



Análisis de datos

Jalil Varela Manjarres

Universidad del Valle
LA-CoNGA physics

28 de febrero de 2021

Contenido

- 1 Introducción
- 2 Extracción de datos
- 3 Estadística por géneros
- 4 Estadística por tiempo
- 5 Estadística por especie
- 6 Conclusiones

¿Que se busca hacer?

- En el presente trabajo se busca poner a prueba el conocimiento adquirido, sobre la manipulación y el tratamiento de una gran cantidad de datos.
- Por su versatilidad a la hora de almacenar, graficar y analizar diferentes tipos de datos, se utilizan las siguientes librerías.
 - pandas
 - matplotlib
 - numpy
- Se utiliza además como fuente de datos un archivo csv provisto en clase, denominado 'surveys.csv', sin el conocimiento previo de que tipo de datos almacena y con que motivo se almacenaron los datos, con el objetivo final de deducir la mayor cantidad de información a partir de estos.

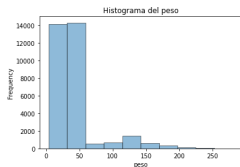
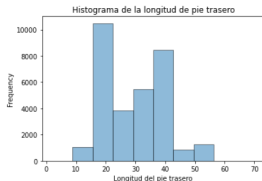
Extracción de datos

- La librería pandas, nos permite extraer y almacenarlos en una variable tipo DataFrame, que tiene las propiedades de una tabla.
- El sistema reconoce los caracteres y los organiza de acuerdo a diferentes filas y columnas, las cuales estan asignadas en el archivo csv original, se obtiene algo de la siguiente forma.

	record_id	month	day	year	plot_id	species_id	sex	hindfoot_length	weight
0	1	7	16	1977	2	NL	M	32.0	NaN
1	2	7	16	1977	3	NL	M	33.0	NaN
2	3	7	16	1977	2	DM	F	37.0	NaN
3	4	7	16	1977	7	DM	M	36.0	NaN
4	5	7	16	1977	3	DM	M	35.0	NaN

Análisis por columna

- Es posible analizar la estadística de una sola de las columnas de la tabla resultante, para el caso de estudio, vemos que hay dos propiedades en la encuesta, que son ("hindfoot _length", "weight").
- Un histograma que nos de una idea de como se distribuyen estas características en todas las muestras.



- Se obtienen los valores estadísticos de la columna, como la media, la desviación estándar, el valor mínimo y el valor máximo.

	Longitud del pie trasero	peso
count	31438.000000	32283.000000
mean	29.287932	42.672428
std	9.564759	36.631259
min	2.000000	4.000000
25%	21.000000	20.000000
50%	32.000000	37.000000
75%	36.000000	48.000000
max	70.000000	280.000000

- Esto nos ayuda a dar un primer vistazo al comportamiento estadístico de toda la encuesta.

Estadística por géneros

- Para entender mejor los datos, podemos empezar separándolos por sexo y viendo la estadística general que siguen las propiedades de los encuestados, para conocer sus similitudes y diferencias.
- Comparación de la estadística de las longitudes del pie trasero para diferentes sexos.
- Comparación de la estadística del peso para los sexos.

	Longitud del pie trasero F	Longitud del pie trasero M
count	14894.000000	16476.000000
mean	28.836780	29.709578
std	9.463789	9.629246
min	7.000000	2.000000
25%	21.000000	21.000000
50%	27.000000	34.000000
75%	36.000000	36.000000
max	64.000000	58.000000

	Peso F	Peso M
count	15303.000000	16879.000000
mean	42.170555	42.995379
std	36.847958	36.184981
min	4.000000	4.000000
25%	20.000000	20.000000
50%	34.000000	39.000000
75%	46.000000	49.000000
max	274.000000	280.000000

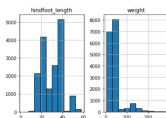
Estadística por géneros

Genero Masculino:

- Se encuentra la correlación lineal entre las propiedades, para el sexo masculino

correlación lineal:0,701169

- Podemos ver la frecuencia en los datos de estas dos variables solo para el genero masculino con un histograma



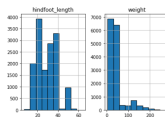
- Es de utilidad tener una visión gráfica de la frecuencia de las propiedades de los encuestados, esto es, un histograma
- Estas estadísticas nos dan una idea de las características medias por sexo y su correlación.

Genero Femenino:

- Se encuentra la correlación lineal entre las propiedades

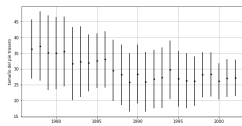
correlación lineal:0,665526

- Para el sexo femenino tenemos el siguiente histograma de las propiedades

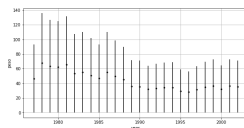


Estadística por tiempo

- Valores medios con la desviación estándar del tamaño del pie trasero en función de los años



- Valores medios del peso trasero en función de los años



Se calcularon las correlaciones lineales entre el tiempo y el tamaño del pie para los días(C_1), años(C_2) y meses(C_3)

$$C_1 = -0,001469$$

$$C_2 = -0,01371$$

$$C_3 = -0,276595$$

y Las correlaciones entre el tiempo y el peso, para días(C'_1), meses(C'_2) y años(C'_3).

$$C'_1 = -0,008636$$

$$C'_2 = -0,00297$$

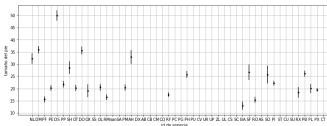
$$C'_3 = -0,276595$$

Estadística por especie

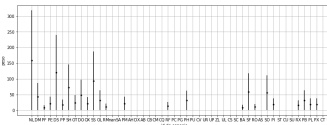
- ❶ Se obtuvo una tabla con los valores medios asociados al peso y al tamaño del pie trasero de cada especie

	NL	DM	PF	PE	DS	PP	SH	OT	DO	OX	...
pie	32.294227	35.982351	15.583389	20.195545	49.948874	21.751569	28.549618	20.267415	35.607551	19.125	...
peso	159.245660	43.157864	7.923127	21.586508	120.130546	17.173942	73.148936	24.230556	48.870523	21.000	...

- ❷ Se gráfica el comportamiento por especie del tamaño del pie trasero, con la desviación estándar correspondiente



- ❸ Se gráfica el comportamiento por especie del peso, con la desviación estándar correspondiente



- ❹ Estas gráficas nos dan una idea de las características medias de cada una de las especies y que tan bien caracterizan la especie.

Discusión

- Se obtuvo un valor dominante en la longitud del pie trasero para el genero masculino.
- Se observo una clara correlación entre el peso y la longitud del pie trasero, para ambos sexos.
- Se observo que la medición del la longitud del pie trasero es una buena medida para caracterizar las especies, con poca desviación estándar, al contrario que el peso.
- Se obtuvo un baja correlación entre el tiempo y las características de los encuestados, mas sin embargo se observa una clara disminución en el valor medio del tamaño del pie trasero .

Conclusiones

- En la practica se evidencio la facilidad de trabajar con grandes cantidades de datos mediante python, haciendo uso de la librería pandas.
- Se obtuvieron algunos análisis estadísticos, evidenciando la facilidad a la hora de manipular y tratar los datos.