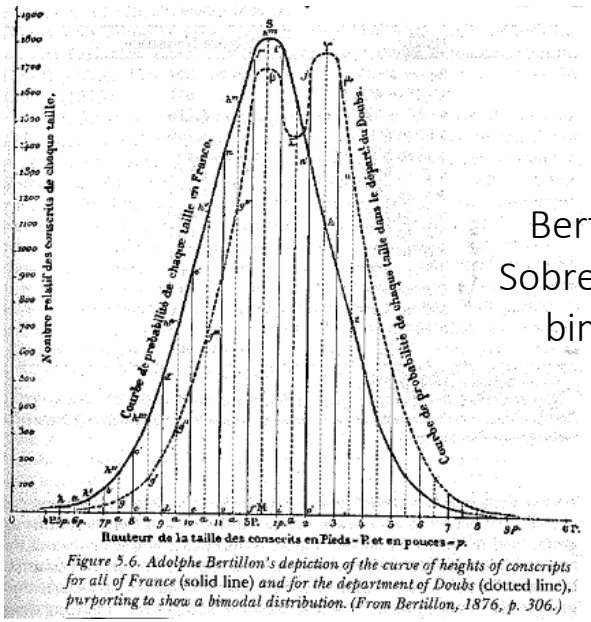


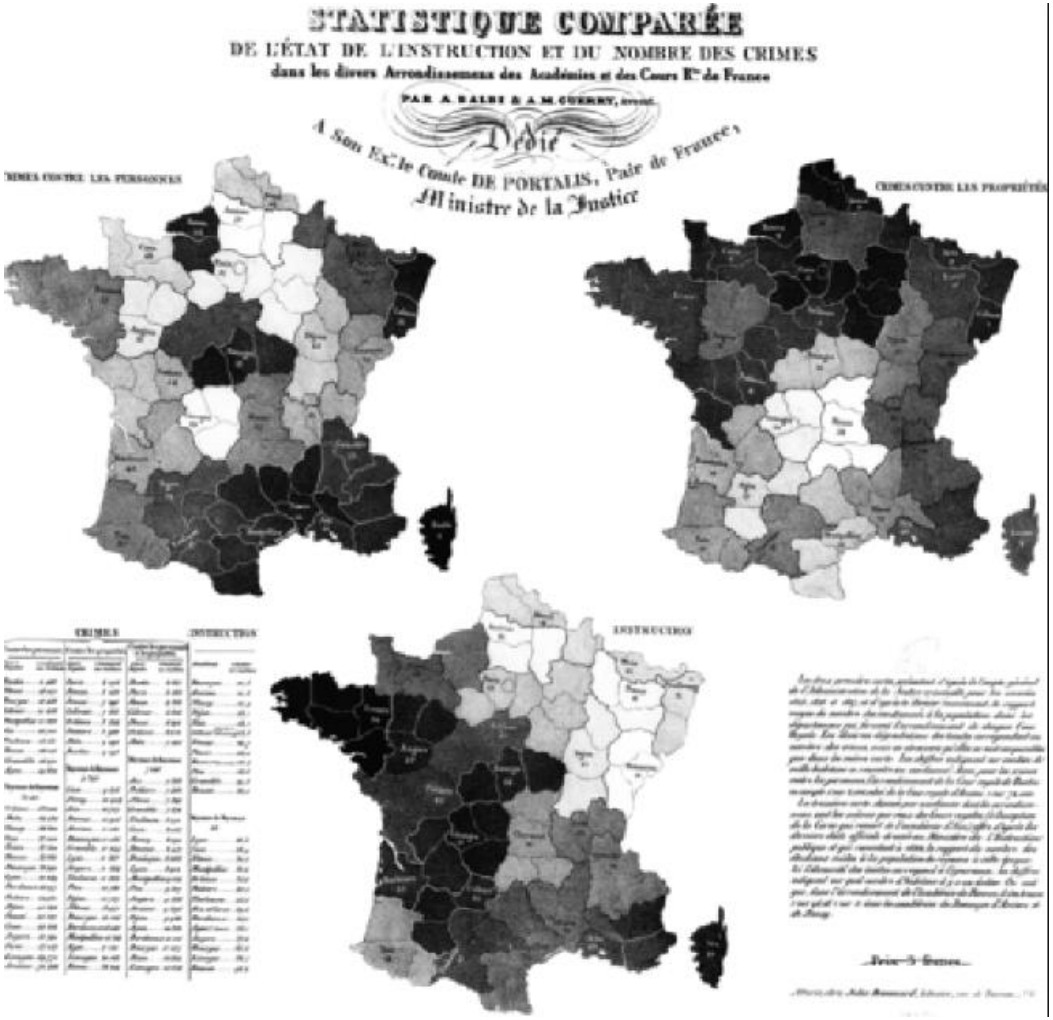
Figure 5.6. Adolphe Bertillon's depiction of the curve of heights of conscripts for all of France (solid line) and for the department of Doubs (dotted line), purporting to show a bimodal distribution. (From Bertillon, 1876, p. 306.)

