CONTRASTE O PRUEBA DE HIPÓTESIS E INTRODUCCIÓN AL ANÁLISIS DE REGRESIÓN LINEAL O AJUSTE DE MÍNIMOS CUADRADOS. NOTAS PARA DOCTORANDOS

Contrast or hypothesis testing and introduction to linear regression analysis or least-squares fitting. Notes for phd students

C. ROCA-FERNÁNDEZ[1], M. MULLOR[2]

Recibido: 25 de septiembre de 2023. Aceptado: 25 de enero de 2024

DOI: http://dx.doi.org/10.21017/rimci.2024.v11.n21.a149

Resumen

Este artículo pretende mostrar una visión integradora sobre dos de las metodologías estadísticas ampliamente utilizadas en diferentes disciplinas académicas y científicas, a veces pudiendo ser complicado llegar a entenderlas sobre todo en los inicios de los estudios universitarios y de la carrera investigadora, tratándose de un artículo enfocado a su entendimiento teórico y aplicación. El contraste o prueba de hipótesis y el análisis de regresión lineal o ajuste de mínimos cuadrados son metodologías que se ponen en práctica en diferentes momentos del desarrollo de un trabajo universitario y científico, por tanto, es importante entender las bases que las fundamentan.

Palabras clave: Inferencia estadística; métodos paramétricos; métodos no paramétricos; regresión; ajuste de modelos.

Abstract

This article aims to show an integrative view of two of the statistical methodologies widely used in different academic and scientific disciplines, sometimes it being able to be complicated to understand, especially at the beginning of the university studies and the research career, it being an article focused on their theoretical understanding and application. The statistical hypothesis testing and the linear regression analysis or least-squares fitting are methodologies that are put into practice at different moments of the development of a university and scientific work, therefore, it is important to understand the bases that support them.

Keywords: Statistical inference; parametric methods; non-parametric methods; regression; fitting models.

I. Introducción

La estadística es una disciplina ampliamente utilizada, siendo a veces difícil su aplicación teórico-práctica en otras disciplinas que no desarrollen algunos de sus conceptos y metodologías, siendo importante su utilización en cualquier ámbito académico-científico. El contraste o prueba de hipótesis y el análisis de regresión lineal o ajuste de mínimos cuadrados son metodologías

1 Máster Universitario en Tecnologías de la Información Geográfica para la Ordenación del Territorio: SIG y Teledetección por la Universidad de Zaragoza, Grado en Geografía por la Universidad de Barcelona. PhD Researcher en Grupo de Investigación GRUMETS, Departamento de Geografía de la Universidad Autónoma de Barcelona. Tesis doctoral sobre las tendencias y los impactos de las sequías en la Península Ibérica a partir del uso de los SIG y de imágenes de teledetección satelitaria (2020–actualidad). ORCID: https://orcid.org/0000-0003-3639-7217 Correo electrónico: Catalina.Roca@uab.cat

2 Grado en Estadística Aplicada y Grado en Sociología por la Universidad Autónoma de Barcelona. Correo electrónico: Marti.Mullor@autonoma.cat

estadísticas que deben acompañar a cualquier estudio con base científica para poder obtener estimaciones fiables del análisis que se esté llevando a cabo. Como se verá más adelante, dependiendo del tipo de análisis será conveniente utilizar una, o la otra, o ambas.

En el artículo se desarrollan los conceptos relacionados con el contraste de hipótesis en base a supuestos de normalidad y habiendo comprobado la correlación entre las variables a analizar: la hipótesis nula (H0), la hipótesis alternativa (H1) y los estadísticos del contraste (Fig. 1)—el nivel de confianza (NC), la significación estadística (p-valor) y el intervalo de confianza (IC)—. En cuanto al análisis de regresión lineal, también se explica la base teórica que lo sustenta en relación con la estimación de los parámetros de la función de la recta de regresión lineal[y = ax + b], y del coeficiente de determinación[r^2]. Además, para cada caso se presenta un ejemplo realizado con el software R.

II. Una hipótesis es una afirmación sobre un parámetro partiendo de la hipótesis nula

La H0 indica que un parámetro de población (media, desviación estándar, varianza, …) es igual a un valor hipotético. Suele ser una afirmación inicial que se basa en análisis previos o en conocimiento especializado. Es lo que se cree como cierto y aceptado hasta el momento en el que se realiza el contraste o prueba de hipótesis, habitualmente que no existe asociación entre las variables o que ésta es explicable por el azar[1][2][3].

La H1 indica que un parámetro de población (media, desviación estándar, varianza, ...) es diferente del valor hipotético de la H0. Es lo que se podría pensar que es cierto y se espera probar que es cierto, y plantea que las variables están asociadas[1][2][3].

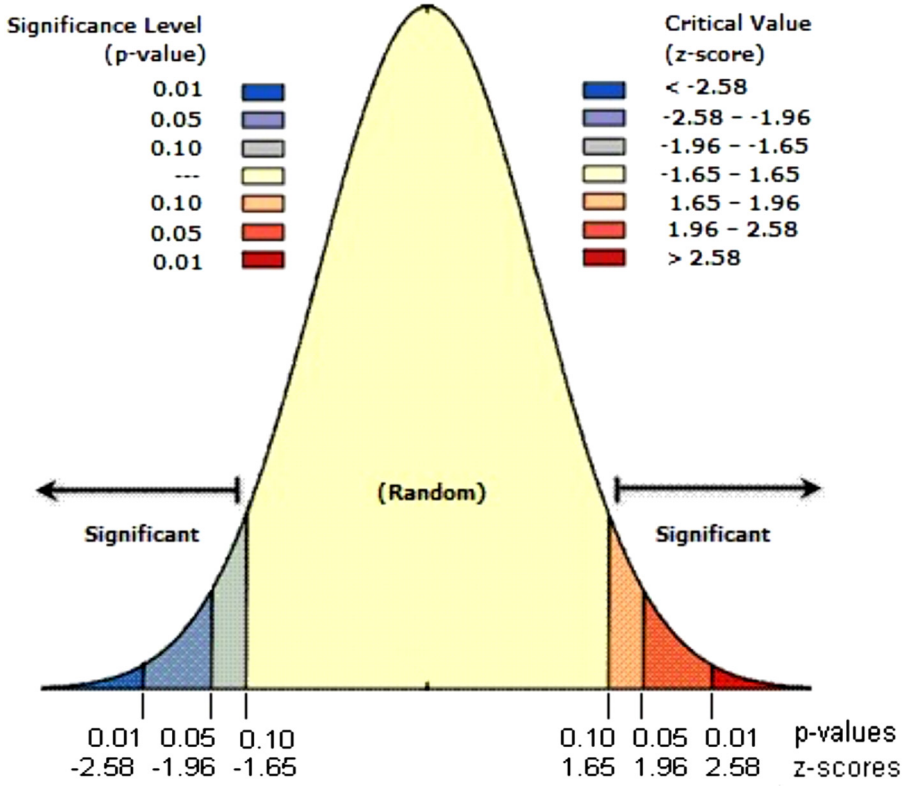

Fig. 1. Representación de los estadísticos del contraste para la inferencia estadística.[10]

