TEMA 1

MÓDULO:
TÉCNICAS AVANZADAS DE DATA MINING

INTRODUCCIÓN AL ANÁLISIS MULTIVARIANTE

FERRAN ARROYO

Licenciado en Empresariales y en Ciencias Actuariales y Financiaras por la UB. Máster Executive en Data Science por la MBIT School. Data Scientist.

Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

ÍNDICE

OBJETIVOS ESPECÍFICOS

- Diferenciar entre Análisis Bivariante y Análisis Multivariante.
- Reconocer las principales técnicas de Análisis Multivariante.
- Conocer todos los elementos que componen un Test de Hipótesis.
- Hacer un Test A/B.

1. INTRODUCCIÓN AL ANÁLISIS MULTIVARIANTE

La tarea de analizar los datos ha existido durante muchos años, pero practicarla era un proceso tremendamente lento hasta hace relativamente pocos años.

Los escasos recursos computacionales antes de los años 60 llevaron a la estadística a desarrollar este campo en su parte teórica.

Mientras que el coste de hacer cálculos en esa época era demasiado elevado (capital, recursos humanos, tiempo, poca fiabilidad), éstos únicamente se limitaban al análisis de datos para casos simples, mientras que el estudio teórico estaba décadas adelantado a su tiempo.

Durante años, los investigadores soñaron con cálculos avanzados para procesar fenómenos complejos, cuyos resultados casi podían predecir el futuro o describir lo que nosotros solos no podíamos, cálculos que, aunque inviables de poner en práctica, eran comprobables y prometedores. Algunos de estos cálculos procedían de modelos multivariantes.

Los investigadores siguieron desarrollando estas teorías con la perspectiva de que los costes asociados a la realización de esos monstruosos cálculos disminuyeran. Y entonces, aparecieron las computadoras. Varias técnicas de Análisis Multivariante se hicieron accesibles a las organizaciones y, más tarde, a todo el mundo con la llegada del ordenador personal.

En este apartado, aprenderás las principales diferencias que hay entre el análisis de datos univariante, el análisis de datos bivariante, y el análisis de datos multivariante, que es el que vamos a trabajar con profundidad a lo largo del tema.

1.1 DEFINICIÓN Y CONCEPTOS BÁSICOS

El Análisis Multivariante se puede definir como un conjunto de **métodos estadísticos que permiten ver el efecto de más de dos dimensiones de forma simultánea y pueden extraer conclusiones de ello.** Puede entenderse, también, como una expansión del análisis bivariante. A medida que los modelos multivariantes consideran más variables, pueden examinar fenómenos más complejos y encontrar patrones de datos que representen con más precisión el evento que queramos explicar.

Desenredar el nudo existente en las relaciones entre las variables, donde cada una puede estar correlacionada con muchas otras, es el objetivo principal del Análisis Multivariante. En muchos casos, cuanto más elevadas son estas interrelaciones, más difícil es la tarea de detectar relaciones significativas de datos, ya que todas las variables parecen influir en algo, cualquier estructura subyacente o cualquier causa-efecto queda diluida.

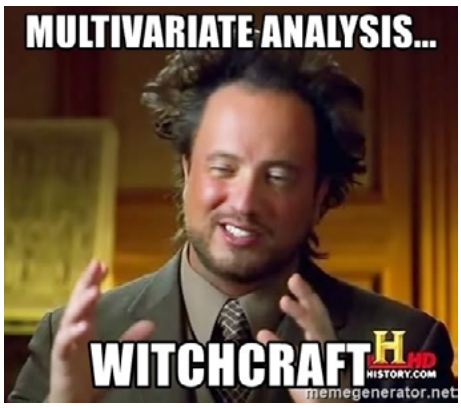
Resolver esta cuestión es una tarea importante del profesional del dato, que debe conocer los datos en profundidad, reducir el ruido y los sesgos que puedan existir en la medida de lo posible, así como otras imperfecciones que puedan existir en los datos que pueden ser tratadas con las técnicas de Análisis Multivariante.

Así pues, podríamos concluir que el **análisis de datos multivariante trata de encontrar patrones en**

un universo muy amplio de variables. Es posible que te preguntes:

- ¿Pero cuáles son esos patrones?
- ¿Qué técnicas estadísticas fueron diseñadas para encontrarlos?
- ¿Cómo es posible poder analizar muchísimas variables a la vez?