Bertillón (1876): Sobre la distribución bimodal de las estaturas

Historia

- Se desarrolló en conjunto con la Sociología
- Rol de la Astronomía:
  - Ley de los Errores de Medición generó la Distribución Normal
- Galton: primer coeficiente de correlación (1888)
- Pearson: coeficiente r (1902) y prueba Chi-cuadrado (1903)
- Gosset (Student): prueba t (1908)
- Fisher: concepto de varianza y Anova (1918)
- Etc.

Correlación inversa entre Educación y Crimen
(Balbi y Guerry, 1829)



Historia en México

- Estadística: siempre persiguió fines utilitarios y económicos
- México prehispánico:
  - “Matrícula de tributos”. Listado de provincias, pueblos, cantidad y calidad de los productos
- México colonial:
  - Mediados siglo XVI: “Suma de Visitas de Pueblos”. Catastro de propiedades indígenas, nómina de tributos y padrón de habitantes de 907 jurisdicciones políticas
  - 1579: “Relaciones Geográficas” de Felipe II. Estadísticas de Nueva Galicia, Nueva Vizcaya y Nuevo León. Siguientes versiones en 1777 y 1791
  - Mediados siglo XVII:“Asuntos de Conventos y Colegios y Hospital Real”. Padrón y mediciones de consumo y provisión de alimentos
  - 1790: “Censo de Revillagigedo”. Población, edad, edo. civil, casta etc. en cuadros
  - 1804: “Tablas Geográfico Políticas del Reino de la N.E.” De Humboldt para su “Ensayo Político sobre el Reino de la Nueva España”
  - 1822: Oficina de Estadística General del Imperio. Iturbide
  - 1831: “Primer Censo General de la República”
  - 1833: Creación del “Instituto Nacional de Geografía y Estadística”
  - 1895: “Primer Censo General de Población”
  - 1900: “Segundo Censo General de Población” y desde entonces decenalmente

Definiciones y términos

- **Dato**: cualquier medición cuantitativa o cualitativa
  - Dato cuantitativo:
    - Puede ser sujeto a todas las operaciones aritméticas (ej. ingreso, detenciones, delitos, juicios, etc.)
  - Dato cualitativo:
    - No se sujeta a todas las operaciones aritméticas (ej. nombre, idioma, país, lugar de residencia, etc.)
- **Elemento**: las entidades u **observaciones** sobre las que se compilan los datos
- **Variable**: característica de interés de los elementos

¿Cuántas variables hay en este cuadro?
¿Y elementos?
¿Y datos?

