

Prácticas de estadística na aula: por que non usar *Epidat*?

GAEL NAVEIRA BARBEITO
SOLY SANTIAGO PÉREZ

Neste artigo preséntase o programa *Epidat* como unha ferramenta de grande utilidade na ensinanza da estatística, de cara ás actividades prácticas da materia. O programa caracterízase fundamentalmente por ser de libre acceso, de fácil manexo e por dispoñer dunha axuda moi completa. Entre os seus módulos inclúense varios de estatística: Mostraxe, Análise descritiva, Inferencia sobre parámetros e Distribucións de probabilidade. *Epidat* está desenvolvido pola Dirección Xeral de Saúde Pública e pode descargarse de forma gratuíta na dirección:

<https://www.sergas.gal/saude-publica/epidat>

Palabras chave: Ensinanza, Estatística, Programa informático.

Statistical Practices in the Classroom: Why Not Use *Epidat*?

In this paper we show the *Epidat* software as an useful tool in teaching Statistics, specially for practical applications. Its main features are it is free, easy to use, and it has a comprehensible help tool. The programme has several statistical modules: Sampling, Descriptive analysis, Inference on parameters and Probability distributions. *Epidat* has been developed by Dirección Xeral de Saúde Pública and is available for free download at the following website:

<https://www.sergas.gal/saude-publica/epidat>

Keywords: Teaching, Statistics, Software.

A estatística está presente en practicamente todos os aspectos da nosa vida, en moitas ocasións dun modo subxacente. Por exemplo, algúns adoitamos comezar o día aplicando a estatística, á hora de elixir o noso vestiario segundo as predicións meteorolóxicas de probabilidade de choiva ou frío, e acostumamos a rematalo interpretando a estatística, botándolle un ollo á prensa que nos informa sobre a taxa de paro ou a intención de voto para as próximas eleccións.

Unha boa cultura estatística é básica para coñecer e comprender o mundo que nos rodea, saber interpretalo e tomar as decisións máis adecuadas en cada caso. Ese coñecemento debe ser promovido dende a ensinanza secundaria, e incluso primaria, para poder acadar unha base sólida que nos sirva e axude o resto da nosa vida. Conseguir esa base implica adquirir os coñecementos teóricos e reforzalos con experiencias prácticas que permitan entender a importancia de propor un obxectivo claro a estudo, valorar a calidade dos datos recollidos e aprender a interpretar correctamente os resultados.

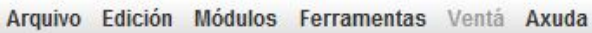
A utilización de calculadoras e ferramentas tecnolóxicas para o tratamento de datos, creación e interpretación de gráficos e elaboración de informes está contemplado no novo currículo da LOMLOE para todos os cursos da ESO [1]. Poñer en práctica estas actividades na aula implica dispoñer de ferramentas informáticas gratuítas, doadas de usar, e que conteñan os métodos estatísticos necesarios. Na Dirección Xeral de Saúde Pública da Xunta de Galicia levamos case trinta anos traballando no desenvolvemento dun programa que cumpre todos estes requisitos, *Epidat*, no que tamén colaboran a Organización Panamericana da Saúde (OPS-OMS) e a Universidade CES de Medellín, Colombia [2, 3]. Aínda que *Epidat* naceu con orientación epidemiolóxica, o seu contido foise ampliando nas sucesivas versións, que xa son 4, e actualmente contén unha boa variedade de métodos tanto epidemiolóxicos coma estatísticos, o que permite o seu uso en distintos contextos. A última versión, *Epidat* 4.2 [4], ten 12 módulos dos cales 4 son específicos de estatística: Mostraxe, Análise descritiva, Inferencia sobre parámetros e Distribucións de probabilidade.


Ademais do seu contido estatístico, *Epidat* 4.2 ten outras características que o converte nunha ferramenta de grande utilidade para apoiar e facilitar a docencia da estatística: é de libre acceso; está programado en Java, o que permite o funcionamento da aplicación en distintos sistemas operativos (Windows, Linux, OS X); o seu manexo é sinxelo e, por último, cabe destacar a calidade da súa axuda, que inclúe exemplos resoltos e comentados.