Así pues, la H0 es un punto de partida para la investigación que no se rechaza, en la que no existe correlación y hay independencia entre las variables ($r = 0$), a menos que los datos de la muestra parezcan evidenciar que la H0 es falsa, demostrando entonces la H1, hipótesis en la que sí existe correlación y dependencia entre las variables ($r \neq 0$). Por tanto, en el inicio se asume la H0, pero se desea probar y demostrar la H1 para rechazar la H0 y aceptar la H1 (contraste o prueba de hipótesis)[1][2][3].

III. Un contraste o prueba de hipótesis utiliza los datos de la muestra para determinar si se puede rechazar la hipótesis nula y aceptar la hipótesis alternativa

Metodología de inferencia estadística para juzgar si una población estadística cumple con una propiedad supuesta compatible con lo observado en una muestra de dicha población, utilizada para la toma de decisiones. A partir de la muestra se investiga la aceptación o rechazo de una afirmación acerca de una o más características de la población, analizando la H0 y la H1 para conocer la probabilidad de que alguna de las hipótesis sea cierta, partiendo de que la H0 lo es. Pero rechazar la H0 no significa obligatoriamente que esta hipótesis no sea cierta, sino que es menos probable de serlo, ya que ninguna prueba de hipótesis es 100 % cierta porque es imposible trabajar con una población completa, siempre se trabaja con una muestra de dicha población basándonos en probabilidades de H0 y de H1 para aceptar o rechazar el parámetro a estimar. Y puesto que la prueba se basa en probabilidades, siempre existe la posibilidad de llegar a una conclusión incorrecta[1][2][3][4].

La zona de rechazo, o también zona crítica (Fig. 1), es la zona de rechazo de la H0, la que demuestra que los resultados son estadísticamente significativos y por tanto se puede aceptar la H1. Está situada en las colas de la distribución delimitadas por p-valor (α para la suma de las dos colas, $\alpha/2$ para cada cola), valor que indica el nivel de significación del contraste de hipótesis que se deberá marcar según un NC preestablecido ($1 - \alpha$), siendo estos dos estadísticos complementarios. Según el parámetro que estemos analizando, valores obtenidos en esta zona de rechazo determinarán más diferencias entre las variables (e.g. diferencias entre medias muestrales), o que las variables son dependientes y que existe correlación significativa entre ellas (e.g. test de correlación)[1][4][5].

Las principales herramientas utilizadas para hacer inferencia estadística son los NC, p-valor y los IC (Fig. 1)[4].

IV. Test de normalidad aplicado a las variables

Metodología estadística que se utiliza para contrastar la hipótesis de normalidad de las variables a analizar (H0). Es necesario conocer si las variables se ajustan o no a una distribución normal para aplicar un test de correlación u otro en la prueba de hipótesis, además de conocer la media y la desviación estándar de las mismas. Entre otros, el test de normalidad de Kolmogorov-Smirnov (KSI) es recomendable para más de 50 observaciones, sino se deberá utilizar el test de Shapiro-Wilk[4][6], y el resultado será p-valor:

- *Si p-valor es superior a 0.05:* se aceptará la H0 y la normalidad de los datos, y se deberá aplicar un test de correlación para variables que se ajusten a una distribución normal.

- *Si p-valor es inferior o igual a 0.05:* se aceptará la H1 y la no normalidad de los datos, y se deberá aplicar un test de correlación para variables que no se ajusten a una distribución normal. Lo más recomendable es normalizar las variables siempre y cuando sea posible.

V. Test de correlación según si son variables que se ajustan o no a una distribución normal de los datos

A. Test de correlación lineal de Pearson[r] entre las variables

Método paramétrico, ya que es necesario que al menos los valores de la variable independiente[x] se ajusten a una distribución normal. Únicamente se puede aplicar si las variables son cuantitativas[7][8][6].

B. Test de correlación por rangos de Kendall[t] y Spearman[rs] entre las variables

Métodos no paramétricos, ya que es recomendable el uso de estos test cuando no se pueda asumir que los datos se ajusten a una distribución normal y no sean normalizables. Se calculan aplicando una transformación a los valores originales de las variables, obteniendo el grado de asociación lineal de clasificaciones por rangos, entendiendo por rango de un valor como el lugar que ocupa dicho valor en el conjunto total de valores de la variable, suponiendo que los valores están ordenados de menor a mayor o viceversa. La diferencia entre ambos test es que el de Kendall[t] se utiliza para variables cualitativas de tipo ordinal, y el de Spearman[rs] para variables cualitativas de tipo ordinal cuando la determinación sea directa y para variables cuantitativas de tipo razón[7][4][8][9][6].

La interpretación del test de correlación de Spearman[rs] es la misma que la del coeficiente de correlación de Pearson[r], considerando el coeficiente de correlación un estimador muestral del coeficiente de correlación poblacional (valor poblacional[ρ])[4] en donde:

- *Si rs es igual a 0:* no existe correlación lineal entre las variables, se asume la H0 y la independencia entre ellas.

- *Si rs es diferente de 0:* sí existe correlación lineal entre las variables, se asume la H1 y la no independencia entre ellas, aunque mantendremos la H0 o nos decantaremos por la H1 si:

  - *p-valor es superior a 0.05 (para un NC del 95 %): correlación no significativa.* Mantendremos la H0, aunque exista correlación entre las variables y nos decantaremos por la no correlación, asumiendo automáticamente r = 0 por ser correlación no significativa y mantendremos la independencia entre ellas.

  - *p-valor es inferior o igual a 0.05 (para un NC del 95 %): correlación significativa.* Nos decantaremos por la H1, porque la H0 no se puede mantener según la muestra, mantendremos la correlación entre las variables, ahora significativa (fiabilidad en los datos), y que existe dependencia entre ellas.