TABLE 1.1 DATA SET FOR 25 MUTUAL FUNDS

Fund Name	Fund Type	Net Asset Value ($)	5-Year Average Return (%)	Expense Ratio (%)	Morningstar Rank
American Century Intl. Disc	IE	14.37	30.53	1.41	3-Star
American Century Tax-Free Bond	FI	10.73	3.34	0.49	4-Star
American Century Ultra	DE	24.94	10.88	0.99	3-Star
Artisan Small Cap	DE	16.92	15.67	1.18	3-Star
Brown Cap Small	DE	35.73	15.85	1.20	4-Star
DFA U.S. Micro Cap	DE	13.47	17.23	0.53	3-Star
Fidelity Contrafund	DE	73.11	17.99	0.89	5-Star
Fidelity Overseas	IE	48.39	23.46	0.90	4-Star
Fidelity Sel Electronics	DE	45.60	13.50	0.89	3-Star
Fidelity Sh-Term Bond	FI	8.60	2.76	0.45	3-Star
Gabelli Asset AAA	DE	49.81	16.70	1.36	4-Star
Kalmar Gr Val Sm Cp	DE	15.30	15.31	1.32	3-Star
Marsico 21st Century	DE	17.44	15.16	1.31	5-Star
Mathews Pacific Tiger	IE	27.86	32.70	1.16	3-Star
Oakmark I	DE	40.37	9.51	1.05	2-Star
PIMCO Emerg Mkts Bd D	FI	10.68	13.57	1.25	3-Star
RS Value A	DE	26.27	23.68	1.36	4-Star
T. Rowe Price Latin Am.	IE	53.89	51.10	1.24	4-Star
T. Rowe Price Mid Val	DE	22.46	16.91	0.80	4-Star
Thornburg Value A	DE	37.53	15.46	1.27	4-Star
USAA Income	FI	12.10	4.31	0.62	3-Star
Vanguard Equity-Inc	DE	24.42	13.41	0.29	4-Star
Vanguard Sht-Tm TE	FI	15.68	2.37	0.16	3-Star
Vanguard Sm Cp Idx	DE	32.58	17.01	0.23	3-Star
Wasatch Sm Cp Growth	DE	35.41	13.98	1.19	4-Star

Source: Morningstar Funds500 (2008).

Definiciones y términos

- Organización de los datos:
  - Pueden ser de 3 tipos: **Transversales**, **series de tiempo** y **panel** (longitudinal)

Ubicación	x
AGS	74
BC	23
BCS	15
CAMPECHE	52
COAHUILA	49
CHIHUAHUA	59

Transversal

$$y_i = \alpha + \beta x_i + u_i$$

Año	x
2010	58
2011	42
2012	38
2013	100
2014	56
2015	54

Serie de tiempo

$$y_t = \alpha + \beta x_t + u_t$$

Panel

$$y_{i,t} = \alpha + \beta x_{i,t} + u_{i,t}$$

Ubicación	Año	x
AGS	2010	100
AGS	2011	73
AGS	2012	84
AGS	2013	43
AGS	2014	22
AGS	2015	71
BC	2010	78
BC	2011	64
BC	2012	43
BC	2013	95
BC	2014	78
BC	2015	54
BCS	2010	90
BCS	2011	63
BCS	2012	92
BCS	2013	10
BCS	2014	77
BCS	2015	93

Definiciones y términos

- **Población**:
  - Todas las observaciones o individuos sobre los que se busca ofrecer una conclusión (N)
    - Poblaciones finitas e infinitas
- **Muestra**:
  - Grupo representativo de la población seleccionada para realizar el análisis estadístico (n)
- **Parámetro**:
  - Medición cuantitativa que describe o sintetiza una característica de la población (μ)
- **Estadístico**:
  - Medición cuantitativa que describe o sintetiza una característica de la muestra (M)

Definiciones y términos

- Ejemplo: ¿cuál es la sentencia promedio de los presos sentenciados en el Penal de Barrientos (Tlalnepantla, Edomex)?