Comezando a traballar con *Epidat* 4

O programa está dispoñible para a súa descarga na páxina web da Dirección Xeral de Saúde Pública (<https://www.sergas.es/Saude-publica/epidat>, sección de Aplicacións e Datos). Non precisa de instalación, soamente é necesario descomprimir o arquivo que se descarga e gardar a carpeta de *Epidat* na localización desexada do ordenador (preferiblemente, con permisos de escritura para que a opción do idioma funcione correctamente). Dentro desta carpeta atópase o arquivo “Epidat.jar” que permite arrancar a aplicación, facendo dobre clic nel.

O entorno de *Epidat* 4.2 é unha ventá (figura 1) composta por:

- Unha barra de menús , que permite, entre outras cousas, acceder aos distintos módulos (menú *Módulos*) e aos documentos de axuda (menú *Axuda*). Tamén permite modificar o idioma no que se amosa a aplicación (menú *Ferramentas*), podendo elixir entre castelán, galego, portugués e inglés.

- Unha barra de ferramentas , que permite xestionar os resultados xerados co programa, abrindo ou gardando un arquivo de resultados; copiando, pegando ou cortando resultados dentro da propia aplicación.
- Unha sección de *Índice de cálculos*, na que se van acumulando os índices de cada cálculo realizado co programa, para facilitar o manexo dos mesmos.
- Unha sección de *Módulos en uso*, na que se van listando todos os módulos que permanecen abertos.
- Unha sección de *Resultados*, na que se van acumulando todos os resultados xerados polo programa, tanto numéricos como gráficos.

Epidat non xestiona bases de datos polo que, á hora de comezar a traballar, o primeiro que se debe facer é seleccionar o módulo e a ventá do módulo coa que se pretende facer o cálculo de interese.

