# Administración de datos

Gabriel Cavada Ch

Julio 2020

# Objetivos de la presentación

- Mostrar una base de datos con “pretensiones” de ser analizada
- Mostrar errores en la estructura
- Mostrar errores en la captura de información
- Mostrar errores en el procesamiento debido a lo anterior

# Ejemplo real

- Base de datos  que pretende describir el estado nutricional de una población escolar de estrato medio alto

# Algunas variables que contiene la base:

| | A | B |
|---|---|---|
| 1 | nombre | Nombre de pila |
| 2 | curso | Curso |
| 3 | fnac | Fecha de Nacimiento |
| 4 | fcuestionario | Fecha de aplicación del cuestionario |
| 5 | edad | Edad en años cumplidos |
| 6 | sexo | Sexo |
| 7 | pesonac | Peso de nacimiento en gramos |
| 8 | semgestac | Semanas de gestación |
| 9 | lacmaterna | Lactancia Materna |
| 10 | durlacmat | Duración de la lactancia materna |
| 11 | enfcardiaca | Presencia de enfermedad cardiaca |
| 12 | cualenf | Qué enfermedad |
| 13 | horasejersem | Horas de ejercicio semanales |
| 14 | horastv | Horas frente a la TV semanales |
| 15 | | |
| 16 | | |

# Un trozo de la base de datos original

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | nombre | curso | fnac | fcuestionario | edad | sexo | pesonac | semgestac | lacmaternı |
| 2 | Patricio | Séptimo Básico | 1995-10-18 | 2008-10-28 | 13 | Masculino | 3200 | 38 | Sí |
| 3 | Diego | Quinto Básico | 1997-10-21 | 2008-10-28 | 11 | Masculino | 3550 | 38 | Sí |
| 4 | Arturo | Segundo Medio | 1992-05-16 | 2008-11-13 | 16 | Masculino | 3.16 | 38 | No |
| 5 | Bárbara | Secto Básico | 1995-11-03 | 2008-11-13 | 13 | Femenino | 2.865 | 36 | No |
| 6 | Fernanda | Secto Básico | 1995-11-03 | 2008-11-13 | 13 | Femenino | 2.725 | 36 | No |
| 7 | Javier A. | Octavo Básico | 1993-08-28 | 2008-10-21 | 15 | Masculino | 2850 | 40 | Sí |
| 8 | Javier | Octavo Básico | 1993-08-28 | 2008-10-23 | 15 | Masculino | 2850 | 40 | Sí |
| 9 | Diego | Secto Básico | 1996-09-10 | 2008-10-23 | 12 | Masculino | 3250 | 40 | Sí |
| 10 | joaquin | Cuarto Básico | 1998-07-01 | 2008-10-22 | 10 | Masculino | 3.8 | 38 | Sí |
| 11 | Joaquin | Cuarto Básico | 1997-07-01 | 2008-11-12 | 11 | Masculino | 3.25 | 37 | Sí |
| 12 | Julian | Primero Basico | 2001-04-06 | 2008-11-12 | 7 | Masculino | 3.9 | 38 | Sí |
| 13 | Alvaro | Cuarto Medio | 1990-08-19 | 2008-11-01 | 18 | Masculino | 3.94 | 38 | Sí |
| 14 | Alfonso Eduardo | Cuarto Básico | 1998-09-20 | 2008-11-17 | 10 | Masculino | 3500 | 40 | Sí |
| 15 | LUCAS MARTIN | Segundo Medio | 2008-09-10 | 2008-10-21 | 0 | Masculino | 3200 | 36 | Sí |
| 16 | TALITHA ANTONIA | Tercero Basico | 1998-12-14 | 2008-10-21 | 9 | Femenino | 2900 | 36 | Sí |
| 17 | tomás | Tercero Basico | 1999-12-10 | 2008-10-20 | 8 | Masculino | 3400 | 40 | Sí |
| 18 | francisca | Quinto Básico | 1997-08-30 | 2008-10-20 | 11 | Femenino | 2900 | 41 | Sí |
| 19 | Maria Jose | Primero Basico | 01/11/2007 | 2008-10-23 | 0 | Femenino | 4.1 | 34 | No |
| 20 | Trinida | Tercero Basico | 1999-07-24 | 2008-10-20 | 9 | Femenino | 3250 | 39 | Sí |
| 21 | Rafael | Primero Basico | 2001-08-25 | 2008-10-20 | 7 | Masculino | 3300 | 38 | Sí |
| 22 | Valentina María | Primero Basico | 2001-02-15 | 2008-10-20 | 7 | Femenino | 3770 | 40 | Sí |
| 23 | nicolas | Cuarto Básico | 1998-06-07 | 2008-10-22 | 10 | Masculino | 3500 | 37 | Sí |