Pero antes se deberá comprobar la correlación entre las variables, ya que no tiene sentido hacer el contraste o prueba de hipótesis entre variables en las que no existe correlación alguna. Una vez calculado el coeficiente de correlación, será necesario calcular p-valor, ya que este estadístico indica si existe correlación significativa o no entre las variables, si me puedo fiar de los datos o no, entendiendo por significativa como aquella afirmación específica respecto a qué tan probable es que algo se deba al azar[4][9].

VI. Los niveles de confianza

Probabilidad expresada en porcentaje, de que el parámetro a estimar a través de una distribución muestral se encuentre en el NC preestablecido, marcando p-valor en la región de rechazo de la distribución (Fig. 1)[10][1][5]. Por ejemplo:

- *Un NC menor (80%):* aunque sería una estimación más precisa, significa menor probabilidad de acierto (80 %) y mayor probabilidad de equivocarnos (20%). Entonces, p-valor se marcaría en 0.2, siendo ésta la región de rechazo de la H0 del 20 % para un NC del 80 %. El resultado de p-valor tendrá que encontrarse en esa región de rechazo, siendo inferior o igual a 0.2 para que el resultado sea estadísticamente significativo. De este modo se podrá rechazar la H0 y aceptar la H1, sino se mantendrá la H0.

- *Un NC mayor (95%):* aunque sería una estimación más segura pero más imprecisa, significa mayor probabilidad de acierto (95 %) y menor probabilidad de equivocarnos (5 %). Entonces, p-valor se marcaría en 0.05, siendo ésta la región de rechazo de la H0 del 5% para un NC del 95 %. El resultado de p-valor tendrá que encontrarse en esa región de rechazo, siendo inferior o igual a 0.05 para que el resultado sea estadísticamente significativo. De este modo se podrá rechazar la H0 y aceptar la H1, sino se mantendrá la H0.

  A cada NC le corresponde un p-valor (e.g. NC = 95%, p-valor = 5%), y cada NC se sitúa

entre unos límites, las desviaciones estándar o valores Z (Fig. 1), valores calculados a partir de aplicar una distribución inversa normal estándar a la probabilidad de error de p-valor, que se utilizarán para estimar el rango o los IC entre los que se encontrará el valor r[10][1][11][4][12].

Por ejemplo:

- *Para un NC del 80 %:* los datos estarán comprendidos entre ± 1.282 desviaciones estándar en torno a la media, o que el 80 % de la varianza se encuentra entre ± 1.282 desviaciones estándar en torno a la media.

- *Para un NC del 95 %:* los datos estarán comprendidos entre ± 1.96 desviaciones estándar en torno a la media, o que el 95 % de la varianza se encuentra entre ± 1.96 desviaciones estándar en torno a la media.

VII. La significación estadística (p-valor)

Valor preseleccionado que nos indica el límite entre rechazar o no la H0 (Fig. 1). Por convenio se fija un NC del 95 % y p-valor en la región de rechazo del 5 % o 0.05, o también un NC del 99 % y p-valor en la región de rechazo del 1 % o 0.01[1][13][14][15][16][17]. Por ejemplo, tomando un NC del 95 %, tendremos dos opciones posibles para p-valor:

- *Si p-valor es superior a 0.05, se mantendrá la H0:* el resultado no permite afirmar con el grado de confianza exigida (p-valor <= 0.05 para un NC del 95 %) que la H1 sea cierta, ya que p-valor observado es superior a p-valor preseleccionado, y el resultado será estadísticamente no significativo.

- *Si p-valor es inferior o igual a 0.05, se rechazará la H0 y se aceptará la H1 con un grado de acierto del 95 % (NC) y una probabilidad de error inferior o igual al 5 %:* el resultado permite afirmar con el grado de confianza exigida (p-valor <= 0.05 para un NC del 95 %) que la H1 es cierta, ya que p-valor observado es inferior o igual a p-valor preseleccionado, y el resultado será estadísticamente significativo (fiabilidad en los datos).

VIII. Los intervalos de confianza

Al igual que p-valor, los IC indican si rechazar o no la H0, además de una manera de estimar, con alta probabilidad, un rango de valores derivado de los estadísticos de la muestra en el que se encuentra el valor ρ, indicando que el valor estimado a partir de la muestra se encuentra en ese determinado IC con un cierto nivel de certeza (NC), con lo que NC más bajos resultarán en IC más pequeños (menor certeza pero mayor precisión en la estimación del parámetro que estemos analizando), y NC más altos resultarán en IC más amplios (mayor certeza pero menor precisión en la estimación)[18][1][5]. Habitualmente se marcan por consenso, en un NC del 95 % en base a supuestos de normalidad, lo que significa que, con un grado de acierto del 95 %, el valor ρ podrá tomar un valor entre el límite inferior y el límite superior del IC resultado. Pero rangos entre el 90 % y el 99 % son comúnmente utilizados en la literatura científica[4][19].

También se debe tener en cuenta que mientras mayor sea el tamaño de la muestra, menor será la variabilidad de los IC y los estimadores serán más precisos[18][1][13][15].

A. Rechazando o no la hipótesis nula

Los IC permiten excluir un valor crítico (0 ó 1) que indique la falta de asociación de la variable con el IC, lo que significa que, si los hallazgos son estadísticamente significativos, es porque el intervalo no contiene entre sus límites inferior y superior el valor crítico. En caso de contenerlo, el parámetro analizado será estadísticamente no significativo[16][12][9][5]. Cuando lo expresado sea una diferencia entre dos variables (e.g. al estimar el IC de medias ρ), el valor 0 será el valor crítico y representaría el punto en el que el evento es igualmente probable en ambos grupos y deberá mantenerse la H0[4]. Lo mismo pasaría en la estimación de los coeficientes de correlación, siendo su intervalo de -1 a +1, en donde el valor 0 sería el valor crítico, representando la no significación estadística y la independencia entre las variables[4]. Si se

trata de un indicador cuya fórmula es un cociente (e.g. al estimar el IC del índice estandarizado de mortalidad), un valor de 1 indicaría que la frecuencia de un determinado evento fue igualmente presentada tanto en el grupo expuesto como en el que no, por lo que 1 es el valor crítico y deberá mantenerse la H0[12].

B. Estimando el intervalo de confianza

Recordando que cada NC se sitúa entre sus desviaciones estándar o valores Z (Fig. 1), valores que se utilizarán para estimar el IC en el que se encontrará el valor ρ, la ecuación para calcular los IC será diferente según el parámetro que se esté analizando (media ρ, índice estandarizado de mortalidad, coeficiente de correlación, pendiente de una recta de regresión lineal, etc.). Habitualmente se utilizarán los valores Z, pero éstos podrán ser sustituidos por algún otro valor obtenido a partir de una función de distribución de probabilidad que permita probar hipótesis sobre el valor ρ.