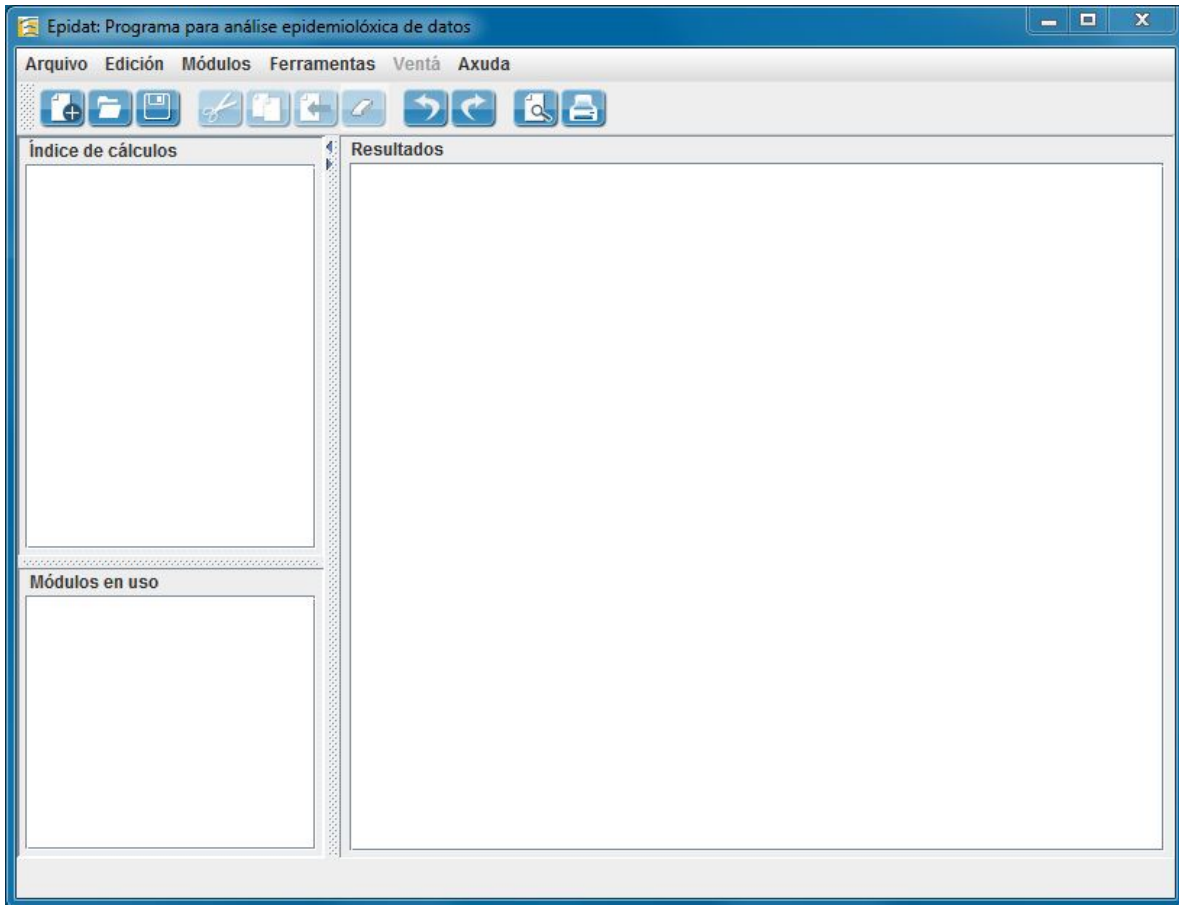


Figura 1: Ventá principal de *Epidat* 4.2.

Un paseo por *Epidat* 4

Nesta sección faremos un percorrido polo programa, empregando un exemplo ficticio que permitirá explorar os catro módulos de estatística: Mostraxe, Análise descritiva, Inferencia sobre parámetros e Distribucións de probabilidade. Tomarase o seguinte exemplo:

Nun colexio de ensinanza secundaria obrigatoria, con alumnado de 12 a 16 anos, principalmente, preténdese levar a cabo unha enquisa sobre factores e condutas que inflúen na saúde (consumo de tabaco, obesidade, práctica de actividade física,...). O colexio dispón de 7 aulas de 1.º de ESO, 6 aulas de 2.º, 6 aulas de 3.º e 5 aulas de 4.º. Por motivos operativos e de escaseza de tempo, pensouse en enquisar a todo o alumnado de dúas aulas escollidas ao azar para cada curso. Isto equivale a facer unha mostraxe aleatoria de aulas, estratificada por curso. Con *Epidat* 4.2 esta selección pódese facer usando o módulo de Mostraxe.

Características principais do módulo de Mostraxe

Inclúe procedementos para determinar tamaños de mostra, facer a selección de mostras probabilísticas con diferentes deseños e realizar o tratamento de datos obtidos mediante unha mostraxe complexa. O seu contido organízase da seguinte maneira:

1. Cálculo de tamaños de mostra:
 - 1.1. Intervalos de confianza.
 - 1.2. Contrastes de hipóteses.
2. Selección de mostras:

<ol style="list-style-type: none"> 2.1. Mostraxe simple aleatoria. 2.2. Mostraxe sistemática. 2.3. Mostraxe aleatoria estratificada. 2.4. Mostraxe por conglomerados monoetápica. 	<ol style="list-style-type: none"> 2.5. Mostraxe por conglomerados bietápica. 2.6. Mostraxe por conglomerados monoetápica estratificada. 2.7. Mostraxe por conglomerados bietápica estratificada.
---	--
3. Estimación con mostras complexas.

En toda investigación por mostraxe, os resultados estarán afectados polo denominado *erro de mostraxe*, que se debe ao feito de que os resultados observados nunha mostra se xeneralizan a toda a poboación da que procede. Esta é a esencia da inferencia estatística, que nos proporciona as técnicas para cuantificar o erro aleatorio cometido, o cal depende, como nos di a intuición e nos confirma a estatística, do tamaño da mostra: canto máis grande sexa a mostra, menor será o erro cometido. Pois ben, o cálculo do tamaño de mostra necesario para realizar un estudo baséase, de xeito xeral, na idea de limitar o erro que estamos dispostos a cometer. O método concreto que se debe utilizar depende do obxectivo do estudo que estamos planificando e do procedemento estatístico que se vai empregar para dar resposta a ese obxectivo: estimación dun parámetro (por exemplo, *proporción de alumnado do centro que fuma*) ou contraste de hipóteses (por exemplo, *responder á pregunta: hai máis fumadores homes ou mulleres?*). *Epidat* inclúe un amplo abano de opcións para determinar tamaños de mostra nos dous contextos descritos: estimación dun parámetro cunha determinada precisión, que equivale a limitar a amplitude dun intervalo de confianza, ou realización dun contraste de hipóteses cunha determinada potencia. En todos os casos é tamén posible realizar o cálculo inverso, precisión ou potencia en función do tamaño de mostra.