# La lucha del bioestadístico: ¿Es la información confiable?

| curso | Freq. | Percent | Cum. |
|---|---|---|---|
| Cuarto Básico | 21 | 8.68 | 8.68 |
| Cuarto Medio | 6 | 2.48 | 11.16 |
| Octavo Básico | 20 | 8.26 | 19.42 |
| Primero Basico | 26 | 10.74 | 30.17 |
| Primero Medio | 11 | 4.55 | 34.71 |
| Quinto Básico | 39 | 16.12 | 50.83 |
| Secto Básico | 24 | 9.92 | 60.74 |
| Segundo Basico | 24 | 9.92 | 70.66 |
| Segundo Básico | 4 | 1.65 | 72.31 |
| Segundo Medio | 13 | 5.37 | 77.69 |
| Sexto Básico | 3 | 1.24 | 78.93 |
| Séptimo Básico | 19 | 7.85 | 86.78 |
| Tercero Basico | 24 | 9.92 | 96.69 |
| Tercero Medio | 8 | 3.31 | 100.00 |
| Total | 242 | 100.00 | |

# La lucha del bioestadístico: ¿Es la información confiable?

| curso | Freq. | Percent | Cum. |
|---|---|---|---|
| Cuarto Básico | 21 | 8.68 | 8.68 |
| Cuarto Medio | 6 | 2.48 | 11.16 |
| Octavo Básico | 20 | 8.26 | 19.42 |
| Primero Basico | 26 | 10.74 | 30.17 |
| Primero Medio | 11 | 4.55 | 34.71 |
| Quinto Básico | 39 | 16.12 | 50.83 |
| Secto Básico | 24 | 9.92 | 60.74 |
| Segundo Basico | 24 | 9.92 | 70.66 |
| Segundo Básico | 4 | 1.65 | 72.31 |
| Segundo Medio | 13 | 5.37 | 77.69 |
| Sexto Básico | 3 | 1.24 | 78.93 |
| Séptimo Básico | 19 | 7.85 | 86.78 |
| Tercero Basico | 24 | 9.92 | 96.69 |
| Tercero Medio | 8 | 3.31 | 100.00 |
| Total | 242 | 100.00 | |

# La lucha del bioestadístico: ¿Es la información confiable?

edad

| Percentiles | | Smallest | | |
|---|---|---|---|---|
| 1% | -27 | -34 | | |
| 5% | 6 | -29 | | |
| 10% | 7 | -27 | Obs | 245 |
| 25% | 8 | -1 | Sum of Wgt. | 245 |
| 50% | 11 | | Mean | 60.86531 |
| | | Largest | Std. Dev. | 521.3283 |
| 75% | 13 | 217 | | |
| 90% | 16 | 218 | Variance | 271783.2 |
| 95% | 18 | 5789 | Skewness | 10.90019 |
| 99% | 218 | 5791 | Kurtosis | 120.0406 |