Es posible que hace algunos siglos, esta hubiese sido la respuesta que te habría dado más de uno:



Fuente: https://memegenerator.net/instance/58811164/science-guy-multivariate-analysis-witchcraft

A día de hoy, alguno te sigue dando esta respuesta. Sin embargo, actualmente sabemos que eso no es así (en principio). Demos un paso más y veamos las posibilidades que nos ofrece de este tipo de análisis.

1.2 PRINCIPALES TÉCNICAS DE ANÁLISIS MULTIVARIANTE

Existen dos categorías dentro del Análisis Multivariante, cada una de las cuales persigue un tipo diferente de relación existente entre las variables: dependencia e interdependencia.

- La **dependencia** se relaciona con situaciones de causa y efecto y trata de ver si un conjunto de variables puede describir o predecir los valores de otras. Un ejemplo clásico de técnica en este ámbito es la **regresión.**
- La **interdependencia** se refiere a la interrelación estructural y tiene por objeto comprender las pautas subyacentes de los datos. Un ejemplo clásico de técnica en este ámbito es el **clustering.**

Existen dos tipos de técnicas dentro del Análisis Multivariante. Las Técnicas de dependencia y las Técnicas de independencia:

- En las **técnicas de dependencia,** el modelo se alimenta con datos de entrada, especificando qué variables son independientes y cuáles son dependientes. Las variables dependientes son las que el modelo tratará de predecir o explicar. Las variables independientes son las que indican en qué medida afectan a la variable dependiente. Así pues, el **objetivo** de todas las técnicas de dependencia es **establecer una relación causa-efecto.** Las diferencias más notables entre ellas son el número de variables independientes que soportan y la naturaleza de las variables involucradas.
- Las técnicas de interdependencia no tienen por objeto resolver los problemas de causa y

efecto, sino comprender la estructura subyacente de los datos.

1.2.1 REGRESIÓN MÚLTIPLE

- Variable dependiente: una variable cuantitativa.
- Naturaleza de las variables independientes: cualquiera.

La regresión múltiple es una técnica de dependencia, **y es una buena opción cuando se estipula que, únicamente, existe solo una variable dependiente que sea cuantitativa.**

El resultado de aplicar una regresión múltiple es el grado de impacto que cada variable independiente tiene en la variable dependiente. Ese resultado también conduce a una función de estimación, en la que acepta valores para las variables independientes y devuelve el valor esperado para la variable dependiente.

Esta técnica podría utilizarse, por ejemplo, para predecir la rentabilidad de diferentes tiendas en función de sus características (como el número de vendedores, el número de horas de apertura, la renta media del barrio en que se encuentra...). Este análisis llevaría a una comprensión más profunda de la causa para que una tienda venda más que otra, lo que podría impulsar los cambios administrativos en las características de las tiendas hacia valores que den una mayor rentabilidad.

1.2.2 ANÁLISIS DISCRIMINANTE

- Variable dependiente: una variable cuantitativa.
- Naturaleza de las variables independientes: variables métricas.

Esta técnica de dependencia es muy similar a los famosos clasificadores que se utilizan en machine learning. **Es una opción interesante cuando sólo hay una variable dependiente que no es cuantitativa.** En esta técnica, a la variable independiente se la suele llamar “clase” o “etiqueta”. El objetivo es comprender la característica de los datos que pertenecen a cada clase.

Esta técnica podría utilizarse, por ejemplo, para crear un modelo que analizara las características de los fragmentos musicales (variables dependientes), mientras que cada pieza se asigna a un género musical (variable independiente). Si el modelo tiene éxito, será capaz de clasificar el género de nuevas canciones de forma correcta para los que nunca ha visto su género.

Una de las principales **limitaciones del análisis discriminante** es que **no es óptimo cuando algunas de las variables independientes no es cuantitativa.**

1.2.3 ÁNÁLISIS FACTORIAL.

- **Objetivo:** Entender qué variables están altamente correlacionadas con otras.