Tras determinar o tamaño de mostra necesario, o seguinte paso no deseño dun estudo é seleccionar os individuos que se incluírán no mesmo. Para esta tarefa, *Epidat 4.2* proporciona distintos procedementos probabilísticos de mostraxe que abarcan dende o máis sinxelo, a mostraxe aleatoria simple, ata algún dos máis complexos, como é a mostraxe por conglomerados bietápica estratificada. En todos os casos, os datos pódense cargar de forma automática a partir dun arquivo coa listaxe de elementos que compoñen a poboación (marco de mostraxe) e a mostra resultante pódese gardar nun arquivo en formato de *Excel* ou de *Calc* que terá a mesma estrutura que o orixinal (en canto ás variables recollidas) pero só incluírá os elementos que resulten seleccionados. Tamén é posible, cando non se dispón do marco de mostraxe detallado, indicar o tamaño da poboación e obter números aleatorios que identifican os elementos da mostra.

Por último, o submódulo de estimación con mostras complexas permite estimar medias e proporcións, cos seus intervalos de confianza, a partir de mostras complexas, e dá a posibilidade de definir filtros.

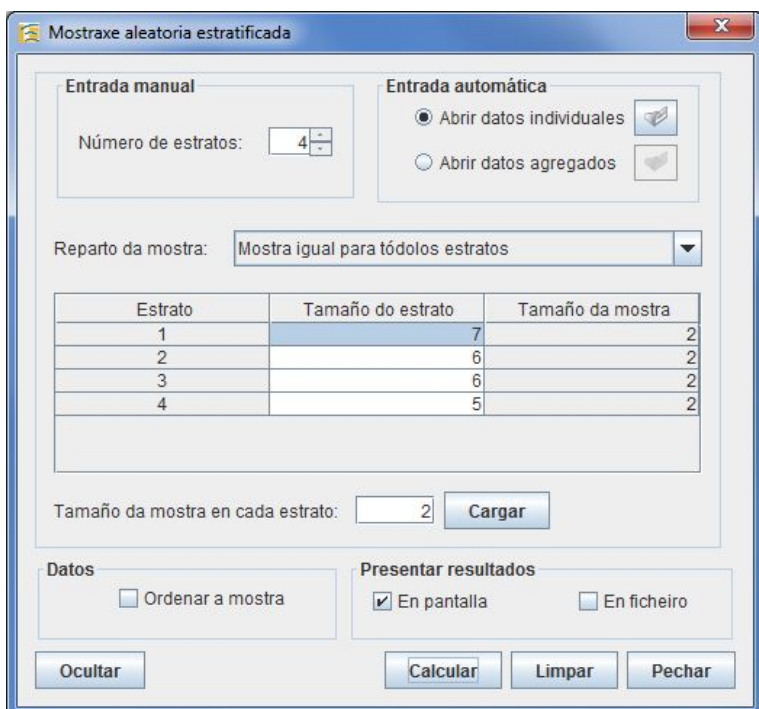
No exemplo que nos ocupa, prescindimos do cálculo de tamaño de mostra, polos motivos expostos anteriormente, e procedemos á selección aleatoria de dúas aulas en cada curso. A opción elixida para o cálculo é a **Mostraxe aleatoria estratificada** (figura 2). A entrada de datos realízase de forma manual, indicando o número de estratos, que é o número de cursos do colexio (4), e o tamaño de cada estrato, que é o número de aulas de cada curso (7 de 1.º, 6 de 2.º, 6 de 3.º e 5 de 4.º). O tamaño de mostra a seleccionar sería 2 en cada curso ou estrato. A selección proporcionada por *Epidat* móstrase na figura 2.

Dado que as clases se nomean con letras, as aulas seleccionadas corresponden a 1.º B, 1.º F, 2.º B, 2.º E, 3.º D, 3.º E, 4.º A e 4.º D.

Tras enquisar a todo o alumnado das aulas que caeron na mostra ($n = 190$), os datos recollidos grávanse nunha folla de cálculo, coma Excel ou Calc, de xeito que cada fila represente a unha persoa e cada columna a unha variable do cuestionario. Á marxe do programa que se use para analizar os datos, é recomendable codificar con números as variables categóricas (por exemplo: 1-home, 2-muller, para o sexo) o que facilitará a gravación dos cuestionarios e reducirá os erros.