# La lucha del bioestadístico: ¿Es la información confiable?

```
. list  fnac fcuestionario edad if edad<4

     +-------------------------------+
     |       fnac   fcuestio~o   edad |
     |-------------------------------|
  14. | 2008-09-10   2008-10-21      0 |
  18. |      39387   2008-10-23      0 |
  78. | 1993-01-26   1967-01-02    -27 |
  79. | 1995-10-31   1967-01-02    -29 |
  80. | 2000-02-28   1967-01-02    -34 |
     |-------------------------------|
 147. | 2008-10-20   2008-10-20      0 |
 218. | 2008-11-28   2008-10-22     -1 |
     +-------------------------------+
```

```
. list  fnac fcuestionario edad if edad>18

     +-------------------------------+
     |       fnac   fcuestio~o   edad |
     |-------------------------------|
  72. |              2008-10-19    108 |
  73. | 1991-11-07   2208-11-13    217 |
  74. | 1990-03-05   2208-11-13    218 |
  96. |              2008-10-22    108 |
 136. |              2008-10-23    108 |
     |-------------------------------|
 153. |              2008-10-20    108 |
 171. | 1999-03-12      2149485   5789 |
 199. | 1997-02-11      2149485   5791 |
     +-------------------------------+
```

# La lucha del bioestadístico: ¿Es la información confiable?

| sexo | Freq. | Percent | Cum. |
|---|---|---|---|
| Femenino | 112 | 45.71 | 45.71 |
| Masculino | 129 | 52.65 | 98.37 |
| No Disponible | 3 | 1.22 | 99.59 |
| zx | 1 | 0.41 | 100.00 |
| Total | 245 | 100.00 | |

```
. list nombre sexo if sexo=="No Disponible"

     +---------------------------------+
     |          nombre            sexo |
     |---------------------------------|
  72. | Ignacio Antonio   No Disponible |
  96. |       Alejandro   No Disponible |
 136. |       Cristobal   No Disponible |
     +---------------------------------+

. list nombre sexo if sexo=="zx"

     +----------------+
     |  nombre   sexo |
     |----------------|
 109. | Antonia     zx |
     +----------------+
```

# La lucha del bioestadístico: ¿Es la información confiable?

```
. sum pesonac,d

                            pesonac
-------------------------------------------------------------
      Percentiles      Smallest
 1%          2.33           2.1
 5%          2.78          2.23
10%             3          2.33       Obs                 236
25%        3.7835          2.61       Sum of Wgt.         236

50%          3000                     Mean           2348.04
                        Largest       Std. Dev.      1993.641
75%          3410          4300
90%          3680          4500       Variance       3974604
95%          3860          4500       Skewness       3.715036
99%          4500         22000       Kurtosis       40.99131
```

# La lucha del bioestadístico: Es la información confiable?

```
. tab  lacmaterna

   lacmaterna |      Freq.     Percent        Cum.
--------------+-----------------------------------
           No |         38       15.51       15.51
No Disponible |          3        1.22       16.73
           Sí |        204       83.27      100.00
--------------+-----------------------------------
        Total |        245      100.00
. sum   durlacmat, d

                        durlacmat
-------------------------------------------------------------
      Percentiles      Smallest
 1%           -99            -99
 5%           -99            -99
10%           -99            -99       Obs                 241
25%             2            -99       Sum of Wgt.         241

50%             4                      Mean           -11.22407
                        Largest        Std. Dev.      38.19914
75%             6             12
90%             8             12       Variance       1459.175
95%            10             12       Skewness      -1.844665
99%            12             40       Kurtosis       4.478617
```

# La lucha del bioestadístico: Es la información confiable?

```
. tab  durlacmat  lacmaterna, missing
```

| durlacmat | lacmaterna | | | Total |
|---|---|---|---|---|
| | No | No Dispon | Sí | |
| -99 | 38 | 0 | 0 | 38 |
| 1 | 0 | 0 | 17 | 17 |
| 2 | 0 | 0 | 16 | 16 |
| 3 | 0 | 0 | 34 | 34 |
| 4 | 0 | 0 | 32 | 32 |
| 5 | 0 | 0 | 12 | 12 |
| 6 | 0 | 0 | 40 | 40 |
| 7 | 0 | 0 | 16 | 16 |
| 8 | 0 | 0 | 13 | 13 |
| 9 | 0 | 0 | 9 | 9 |
| 10 | 0 | 0 | 7 | 7 |
| 11 | 0 | 0 | 1 | 1 |
| 12 | 0 | 0 | 5 | 5 |
| 40 | 0 | 0 | 1 | 1 |
| . | 0 | 3 | 1 | 4 |
| Total | 38 | 3 | 204 | 245 |