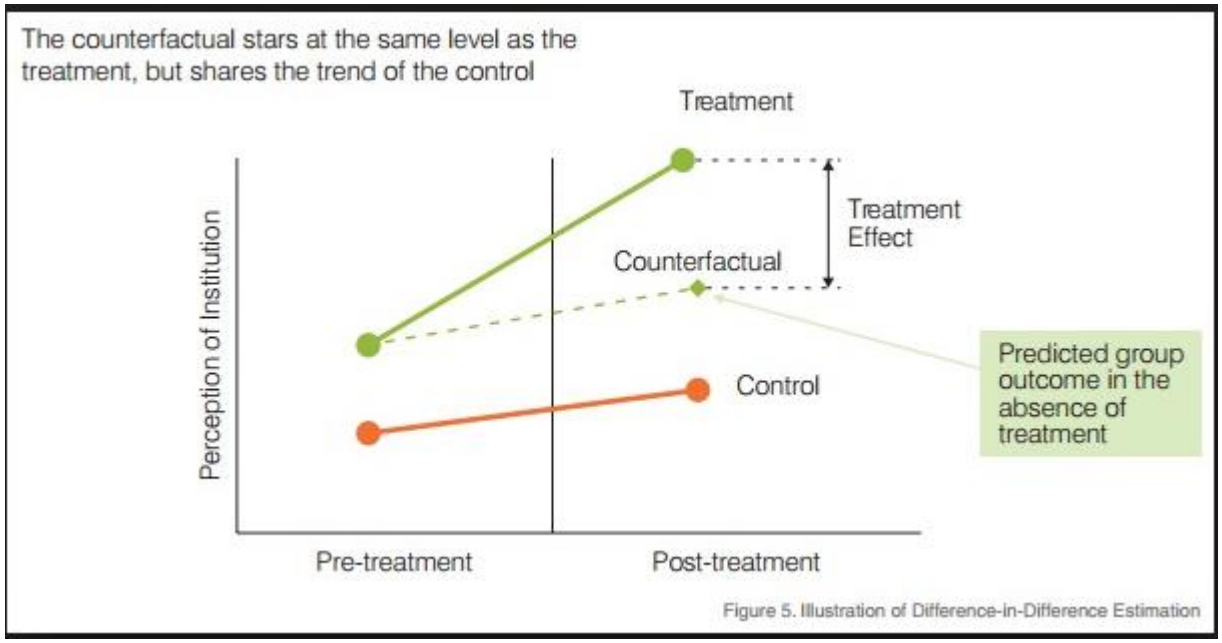
  - **Población**: todos los reclusos en ese penal
  - **Muestra**: una parte de los reclusos que representan al conjunto completo o población
  - **Variable**: la duración de las sentencias
  - **Elemento** u **observación**: un recluso
  - **Dato**: la duración de la sentencia de 1 recluso
  - Los **datos**: el conjunto de valores en la muestra de reclusos
  - El **parámetro**: la media (m) de duración de las sentencias de todos los reclusos
  - El **estadístico**: la media (M) muestra de la duración de las sentencias

Definiciones y términos

- Otros conceptos estadísticos frecuentemente utilizados en PP:

  - **Sesgo:** cuando las observaciones que se realizan de una población favorecen a unos individuos sobre otros
    - Muestra sesgada = muestra no representativa

  - **Contrafactual**: lo que habría sucedido de no haber intervenido



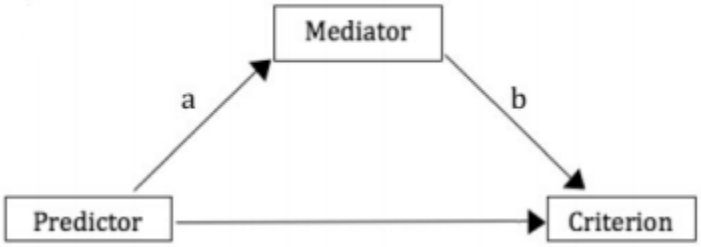
Fuente: UNODC

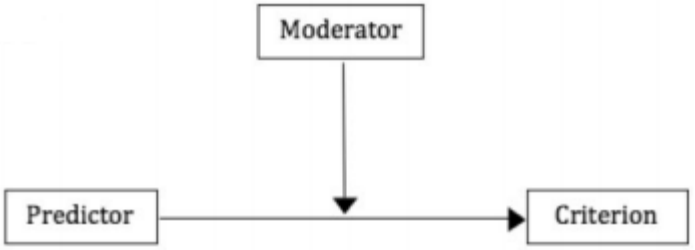
Tipos de variables (conceptualmente)

- Variable **aleatoria**: variable que toma diferentes valores como consecuencia de un muestreo o experimento aleatorio
- Variable **dependiente**: el efecto bajo estudio (y)
- Variable **independiente**: la causa del efecto bajo estudio (x)

$$y_i = \alpha + \beta x_i + u_i$$

- Variable de **control**: tercera variable neutralizada en la ecuación
- Variable **mediadora:** tercera variable que condiciona la asociación entre "x" y "y".
  - Utilidad: nos dice por qué "x" causa "y"
  - "x" afecta a "y" sólo cuando "z" está en la ecuación
- Variable **moderadora**: tercera variable que modula la relación entre "x" y "y"
  - Utilidad: nos dice en qué condiciones "x" causa "y"
  - "x" afecta a "y" como función de "z"





Tipos de variables (matemáticamente)

- Variables **discretas**: consisten de categorías o valores indivisibles
  - Sensación de inseguridad: 0 = No y 1 = Si
  - Sensación de inseguridad: 0 = Nada, 1 = Poco, 2= Algo, 3 = Mucho
  - A veces conocidas como variables categóricas
  - Nivel de medición: nominal y ordinal
- Variables **continuas**: consisten de valores divisibles
  - Tasa general de sentenciados condenatoriamente: 0 a ∞
  - Nivel de medición: intervalo y razón o proporciones