Unha vez informatizados os datos, o primeiro paso na análise dos mesmos sempre debe ser unha análise descritiva, para explorar a distribución de valores de todas as variables, ben por medio de táboas ou de gráficos. Este paso permite, non só resumir a información recollida, senón comprobar que os datos se gravaron correctamente ou descartar variables cun elevado número de valores ausentes (por non resposta).

No exemplo, comezamos a análise describindo a distribución dos alumnos da mostra segundo a clase á que pertencen. Para iso emprégase o módulo de Análise descritiva.



[1] Mostraxe aleatoria estratificada:

Número dos suxeitos seleccionados:

Estrato 1:

6 2

Estrato 2:

2 5

Estrato 3:

5 4

Estrato 4:

4 1

Figura 2: Ventá para a selección dunha mostra aleatoria estratificada e resultados da selección. Módulo de Mostraxe de *Epidat* 4.2.

Características principais do módulo de Análise descritiva

Ofrece a posibilidade de organizar e resumir un conxunto de datos mediante táboas, gráficos e medidas numéricas de resumo. O seu contido detallado é o seguinte:

- | | |
|------------------------------|-------------------------------|
| 1. Táboas de frecuencias. | 5. Gráficos: |
| 2. Táboas de continxencia. | 5.1. Barras. |
| 3. Estatísticos descritivos. | 5.2. Sectores. |
| 4. Correlación. | 5.3. Liñas. |
| | 5.4. Dispersión. |
| | 5.5. Histograma. |
| | 5.6. Diagrama de caixas. |
| | 5.7. Intervalos de confianza. |

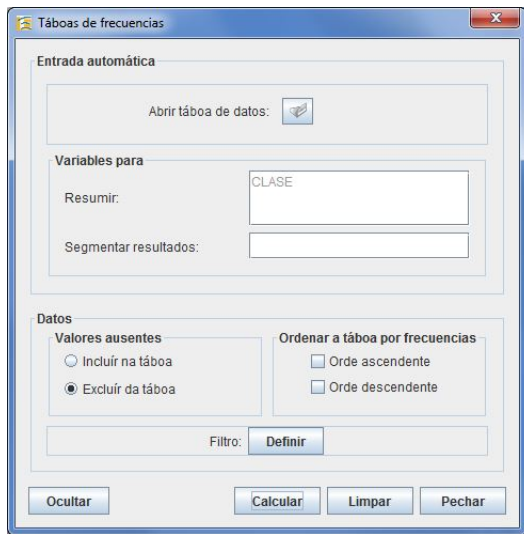
Exceptuando os gráficos de liñas e de intervalos de confianza, todas as opcións deste módulo teñen un funcionamento semellante:

- A entrada de datos realízase unicamente de xeito automático, identificando, na base de datos en uso, as variables necesarias para a análise que se pretende realizar.
- É posible segmentar os resultados ou definir filtros para analizar un subconxunto dos datos.
- Os gráficos xerados dispoñen dun editor, presente en todo o programa, que permite personalizalos e gardalos con formato imaxe (jpg ou bmp).

Respecto ás dúas opcións deste módulo que non comparten o funcionamento xeral do resto (gráficos de liñas e de intervalos de confianza), a diferenza fundamental radica en que o programa non fai ningún resumo dos datos, senón que representa directamente os valores que introduce o usuario. Por este motivo, tamén é posible introducir os datos manualmente, ademais de importalos a partir dun arquivo.

Volvendo ao exemplo da enquisa, imos analizar a distribución da mostra por clase calculando unha táboa de frecuencias coa correspondente opción do programa (figura 3). Os datos cárganse por medio dun asistente ao que se accede pulsando no botón “Abrir táboa de datos”, e que permite seleccionar o arquivo, a folla onde se atopan os datos, e as variables a resumir (CLASE). Os resultados que se obteñen co programa pódense ver tamén na figura 3. Os datos mostran o número de persoas que cubriron a enquisa en cada unha das aulas seleccionadas, e a porcentaxe que supoñen con respecto ao total da mostra ($n = 190$).