El análisis factorial es una técnica de interdependencia y tiene por objeto **reducir la dimensionalidad de los datos mediante la reducción del número de variables.** Con ello, se pueden detectar grupos de variables con alta correlación, utilizados como base para crear una nueva variable que pueda reemplazar a las variables originales con poca pérdida de información.

Un uso clásico para este tipo de técnica consiste en utilizarla como **paso previo al procesamiento para transformar los datos antes de utilizar otros modelos.**

Cuando los datos tienen demasiadas variables, el rendimiento de las técnicas multivariantes tiende a ser subóptimo, ya que los modelos son más difíciles de encontrar. Al utilizar el análisis factorial para condensar la información en un conjunto más pequeño de nuevas variables, los patrones se vuelven menos diluidos y más fáciles de analizar.

1.2.4 ANÁLISIS CLUSTER.

- **Objetivo:** Encontrar patrones en las observaciones.

El Análisis Cluster es una técnica de interdependencia y tiene por **objeto detectar grupos de observaciones que tienen valores similares en sus variables.** Esta técnica no es una exclusividad del Análisis Multivariante, ya que, incluso los datos a nivel unidimensional pueden ser agrupados. Sin embargo, esta técnica es más interesante utilizarla con más de una dimensión, ya que ofrece un output mucho más rico que en el caso de una sola dimensión.

Los clusters se suelen utilizar para entender la distribución de las variables que componen las observaciones. Al encontrar puntos situados a una distancia similar, se pueden realizar razonamientos acerca de la cercanía o la lejanía de los puntos para aumentar el conocimiento global del set de datos con el que estamos trabajando.

Esta técnica podría utilizarse, por ejemplo, para caracterizar un grupo de consumidores que tienen características similares y que compran ciertos productos muy a menudo (comúnmente denominado perfil). La empresa puede, entonces tomar medidas para que esos productos sean más accesibles para esos consumidores potenciales.

1.3 EJEMPLO PRÁCTICO.

Consideremos, como un pequeño ejemplo, el famoso modelo de regresión.

Imaginemos que tenemos dos variables y queremos ver la relación que existe entre ellas. Un método sencillo de verlo sería con una regresión bivariante. Podría utilizarse, por ejemplo, para ver cómo la altura de un nadador se correlaciona con su velocidad. Haciendo una regresión bivariante, el analista podría determinar que los nadadores más altos tienden a nadar más rápido. Aunque es correcto, sabemos que la altura no es lo único que influye en la velocidad, por lo que el modelo bivariado explica este fenómeno sólo en una pequeña parte.

Por el contrario, una regresión multivariante (también llamada regresión múltiple) podría tener en cuenta muchas más variables, como el peso, la edad, la ingesta de carbohidratos, la ingesta de proteínas, la cantidad de horas de entrenamiento, la cantidad de horas de descanso, y muchas otras. En teoría, cuanto mayor sea el número de variables, más precisa será la regresión que pueda representar el fenómeno de la natación, hasta el punto de poder determinar con precisión la velocidad de un nuevo nadador, según sus características, con un error mucho menor que en el caso de la regresión bivariante.

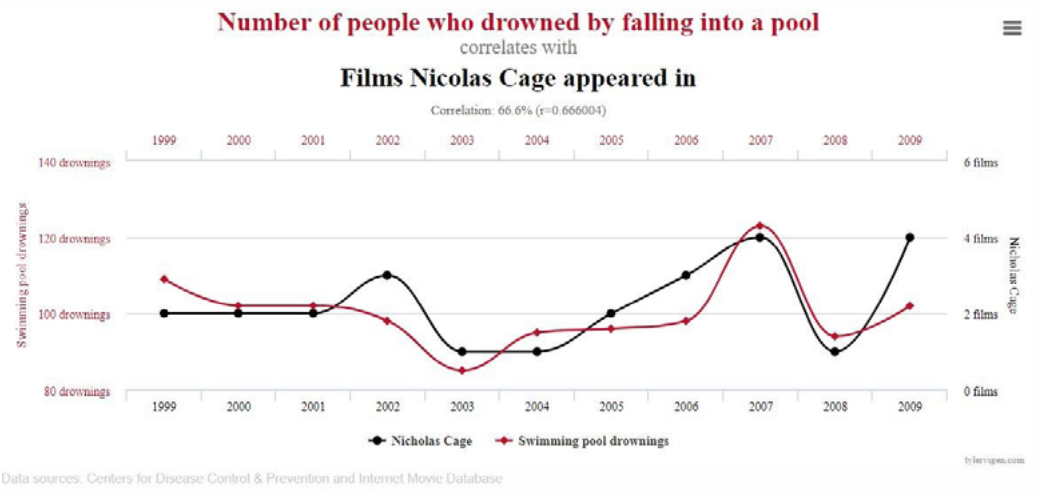
Sin embargo, hay que tener en cuenta que, aunque es muy importante disponer de muchas variables para poder lograr resultados robustos, los modelos hay que construirlos con cierta cautela y seleccionar,