Los IC para estimar coeficientes de correlación ρ entre las variables se pueden obtener de varias maneras. Lo más habitual es hacerlo a partir de la transformación Z de Fisher aplicada al coeficiente de correlación muestral (Z_{rxy})[20][21][4][22][23]. Otra manera de calcularlos es utilizando la t-Student, función estadística que también se aplica en la construcción de p-valor, calculando el valor T en lugar del valor Z asociado al NC que necesitemos, y considerando los grados de libertad de la muestra[20][7][4]; aunque si son muestras inferiores a 30 observaciones, se necesitará el cálculo de la t-Student[4]. Los grados de libertad son n_observaciones que pueden variar libremente al estimar parámetros estadísticos, en donde el último valor de la muestra no puede variar, porque es el número de relaciones establecidas[1][20][7]. Por tanto, los grados de libertad para una variable son n – 1, en caso de estar comparando dos variables serán n – 2, y así sucesivamente.

c. Ejemplos

Un estudio detectó mayor mortalidad en el postoperatorio por fibrilación auricular con un riesgo relativo a la mortalidad de 3, con un NC del 95% e IC 2 – 4, representando que la presencia de la arritmia en la muestra triplicó la probabilidad de morir en relación con quienes no la tuvieran. Por tanto, la estimación del IC indica que, con un grado de acierto del 95 %, el valor ρ del riesgo relativo a la mortalidad podrá tomar un valor entre 2 y 4. Además, como el rango del intervalo no contiene el valor 1, se puede afirmar que el resultado es estadísticamente significativo[24].

Tenemos una muestra de 100 neumáticos en la que la duración en km se ajusta a una distribución normal N(48000,3000), y necesitamos saber el rango de valores en el que se encontrará la duración media ρ para un NC del 80 %, recordando que se trabaja con variables muestrales, porque es imposible trabajar con una población completa[25]:

- Calcularemos el error estándar de la muestra, que no es más que la desviación estándar de la distribución muestral. Para ello se dividirá la desviación estándar (3000) entre la raíz cuadrada del tamaño de la muestra (100), dando como resultado 300.

- Calcularemos el margen de error, multiplicando el valor Z del 80 % (± 1.282) por 300, dando como resultado ± 384.60.

- Finalmente, el resultado anterior (± 384.60) se sumará y se restará a la media muestral (48000), dando como resultado un IC entre 47615.40 km y 48384.60 km.

- Entonces, aunque la duración media muestral en km de estos neumáticos es de 48000, con un grado de acierto del 80 % se puede afirmar que la duración media ρpodrá tomar un valor entre 47615.40 km (duración media mínima) y 48384.60 km (duración media máxima).

IX. La significación estadística (p-valor) y los intervalos de confianza

Son técnicas de inferencia estadística que contribuyen a la precisión de cualquier investigación. Ambas están estrechamente relacionadas, siendo raro que, para una prueba de hipótesis, un p-valor ofrezca un resultado significativo y un IC no. En cualquier caso, se mantendrá la H0 si los resultados son estadísticamente no significativos, o se

rechazará la H0 y se aceptará la H1 cuando sean significativos.

p-valor y los IC se complementan, pero no se sustituyen, debiendo ser ambas técnicas consideradas en el análisis para la toma de decisiones de nuestra investigación. Calculando p-valor se puede aceptar o rechazar un nivel de significación, que será la probabilidad de cometer errores y de que el resultado se deba al azar. La información contenida en un IC es más detallada que la contenida en p-valor, permitiendo estimar con un grado de acierto el rango en el que se puede encontrar el valor ρ analizado a partir de la muestra[18][1][13][15]. Por ejemplo, calculando el coeficiente de correlación r y p-valor, obtenemos una estimación puntual del grado de asociación entre las variables y del nivel de significación estadística. Pero calculando el IC, además podremos saber con un cierto nivel de certeza (NC) entre qué valores podrá variar el coeficiente de correlación ρ.

X. Errores de tipo I y II, y probabilidades de cometerlos

A. Error de tipo I o error alfa que se comete al rechazar la hipótesis nula

Error que se comete cuando se rechaza la H0 siendo ésta verdadera, tratándose de un falso positivo. La probabilidad de cometerlo es α, posibilidad de llegar a una conclusión incorrecta, y viene determinado por el nivel de significación p-valor (α), recordando que es complementario al NC (1 – α), en donde la decisión correcta de no rechazar la H0 siendo ésta verdadera sería la probabilidad de 1 – α. Por tanto, α es el nivel de significación de la prueba estadística y la máxima probabilidad de cometer el error de tipo I[1][15][17].

Si en la prueba de hipótesis p-valor es inferior o igual al valor preseleccionado (e.g. 0.05 para un NC del 95 %), aceptando entonces la H1, estaremos dispuestos a asumir una probabilidad de equivocarnos del 5 % al rechazar la H0. Pero si p-valor es superior al valor preseleccionado, entonces se mantendrá la H0 y no existirá ninguna probabilidad de cometer el error de tipo I[26][17][27].

Para reducir este tipo de error deberemos aumentar el NC y reducir p-valor. Pero se debe tener en cuenta que, si se reduce p-valor, aunque las probabilidades de error serán menores, también lo serán las probabilidades de detectar resultados estadísticamente significativos si existen realmente, debido a que la zona de rechazo de la H0 es menor, por lo que se estaría cometiendo un error de tipo II, manteniendo la H0 cuando en realidad es falsa. Por tanto, cuanto menor sea p-valor, menor será el riesgo de cometer el error de tipo I y menor será el riesgo de establecer hipótesis falsas en la población de estudio, pero también habrá menos probabilidades de encontrar posibles resultados estadísticamente significativos verdaderos[15][17][19].

Por convenio se fija este tipo de error inferior o igual al 5 % o 0.05 para un NC del 95 %, aunque también es común marcarlo en el 1 % para un NC del 99 %, sobre todo en ensayos clínicos[4][17][19].

B. Error de tipo II o error beta que se comete al mantener la hipótesis nula

Error que se comete cuando no se rechaza la H0 siendo esta falsa, tratándose de un falso negativo. La probabilidad de cometerlo es β, posibilidad de llegar a una conclusión incorrecta, que se marcará según un nivel de potencia prestablecido (1 – β), siendo estos dos estadísticos complementarios, en donde la decisión correcta de rechazar la H0 siendo esta falsa sería la probabilidad de 1 – β (zona de rechazo de la H0, la que demuestra que los resultados son estadísticamente significativos y por tanto se puede aceptar la H1). Así pues, 1 – β es el nivel de potencia de la prueba estadística, valor que indica el nivel de significación del contraste de hipótesis, y β es la máxima probabilidad de cometer el error de tipo II[1][15][17][19].