No cuestionario da nosa enquisa preguntábase pola práctica de actividade física extraescolar, con dúas opcións de resposta: non (valor 0) ou si (valor 1). Para estimar a porcentaxe de alumnado do centro que practica algunha actividade, temos que calcular a porcentaxe observada na mostra e acompañala cun intervalo de confianza. Tamén pode ser interesante comparar os resultados entre rapaces e rapazas mediante unha proba de hipótese. Para isto emprégase o módulo de Inferencia sobre parámetros.



[2] Táboas de frecuencias:

Entrada automática:

Arquivo de traballo: C:\Arquivos de programa\Epidat 4\Revista GAMMA\Enquisa.xlsx
 Táboa: Data

Variables:
 Resumir: CLASE

Resultados:

Frecuencias para a variable CLASE:

VALOR	Frecuencia	Porcentaxe	Porcentaxe acumulada
1ºB	21	11,05	11,05
1ºF	25	13,16	24,21
2ºB	29	15,26	39,47
2ºE	25	13,16	52,63
3ºD	24	12,63	65,26
3ºE	22	11,58	76,84
4ºA	19	10,00	86,84
4ºD	25	13,16	100,00
TOTAL	190	100,00	

Figura 3: Ventá para obter táboas de frecuencias e distribución da mostra por clase. Módulo de Análise descritiva de *Epidat* 4.2.

Características principais do módulo de Inferencia sobre parámetros

Permite facer inferencia sobre unha ou dúas poboacións e contrastar a normalidade dos datos. O seu contido detallado é o seguinte:

- | | |
|---|--|
| <ol style="list-style-type: none"> 1. Unha poboación: <ol style="list-style-type: none"> 1.1. Media. 1.2. Proporción. 1.3. Percentís. 1.4. Correlación. 1.5. Taxa de incidencia. 1.6. Índice de posición. | <ol style="list-style-type: none"> 2. Dúas poboacións: <ol style="list-style-type: none"> 2.1. Medias independentes. 2.2. Medias emparelladas. 2.3. Proporcións independentes. 2.4. Proporcións emparelladas. 2.5. Taxas de incidencia. 3. Comparación non paramétrica: <ol style="list-style-type: none"> 3.1. Medias independentes. 3.2. Medias emparelladas. 4. Contraste de normalidade. |
|---|--|

As opcións do submódulo de **Unha poboación** e **Dúas poboacións** permiten obter estimacións puntuais, intervalos de confianza e realizar contrastes de hipótese para o parámetro en cuestión. Nestes casos os datos de entrada poden introducirse manualmente, xa resumidos, ou resumirse a partir dun arquivo (con formato *Excel* ou *Calc*) que conteña a información individual.

As opcións de **Comparación non paramétrica** permiten comparar as medias de dúas poboacións sen asumir ningunha distribución para os datos. A entrada de datos realízase única e exclusivamente de forma automática, do mesmo xeito que na opción de **Contraste de normalidade**. Esta última opción permite analizar a normalidade dos datos por medio de dúas probas de hipótese (Shapiro-Francia e Asimetría-Curtose) e unha representación gráfica (gráfico cuantil-cuantil).

Cos datos do exemplo podemos estimar a porcentaxe de alumnado que fai actividade física extraescolar usando a opción de inferencia sobre unha proporción de *Epidat* 4.2 (figura 4). Os datos pódense cargar de xeito automático a partir do arquivo da enquisa e o programa mostra a información resumida: o número de persoas que respondeu a esta pregunta é 175 (tamaño de mostra) e, deles, fan algunha actividade 131 (número de casos). Estes datos resumidos tamén se poderían introducir manualmente. Os resultados que proporciona *Epidat* (figura 5) indican que un 75% do alumnado do centro practica actividade física extraescolar, pero este valor é unha estimación baseada na mostra; o intervalo de confianza indícanos que a porcentaxe verdadeira está entre 68% e 81% cunha confianza do 95%.

Por outra banda, temos a hipótese de que os rapaces fan máis actividade que as rapazas e, para comprobalo, imos á opción de comparación de proporcións independentes. De novo cargamos os datos a partir do arquivo da enquisa e visualizamos na ventá de *Epidat* 4.2 os datos resumidos (figura 4); a mostra 1 corresponde aos rapaces (sexo = 1) e a mostra 2 ás rapazas (sexo = 2). Seleccionando como opción de cálculo un contraste de hipótese unilateral dereito obtemos os resultados que se mostran na figura 5. As estimacións puntuais, 81% en homes e 69% en mulleres, corroboran a nosa hipótese, pero de novo estes datos están afectados polo erro de mostraxe polo que debemos ter en conta os resultados da proba de hipótese. O valor p da proba é a probabilidade de obter os resultados observados na mostra supoñendo que na poboación, todo o colexio, é certa a hipótese nula de que non hai diferenza entre nenos e nenas. Como esta probabilidade é pequena (0,037), menor que o nivel de significación 0,05, rexeitamos a hipótese nula en favor da alternativa ($P_1 - P_2 > 0$, é dicir, $P_1 > P_2$). En resumo, concluímos que a proporción de alumnado que practica actividade física é maior nos rapaces.