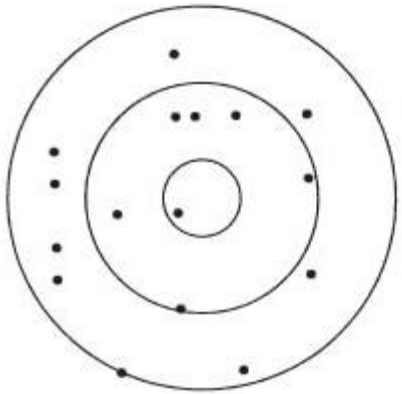
Nominal (discreta)	**Ordinal** (discreta)	**Intervalo** (continua)	**Razón** (continua)
Categorías no ordenadas Ej. ideología	Categorías ordenadas en intervalos no necesariamente equivalentes Ej. ingreso en SMV	Valores ordenados en intervalos equivalentes y sin cero absoluto Ej. temperatura	Valores ordenados en intervalos equivalentes con un cero absoluto Ej. ingreso en $

Requisitos de las mediciones

- **Validez**: que lo que medimos sea realmente lo que queremos medir
  - Pregunta: ¿Estás midiendo el concepto que quieres medir?

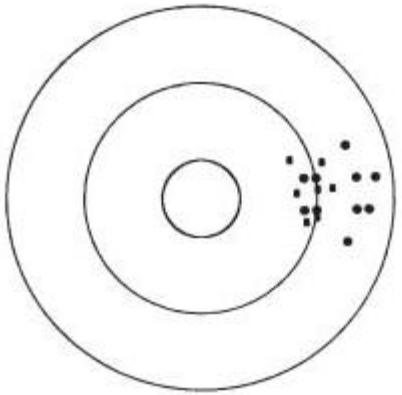
- **Confiabilidad**: independientemente del concepto… ¿si mides lo mismo varias veces, obtienes los mismos valores cada vez? ¿Lo estoy midiendo bien?
  - Una medida es confiable si la respuesta es afirmativa

Validez y confiabilidad de las mediciones



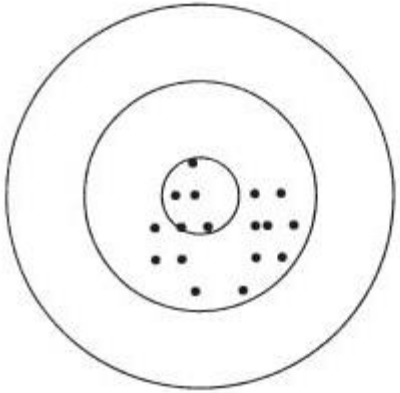
Neither valid nor reliable

The research methods do not hit the heart of the research aim (not 'valid') and repeated attempts are unfocussed



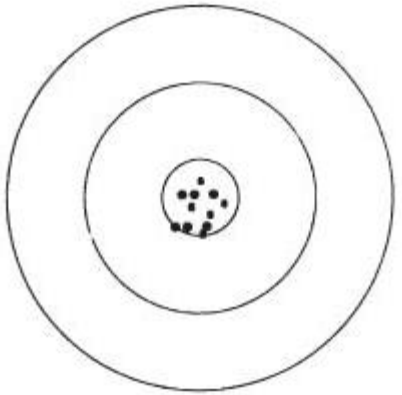
Reliable but not valid

The research methods do not hit the heart of the research aim, but repeated attempts get almost the same (but wrong) results



Fairly valid but not very reliable

The research methods hit the aim of the study fairly closely, but repeated attempts have very scattered results (not reliable)



Valid and reliable

The research methods hit the heart of the research aim, and repeated attempts all hit in the heart (similar results)

Tarea

- Visitar Sistema para la consulta de Estadísticas Históricas de México
  - Liga: http://dgcnesyp.inegi.org.mx/ehm/ehm.htm
- Buscar entre los Indicadores de los Objetivos de Desarrollo del Milenio:
  - La variable(s) con el registro más antiguo
  - La variable(s) con la serie de tiempo más larga
  - Buscar sus definiciones en el Glosario
  - Comentar si se tratan de mediciones válidas y confiables
- Anderson et al.: Resolver ejercicios (*suplementary exercises*) 1 a 10
- Entrega: jueves en el laboratorio