Los análisis estadísticos más comunes están diseñados para controlar el error de tipo I, por lo que el cálculo de la potencia de una prueba estadística también se puede entender como una medida de confianza del análisis que hemos realizado, principalmente cuando el resultado del contraste de hipótesis con p-valor ha sido no significativo[26][27]. Por convenio se fija este tipo de error inferior o igual al 20 % o 0.2, debiendo ser el nivel ideal de la potencia estadística igual o superior al 80 % o 0.8[27][19].

Si en la prueba de hipótesis, β es superior al valor preseleccionado (e.g. 0.2 para un nivel de

potencia del 80 %), aceptando entonces la H0, estaremos dispuestos a asumir una probabilidad de equivocarnos de más del 20 % al rechazar la H1. Pero si β es inferior o igual al valor preseleccionado, entonces se aceptará la H1 y no existirá ninguna probabilidad de cometer el error de tipo II[27][19].

Para reducir este tipo de error se deberá aumentar el tamaño de la muestra, habiendo de ser lo suficientemente grande o potente como para detectar resultados estadísticamente significativos verdaderos, ya que cuanto mayor sea ésta, menor será el error de tipo II. También se podría considerar aumentar el valor predeterminado de p-valor, dependiendo del NC marcado previamente y del tamaño de la muestra, si ésta es lo suficientemente grande o no, aunque no es muy recomendable, porque con un p-valor mayor las probabilidades de error serán mayores, y también lo serán las probabilidades de detectar resultados estadísticamente significativos aunque no existan realmente, debido a que la zona de rechazo de la H0 será mayor, y se estaría cometiendo un error de tipo I, aceptando la H1 cuando en realidad es falsa. Por tanto, cuanto menor sea β, menor será el riesgo de cometer el error de tipo II y menor será el riesgo de establecer hipótesis falsas en la población de estudio, pero también habrá más probabilidades de encontrar posibles resultados estadísticamente significativos falsos[15][17][19].

C. Errores de tipo I y II, y otros factores determinantes

Existe cierto compromiso entre ambos tipos de error y los riesgos de cometerlos están inversamente relacionados, por lo que cuanto menor sea alfa, mayor será beta y viceversa. Por tanto, alfa y la potencia estadística están directamente relacionadas[27][19].

En relación con el tamaño muestral, si es reducido, disminuirán las posibilidades de observar relaciones con otras variables. De todos modos, podría parecer que una muestra es mejor cuanto más grande sea, ya que cuanto más grande es, las estimaciones son más fiables y con menor riesgo de error. Pero esto no tiene por qué ser cierto. Si los datos de la muestra no están bien recogidos, la muestra será incorrecta sea ésta más pequeña o más grande; aunque por lo general, más datos proporcionarán más información y una mejor oportunidad estadística de distinguir los valores[1][4][27][19].

Existen también otros dos factores que influyen en la comparación de variables y por tanto en la probabilidad de cometer errores: la diferencia entre las medias muestrales y la dispersión de las variables[4]. Tendremos más probabilidades de que los resultados sean estadísticamente significativos si tenemos una mayor diferencia entre las medias muestrales de las variables y/o si tenemos una menor dispersión de las variables (varianza y dispersión estándar más próximas a la media muestral)[1][4][27]:

- *En el caso de las medias muestrales:* una diferencia entre ellas de 0 sería el valor ideal para mantener la H0, ya que no se podría mantener que las medias ρ fuesen distintas si las medias muestrales son iguales.

- *En el caso de la dispersión de las variables – variabilidad en que los valores de la variable tienden a extenderse alrededor de la media – :* cuanto menor, más probabilidades habrá que el estadístico de contraste caiga en la región de rechazo de la H0, porque sus límites se acercarán hacia el centro de la distribución muestral debido a la menor dispersión en torno a la media, y la región de rechazo será mayor.

XI. Contraste o prueba de hipótesis – Ejemplo con el software R

```
#1. Crear objeto en R:
    setwd("E:/R_Contrastes_Pruebas_Hipotesis")
    prueba<-read.csv("Contrastes_Pruebas_Hipotesis.csv",header=TRUE,sep=";",dec=".")
    prueba

#2. Instalar paquete para testar la normalidad de los datos:
    install.packages("nortest",dep=TRUE)
    library(nortest)
```

```
#3. Test de normalidad KSI aplicado a las variables:
    lillie.test(prueba$R_GLOBAL1_Avg) #de manera individual a la columna invocada
    apply(prueba,2,lillie.test) #de manera conjunta a todas las columnas del objeto (indicar '2' en
    la función)

#4. Media y desviación estándar (y varianza):
    mean(prueba$R_GLOBAL1_Avg)
    apply(prueba,2,mean)
    sd(prueba$R_GLOBAL1_Avg)
    apply(prueba,2,sd)
    #var(prueba$R_GLOBAL1_Avg)
    #apply(prueba,2,var)

#5. Correlación de Pearson (por defecto) entre las variables:
    cor(prueba)
    cor_1<-cor(prueba)
    cor(prueba$R_GLOBAL1_Avg,prueba$R_GLOBAL_SolarGis)
    cor_2<-cor(prueba$R_GLOBAL1_Avg,prueba$R_GLOBAL_SolarGis)
    #cor(prueba,method="spearman") #para aplicar la correlación de Spearman (indicar en la función)
    #cor(prueba,method="kendall") #para aplicar la correlación de Kendall (indicar en la función)

#6. Test de correlación de Pearson (por defecto) entre las variables:
    cor.test(prueba$R_GLOBAL1_Avg,prueba$R_GLOBAL_SolarGis)
    contrastH_1<-cor.test(prueba$R_GLOBAL1_Avg,prueba$R_GLOBAL_SolarGis)
    #cor.test(prueba$R_GLOBAL1_Avg,prueba$R_GLOBAL_SolarGis,method="spearman")
    #cor.test(prueba$R_GLOBAL1_Avg,prueba$R_GLOBAL_SolarGis,method="kendall")
```

a. Explicación de los resultados

- Tamaño de la muestra n_observaciones: 69 observaciones.
- Test de normalidad KSI aplicado a las variables: las variables se ajustan a una distribución normal de los datos.
- Coeficiente de correlación de Pearson entre las variables: r = 0.895.
- Test de correlación de Pearson entre las variables: por defecto, R define el NC en el 95 % y p-valor en el 5 % (0.05):
- p-valor < 0.05: rechazamos la H0 y aceptamos la H1, porque p-valor es inferior a 0.05.
- IC 0.836 – 0.934: rechazamos la H0 y aceptamos la H1, porque el IC no contiene en su interior el valor crítico 0.