O valor p da proba de hipótese realizada é unha probabilidade que se calcula a partir da distribución do estatístico de contraste, unha normal estándar. O dito valor é a probabilidade de que o estatístico Z sexa maior que o valor obtido con esta mostra, que foi 1,786:

$$\text{Valor } p = \text{Probabilidade } \{Z > 1,786\}$$

Podemos calcular esta probabilidade usando o módulo de Distribucións de probabilidade de *Epidat* 4.2.

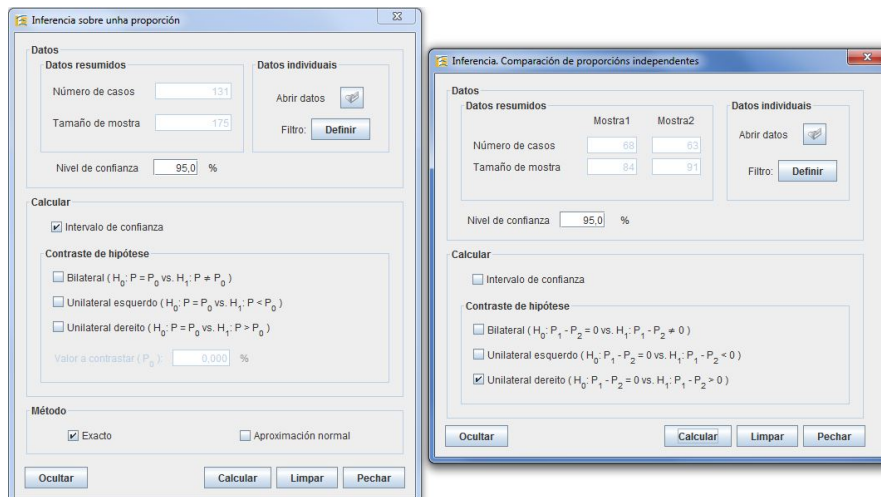


Figura 4: Ventás para facer inferencia sobre unha e dúas proporcións. Módulo de Inferencia sobre parámetros de *Epidat* 4.2.

[3] Inferencia sobre unha proporción:

Resultados:

Método exacto

Intervalo de confianza (95,0%)

Porcentaxe (%)	Límite inferior (%)	Límite superior (%)
74,857	67,754	81,098

[4] Inferencia. Comparación de proporcións independentes:

Resultados:

Poboación	Porcentaxe (%)
Nenos	80,952
Nenas	69,231

Intervalo de confianza (95,0%)

Diferenza de proporcións	Límite inferior	Límite superior
0,117	-0,009	0,244

Proba de comparación de proporcións

Contraste	Estatístico z	Valor p
Unilateral dereito	1,786	0,037

Figura 5: Resultados obtidos con *Epidat* 4.2 para a porcentaxe de alumnado que realiza algunha actividade física extraescolar.

Características principais do módulo de Distribucións de probabilidade

Permite calcular valores da función de distribución e a súa inversa para un amplo abano de distribucións discretas e continuas, así como representar graficamente as funcións de densidade e de distribución. Tamén permite xerar mostras simuladas para cada unha das distribucións incluídas. As opcións deste módulo son:

- | | |
|-------------------------------|-------------------------------|
| 1. Cálculo de probabilidades: | 2. Xeración de distribucións: |
| 1.1. Distribucións discretas. | 2.1. Distribucións discretas. |
| 1.2. Distribucións continuas. | 2.2. Distribucións continuas. |

En calquera das catro opcións, o programa presenta unha lista de distribucións, discretas ou continuas segundo o caso, entre as que o usuario pode elixir a distribución de interese e introducir como datos os seus parámetros. Os resultados inclúen os valores que caracterizan a distribución seleccionada: media, mediana, moda, varianza, asimetría e curtose (cando existen). No caso da opción de cálculo de probabilidades, está integrado o editor de gráficos para poder modificar as funcións de distribución e densidade. En xeración de distribucións, *Epidat* ofrece ao usuario a posibilidade de presentar os valores simulados en pantalla ou gardalos nun arquivo, as dúas opcións non son excluíntes.