# La lucha del bioestadístico: Es la información confiable?

```
. list  cualenf if  enfcardiaca=="Sí"

     +-------------------------------------------------------------------------+
     |                                                                 cualenf |
     |-------------------------------------------------------------------------|
  56. |        COARTACION AORTICA SEVERA OPERADA, ESTENOSIS LEVE VALVULA MITRAL |
 105. | soplo inocente extrasistoles ventriculares \"inocentes\" hasta la fecha |
 214. |                                                       soplo inofensivo |
     +-------------------------------------------------------------------------+
```

```
. tabstat  horasejersem horastv, stat(q mean sd) col(stat)

    variable |       p25       p50       p75      mean        sd
-------------+--------------------------------------------------
horasejersem |         4         6         9  2.565126  66.18048
     horastv |         1         1         2  1.712605  6.806122
----------------------------------------------------------------
```

# La lucha del bioestadístico: Es la información confiable?

```
. sum  horasejersem horastv,d

                    horasejersem
-------------------------------------------------------------
      Percentiles      Smallest
 1%             1          -1012
 5%             2              1
10%             2              1       Obs                 238
25%             4              1       Sum of Wgt.         238

50%             6                      Mean           2.565126
                        Largest        Std. Dev.      66.18048
75%             9             16
90%            12             20       Variance       4379.855
95%            15             20       Skewness      -15.23231
99%            20             40       Kurtosis       234.0354

                       horastv
-------------------------------------------------------------
      Percentiles      Smallest
 1%             0            -34
 5%            .5            -23
10%             1              0       Obs                 238
25%             1              0       Sum of Wgt.         238

50%             1                      Mean           1.712605
                        Largest        Std. Dev.      6.806122
75%             2             14
90%             2             14       Variance        46.3233
95%             3             30       Skewness       8.772069
99%            14             90       Kurtosis       125.2766
```

# CODIFICACIÓN DE VARIABLES

Las variables numéricas (peso, talla, edad) regístrelas como tal, como números.
Las variables ordinales (EVA, calidad de vida) también regístrelas como tal.
Las variables categóricas regístrelas usando lógica booliana: 0 si la respuesta es negativa y 1 si la respuesta es positiva

| | femenino | en años | en kg | en mt | | | | |
|---|---|---|---|---|---|---|---|---|
| id | sexo | edad | peso | talla | hta | db | cardiopatia | muerto |
| 1 | 1 | 55 | 70 | 1.66 | 1 | 0 | 0 | 0 |
| 2 | 1 | 53 | 73 | 1.72 | 0 | 0 | 1 | 0 |
| 3 | 0 | 60 | 72 | 1.8 | 0 | 1 | 1 | 0 |
| 4 | 1 | 59 | 80 | 1.63 | 1 | 1 | 0 | 1 |

•Si un programa estadístico es alimentado con basura, éste devolverá basura procesada

•Una excelente idea de investigación y mucho tiempo invertido en capturar la información puede reducirse a la nada si no se cuida la calidad de los datos

## CALIDAD DE LOS DATOS

LOS ERRORES APARENTEMENTE PEQUEÑOS DEL MUESTREO, LA MEDICIÓN Y EL REGISTRO DE DATOS PUEDEN ACABAR CON CUALQUIER ANÁLISIS. **R. A. FISHER**, ESTUDIOSO DE LA GENÉTICA Y FUNDADOR DE LA ESTADÍSTICA MODERNA, NO SÓLO DISEÑABA Y ANALIZABA LA CRÍA DE ANIMALES, SINO QUE TAMBIÉN **LIMPIABA SUS JAULAS** Y **CUIDABA DE ELLOS**, PORQUE SABÍA QUE LA PÉRDIDA DE UN ANIMAL INFLUIRÍA EN SUS RESULTADOS.