Se puede afirmar que existe correlación lineal fuerte y directa (signo positivo) entre las dos variables muestrales (r = 0.895), con p-valor inferior a 0.05, e IC 0.836 (correlación mínima) – 0.954 (correlación máxima) para el coeficiente de correlación ρ. El resultado es estadísticamente significativo (fiabilidad en los datos), porque p-valor es inferior a 0.05 y el IC no contiene en su interior el valor crítico 0, lo cual permite rechazar la H0 y aceptar la H1 con un grado de acierto del 95 % y una probabilidad de cometer el error de tipo I inferior al 5 % (alfa o p-valor), afirmando que existe dependencia entre las variables. Se considera que existe una probabilidad muy pequeña de que la asociación encontrada entre las dos variables se deba al azar.

Si p-valor fuese superior a 0.05, o si el IC contuviera el valor crítico 0, se trataría de una correlación lineal no significativa: automáticamente se asumiría que no existe correlación lineal entre las variables (r = 0), debiendo mantener la H0.

XII. Análisis de regresión lineal o ajuste de mínimos cuadrados

Análisis que permite estimar la relación entre una variable dependiente[y] y una o más variables explicativas o independientes[x], además de generar una predicción o modelar la variable dependiente a partir de los valores conocidos y de sus relaciones con las variables independientes[28][7][29][6].

Se trata de un procedimiento paramétrico (normalidad de los datos) a partir del cual establecemos y cuantificamos la naturaleza de la relación existente entre las variables, que implica la estimación de los parámetros de la función de la recta de regresión lineal[y = ax + b], y del coeficiente de determinación[r^2][28] ;[30][31][6][32], en donde:

- *y:* es la variable explicada o dependiente.
- *x:* es la variable explicativa o independiente.
- *Los coeficientes[a] y[b]:* garantizan un ajuste de mínimos cuadrados, determinando que el cuadrado de las diferencias entre las variables (residuales) sea mínimo, y explicando la posible relación lineal entre la variable dependiente y la variable independiente, en donde:
  - *a:* es la pendiente de la recta de regresión, el coeficiente de la variable independiente, siendo el incremento de *y* por cada unidad que se incrementa *x*.
  - *b:* es la constante de la recta de regresión, el coeficiente interceptal o intercepto, allí donde la recta corta el eje *y*, siendo el valor que toma *y* cuando *x* es igual a 0.
- r^2*:* es la varianza expresada en porcentaje de la variable dependiente[y] que depende o que es explicada por la variable independiente[x], siendo un valor de 0 a 1, en donde:
  - $r^2 = 1$*:* mejor ajuste de mínimos cuadrados, en el que el cuadrado de las diferencias entre las variables (residuos) es 0 y el ajuste es perfecto.
  - *Por ejemplo, si* r^2 *es igual a 0.11:* el 11 % de la varianza de la variable dependiente[y] es controlada o explicada por la variable independiente[x].

Se recomienda el uso de r^2 ajustado por considerarse como la corrección del sesgo de r^2 sin ajustar, ya que a medida que se añaden variables independientes a un análisis de regresión lineal, r^2 sin ajustar tiende a aumentar, porque no tiene en cuenta el número de variables incluidas en el modelo, sobreestimando así la predicción. Aunque en la varianza y covarianza para el cálculo de r^2 sin ajustar ya se consideran n_observaciones, r^2 ajustado considera n_observaciones en la corrección. De este modo, r^2 ajustado considera el efecto acumulativo de la cantidad de observaciones y de las variables independientes contempladas en el análisis de regresión[33][29][6].

La metodología del ajuste entre las variables será: testar la normalidad de los datos, generar el diagrama de dispersión, obtener la función de la recta de regresión lineal[y = ax + b], evaluar el ajuste con los coeficientes de regresión[a,b], evaluar la bondad del ajuste con el coeficiente de determinación[r^2], y realizar el contraste o prueba de hipótesis (H0 vs H1) como test independiente entre las variables.

La dispersión deberá analizarse mediante el diagrama de dispersión o el análisis de covarianza —covariabilidad que existe entre los datos de dos variables—, que parte de los conceptos de regresión lineal para el ensayo de la H0 en donde las dos medias muestrales sean iguales[34]. En líneas generales, si las observaciones de ambas variables están más alejadas entre sí, existirá una mayor distribución entre ellas y una mayor dispersión de los datos, un peor ajuste de mínimos cuadrados y una menor correlación lineal. Pero si las observaciones de ambas variables están más próximas entre sí, existirá una menor distribución entre ellas y una menor dispersión de los datos, un mejor ajuste de mínimos cuadrados y una mayor correlación lineal[30][29][31][8][6].

XIII. Análisis de regresión lineal o ajuste de mínimos cuadrados – Ejemplo con el software R

```
#7. Estimar la relación entre la variable dependiente[y] y la variable independiente[x]:
    plot(prueba$R_GLOBAL1_Avg,prueba$R_GLOBAL_SolarGis) #representación gráfica de las variables
    lm(R_GLOBAL_SolarGis~R_GLOBAL1_Avg,prueba) #devuelve los coeficientes de regresión[a,b]
    del ajuste para la función de la recta de regresión lineal[y = ax + b]
    #lm(prueba$R_GLOBAL_SolarGis~prueba$R_GLOBAL1_Avg)
    coef_re_ab<-lm(R_GLOBAL_SolarGis~R_GLOBAL1_Avg,prueba)
```

```
    #lm(R_GLOBAL_SolarGis~R_GLOBAL1_Avg+n_VarsIndepes,prueba) #para añadir más variables
    independientes que expliquen la posible relación
    #lm(R_GLOBAL_SolarGis~.,prueba) #con el '.' en la función, lm() considera todas las
    variables independientes
    summary(coef_re_ab) #devuelve un resumen estadístico, en este caso del ajuste
    est_ajuste<-summary(coef_re_ab)
    abline(coef_re_ab) #representación gráfica del modelo de regresión lineal
    #la función lm() se puede estructurar de diferentes maneras
    #además, las funciones relacionadas se pueden crear invocando los objetos ya creados con
    la función lm(), o invocando lm() y sus parámetros
    #ATENCIÓN ==> en la función lm(), antes del símbolo ~ se pone la variable dependiente[y],
    después del símbolo se ponen la/s independiente/s[x]

#8. Predicción de la variable dependiente[y] a partir de los valores conocidos y de sus
    relaciones con la variable independiente[x]:
    predict(coef_re_ab,prueba)
    v_depe_new<-predict(coef_re_ab,prueba)

#9. Modificar la nueva matriz para facilitar la comparación de la predicción con las variables:
    v_depe_new<-cbind(prueba$R_GLOBAL1_Avg,prueba$R_GLOBAL_SolarGis,v_depe_new)
    colnames(v_depe_new)[1:2]<-c("R_GLOBAL1_Avg","R_GLOBAL_SolarGis")
    v_depe_new<-as.data.frame(v_depe_new)
    class(v_depe_new)
    v_depe_new
```