Para calcular o valor p , abrimos a ventá de Cálculo de probabilidades, Distribucións continuas, seleccionamos a distribución normal e introducimos os seus parámetros: media = 0, desvío estándar = 1. A continuación eliximos calcular unha probabilidade e damos o valor do punto: 1,786. Nos resultados móstranse varias probabilidades asociadas con ese punto; como podemos comprobar, o valor p da nosa proba de hipótese correspóndese coa cola dereita: 0,037 (figura 6).

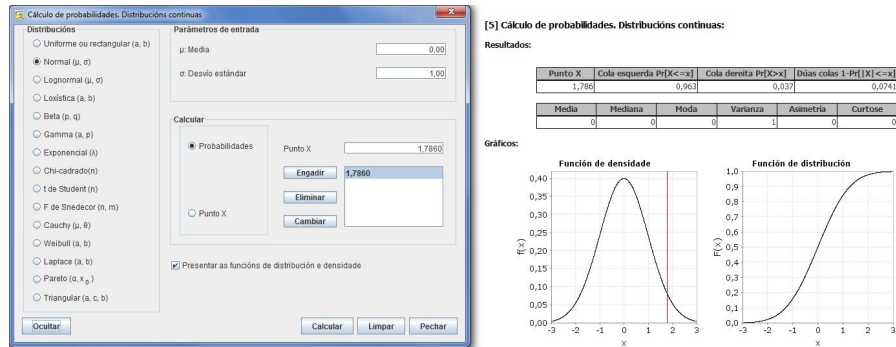


Figura 6: Ventá de cálculo de probabilidades para distribucións continuas e resultados. Módulo de Distribucións de probabilidade de *Epidat* 4.2.

Aquí remata este breve paseo por *Epidat* 4.2 no que tratamos de mostrar algunhas das posibilidades do programa para facer prácticas de estadística. Ademais dos métodos que inclúe, o feito de ser gratuito, válido para distintos sistemas operativos, e de manexo sinxelo, convérten a *Epidat* nunha ferramenta destacada para o seu uso na aula.

Para calquera cuestión relacionada co programa está dispoñible a seguinte dirección de correo á que dirixirse: <soporte.epidat@sergas.es>.

Referencias bibliográficas

- [1] España: *Real Decreto 217/2022, de 29 de marzo, polo que se establece o currículo básico da Educación Secundaria Obligatoria*. Boletín Oficial del Estado (BOE), 76, pp. 41571-41789.
- [2] Hervada Vidal X., M.I. Santiago Pérez, E. Vázquez Fernández, C. Castillo Salgado, E. Loyola Elizondo e L.C. Silva Ayçaguer (2004): “Epidat 3.0: Programa para el análisis epidemiológico de datos tabulados. Versión 3.0”, *Revista Española de Salud Pública*, 78(2), pp. 277-280.
- [3] Santiago Pérez M.I., X. Hervada Vidal, G. Naveira Barbeito, L.C. Silva, H. Fariñas, E. Vázquez, J. Bacallao e O.J. Mújica (2010): “El programa Epidat: usos y perspectivas [Carta al editor]”, *Revista Panamericana de Salud Pública*, 27(1), pp. 80-82.

Referencias en liña

- [4] Consellería de Sanidade, Xunta de Galicia; Organización Panamericana de la Salud (OPS-OMS) e Universidad CES, Colombia (2016): *Epidat: programa para análise epidemiolóxica de datos. Versión 4.2, xullo 2016*. Dispoñible en: <https://www.sergas.gal/Saude-publica/EPIDAT-4-2>.

Gael Naveira Barbeito

Dirección Xeral de Saúde Pública. Consellería de Sanidade. Xunta de Galicia.

<gael.naveira.barbeito@sergas.es>

Soly Santiago Pérez

Dirección Xeral de Saúde Pública. Consellería de Sanidade. Xunta de Galicia.

<soly.santiago.perez@sergas.es>