XIV. Conclusiones

El contraste o prueba de hipótesis, metodología de inferencia estadística, determina si dos variables están relacionadas para poder rechazar la H0 y aceptar la H1, o si no están relacionadas mantener la H0, estableciendo la dependencia o la independencia entre las variables. Además, permite estimar el grado de fuerza de la correlación lineal entre las dos variables con un cierto nivel de confianza o grado de acierto, así como la probabilidad de cometer errores en la estimación.

El análisis de regresión permite estimar el grado de fuerza de la relación lineal entre una variable dependiente y una o más independientes, pudiendo, además, predecir el valor de la variable dependiente a partir de los valores conocidos y de sus relaciones con varias variables explicativas, tratándose de una metodología ampliamente utilizada en los procesos de interpolación. Como por ejemplo cuando es necesario estimar las temperaturas —variable dependiente estimada— a partir de unas coordenadas XY para la cuales las temperaturas son conocidas (puntos verdad-terreno) —variable dependiente observada— y de sus relaciones con diferentes factores condicionantes que afectan a las variables climáticas, pudiendo ser las elevaciones, la distancia al mar, la latitud, etc. —variables independientes observadas—.

Se recomienda el contraste o prueba de hipótesis para correlacionar dos variables y determinar la H0 o la H1, y el análisis de regresión lineal para relacionar una variable dependiente con dos o más variables independientes, y/o para predecir los valores de la variable dependiente a partir de los valores conocidos y de sus relaciones con las variables independientes. Además, en el proceso del análisis de regresión es recomendable realizar el contraste o prueba de hipótesis (H0 vs H1) como test independiente entre las variables, pudiendo emplearse tanto en la estimación de las relaciones iniciales como para conocer las magnitudes de correlación lineal entre la variable dependiente estimada y la observada y las variables independientes. En el proceso del análisis de regresión también se recomienda calcular los estadísticos de error entre la variable dependiente estimada y la observada, ya que cuanto más se acerque el error cuadrático medio (RMSE)

a 0 la predicción de la variable dependiente tendrá una precisión más perfecta.

Referencias

[1] Á. Carreño, "Pruebas de contraste de hipótesis. Estimación puntual y por intervalos". Seden, 10 (10), pp.133-148. 2010. https://www.revistaseden.org/files/10-CAP%2010.pdf

[2] J. Dagnino, "Inferencia estadística: Pruebas de hipótesis". Rev. Chil. Anest., 43(2), pp.125-128. 2014. https://dx.doi.org/10.25237/revchilanestv43n02.10

[3] I. Leenen, "La prueba de la hipótesis nula y sus alternativas: Revisión de algunas críticas y su relevancia para las ciencias médicas". Inv. Ed. Med., Elsev., 1(4), pp.225-234. 2012. https://www.redalyc.org/articulo.oa?id=349736306010

[4] J. Llopis-Pérez, La estadística: Una orquesta hecha instrumento. Curso de estadística. 2021. https://jllopisperez.com

[5] G. Yáñez, R. Behar, "Interpretaciones erradas del nivel de confianza en los intervalos de confianza y algunas explicaciones plausibles". Seiem. 2009. https://www.seiem.es/docs/comunicaciones/GruposXIII/depc/Yanez_Behar_R.pdf

[6] M. E. Szretter-Noste, Apunte de Regresión Lineal [apuntes académicos]. Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales. 2017. http://mate.dm.uba.ar/~meszre/apunte_regresion_lineal_szretter.pdf

[7] S. Gómez-Biedma, M. Vivó, E. Soria, "Pruebas de significación en bioestadística". Rev. Diagn. Biol., Scielo, 4. 2001. https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0034-79732001000400008

[8] M. Rodrigues, *Estadística. Análisis estadístico. Estadística descriptiva e inferencia estadística. Contrastes de hipótesis. Regresión, modelos Anova. Nociones de análisis multivariante. Programas en uso*[apuntes académicos]. Moodle Universidad de Zaragoza. Análisis de la información geográfica: SIG. Máster Universitario en Tecnologías de la Información Geográfica para la Ordenación del Territorio: Sistemas de Información Geográfica y Teledetección (Universidad de Zaragoza). 2017.

[9] M. Rodrigues, *Programación para el análisis espacial: Scripting, Python y R*[apuntes académicos]. Moodle Universidad de Zaragoza. Análisis de la información geográfica: SIG. Máster Universitario en Tecnologías de la Información Geográfica para la Ordenación del Territorio: Sistemas de Información Geográfica y Teledetección (Universidad de Zaragoza). 2017.

[10] ArcMap.¿Qué es una puntuación Z? ¿Qué es un valor P? ESRI. 2021. https://desktop.arcgis.com/es/arcmap/10.3/tools/spatial-statistics-toolbox/what-is-a-z-score-what-is-a-p-value.htm

[11] A. J. González-Pareja, Cálculo de valores críticos [archivo de vídeo]. 2017. https://www.youtube.com/watch?v=VAqOGIz3ApA

[12] J. Recaño-Valverde, *Análisis demográfico*[apuntes académicos]. Campus Virtual Universitat Autònoma de Barcelona. Grau en Geografia, Medi Ambient i Planificació Territorial (Universitat Autònoma de Barcelona). 2020.

[13] M. L. Clark, "Los valores P y los intervalos de confianza: ¿En qué confiar?". Pan. Am. J. Public Health, 15(5), pp.293-296. 2004. https://scielosp.org/pdf/rpsp/2004.v15n5/293-296/es

[14] Z. N. Kain, J. MacLaren, "Valor de P inferior a 0.05: ¿Qué significa en realidad?". Pediatrics, Elsev., 63(3), pp.118-120. 2007. https://www.elsevier.es/es-revista-pediatrics-10-articulo-valor-p-inferior-005-que-13112660

[15] C. Manterola, V. Pineda, "El valor de P y la significación estadística. Aspectos generales y su valor en la práctica clínica". Rev. Chil. Cir., Scielo, 60(1), pp.86-89. 2008. https://dx.doi.org/10.4067/S0718-40262008000100018

[16] M. Molina-Arias, "Qué significa realmente el valor de P?". Rev. Pediatr. Aten. Primaria, Scielo, 19(76):377-81. 2017. https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1139-76322017000500014

[17] S. Pita-Fernández, S. Pértega-Díaz, "Significancia estadística y relevancia clínica". Fisterra, 8, pp.191-195. 2001. https://www.fisterra.com/fichas/interior.asp?idArbol=8&idTipoFicha=8&urlseo=significancia-estadistica-relevancia-clinica

[18] R. Candia, G. Caiozzi, "Intervalos de confianza". Rev. Med. Chile, Scielo, 133, pp.1111-1115. 2005. https://dx.doi.org/10.4067/S0034-98872005000 900017

[19] T. Salvador-Elías, I: Error tipo I y II. II: Nivel de significancia. III: Intervalo de confianza[apuntes académicos]. Universidad Católica Cecilio Acosta. Facultad de Medicina Veterinaria y Zootecnia. Bioestadística. 2018. https://eliasnutri.files.wordpress.com/2018/10/clase-6-error.pdf

[20] E. Cobo, K. Belchin, J. Cortés, J. A. González, P. Muñoz, Bioestadística para no estadísticos. Capítulo 8: Intervalos de confianza[apuntes académicos]. Universitat Politècnica de Catalunya. Barcelonatech. Departament d'Estadística i Investigació Operativa. 2014. https://upcommons.upc.edu/bitstream/handle/2117/186420/08_intervalos_de_confianza-5331.pdf

[21] J. C. Correa, L. V. Pacheco, "Estadística aplicada: Didáctica de la estadística y métodos estadísticos en problemas socioeconómicos. Comparación de intervalos de confianza para el coeficiente de correlación". VII Coloquio Regional de Estadística. XII Seminario de Estadística Aplicada IASI. III Escuela de Verano CEAES. Universidad Nacional de Colombia, Medellín, 20-23 de Julio de 2010. https://www.inec.gob.pa/IASI/docs/announcements/documentos/MemoriasComunicaciones/7%20Pacheco_CorreaComparacionIntervalosConfianzaCoeficienteCorrelacion.pdf

[22] L. V. Pacheco, J. C. Correa, "Comparación de intervalos de confianza para el coeficiente de correlación". Comunicaciones en estadística, Dialnet, 6(2), pp.157-174. 2013. https://www.google.es/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjh44W—Kb1AhXOxoUKHX fUCZkQFnoECAcQAQ&url=https%3A%2F%2Fdialnet.unirioja.es%2Fdescarga%2Farticulo%2F7393721.pdf&usg=AOvVaw3AJdev R2jdFR_3A7tnpT3l

[23] A. Sánchez-Bruno, Á. Borges del Rosal, "Transformación Z de Fisher para la determinación de intervalos de confianza del coeficiente de correlación de Pearson". Psicothema, 17(1), pp.148-153. 2005. http://www.psicothema.es/pdf/3079.pdf

[24] F. J. Candel, M. Matesanz, F. Cogolludo, I. Candel, C. Mora, T. Bescos, M. Martín, I. Vila i Costa, I., "Prevalencia de fibrilación auricular y factores relacionados en una población del centro de Madrid". An. Med. Interna, Scielo, 21(10), pp.477-482. 2004. https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0212-71992004001000003

[25] Píldoras matemáticas. 07: Intervalo de confianza[archivo de vídeo]. 2017. https://www.youtube.com/watch?v=2wugQGs1GNY

[26] J. Dagnino, "Inferencia estadística: Pruebas de hipótesis". Rev. Chil. Anest., 43(2), pp.125-128. 2014. https://dx.doi.org/10.25237/revchilanestv43n02.10

[27] J. Quesada, J. Figuerola, "Potencia de una prueba estadística: Aplicación e interpretación en ecología del comportamiento". Etologuía: Boletín de la Sociedad Española de Etología, Dialnet, 22, pp.19-37. 2010. http://www.ebd.csic.es/jordiplataforma/subidas/Etologuia2010.pdf

[28] J. Gabriel-Molina, M. F. Rodrigo, 6 – El modelo de regresión lineal[apuntes académicos]. Universitat de València. OpenCourseWare. 2021. http://ocw.uv.es/ciencias-sociales-y-juridicas/estadistica-i/tema_6.pdf

[29] A. Novales, Análisis de regresión[apuntes académicos]. Universidad Complutense de Madrid. Departamento de Economía Cuantitativa. 2010. https://www.ucm.es/data/cont/docs/518-2013-11-13-Analisis%20de%20Regresion.pdf

[30] R. Montorio-Llovería, *Creación y gestión de bases de datos de información geográfica: Georreferenciación de imágenes de teledetección*[apuntes académicos]. Moodle Universidad de Zaragoza. Obtención y organización de la información geográfica. Máster Universitario en Tecnologías de la Información Geográfica para la Ordenación del Territorio: Sistemas de Información Geográfica y Teledetección (Universidad de Zaragoza). 2017.

[31] F. Pérez-Cabello, *Tratamiento digital avanzado de imágenes de teledetección[*apuntes académicos]. Moodle Universidad de Zaragoza. Análisis de la información geográfica: Teledetección. Máster Universitario en Tecnologías de la Información Geográfica para la Ordenación del Territorio: Sistemas de Información Geográfica y Teledetección (Universidad de Zaragoza). 2019.

[32] Tests&Trials. Regresión y ajuste de la recta. 2021 https://www.testsandtrials.com/usos-siagro/regresion-y-ajuste-de-la-recta/

[33] E. Martínez, "Errores frecuentes en la interpretación del coeficiente de determinación lineal". Anuario jurídico y económico escurialense, Dialnet, 38, pp.315-331. 2005. https://www.google.es/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjJ8o_u5qb1AhU3A2MBHQk2DYYQFnoECAkQAQ&url=https%3A%2F%2Fdialnet.unirioja.es%2Fdescarga%2Farticulo%2F1143 023.pdf&usg=AOvVaw24UtKUpuQw9C07C WWv7fpk

[34] L. M. Molinero, Análisis de la covarianza. Asociación de la Sociedad Española de Hipertensión. Liga española para la lucha contra la hipertensión arterial. 2002. https://www.alceingenieria.net/bioestadistica/ancova.pdf