adecuadamente, qué variables van a explicar el fenómeno de estudio. Incluir variables no significativas traerá muy pocos beneficios al modelo, por no decir ninguno, pero incluso podría desencadenar que el modelo explicara peor el fenómeno (es decir, que tuviera un error más elevado, que se equivocara más). En consecuencia, la realización de Análisis Multivariante es necesaria tomarla con cautela y no poner el piloto automático.

Continuando con el ejemplo anterior, imaginemos que se incluyó una variable que caracterice el número de series de terror que ve el nadador a la semana. ¿Se podría dar el caso que aquellos nadadores que ven menos series de terror a la semana tienen una velocidad superior y podríamos llegar a concluir que prohibir ver series de terror a los nadadores mejorará su velocidad? ¿Consideras que esto tiene sentido?

SABÍAS QUE...

Cuando encuentres una relación lineal, no creas que has encontrado el santo grial. Es muy importante que estudies y entiendas la relación que existe entre las variables, ya que el hecho de **que exista relación lineal entre las variables no implica que exista causalidad:**



Fuente: [Spurious Correlations: https://www.tylervigen.com/spurious-correlations](https://www.tylervigen.com/spurious-correlations)

2. INTRODUCCIÓN AL CONTRASTE DE HIPÓTESIS

El Contraste de Hipótesis se encuadra **dentro de la inferencia estadística** y ,también, se puede denominar test de hipótesis o prueba de significación.

Éste es un procedimiento para juzgar si una propiedad que se supone en una población estadística, es compatible con lo observado en una muestra de dicha población.

Mediante esta teoría, se aborda el problema estadístico considerando una hipótesis determinada, también conocida como hipótesis nula (H_0) y una hipótesis alternativa (H_1), y se intenta dirimir cuál de las dos es la hipótesis verdadera, tras aplicar el problema estadístico a un cierto número de experimentos. Siempre que queremos hacer afirmaciones sobre la distribución de los datos o sobre si un conjunto de resultados es diferente de otro conjunto de resultados, debemos realizar un contraste de hipótesis.

CITA

"Existen dos posibles resultados: si el resultado confirma la hipótesis, habrás hecho una medición. Si el resultado es contrario a la hipótesis, entonces tendrás un descubrimiento."

Enrico Fermi.

2.1 PARÁMETROS VS ESTADÍSTICOS

Un parámetro es una descripción resumida de una característica o medida fija de la población objetivo. Un parámetro denota el verdadero valor que se obtendría si se realizara un censo en lugar de una muestra.

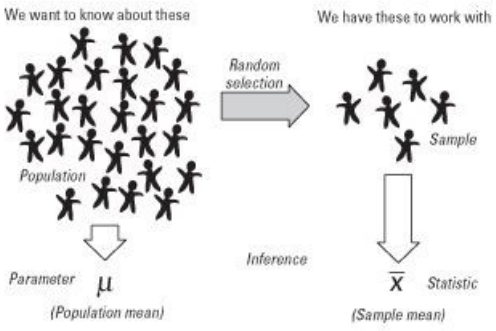
EJEMPLO

- *Media poblacional (μ).*
- *Varianza poblacional (σ^2).*

El estudio de una gran cantidad de datos individuales de una población puede ser farragoso e inoperativo, por lo que se hace necesario realizar un resumen que permita tener una idea global de la población, compararla con otras, comprobar su ajuste a un modelo ideal, realizar estimaciones sobre datos desconocidos de la misma y, en definitiva, tomar decisiones. A estas tareas contribuyen de modo esencial los estadísticos.

EJEMPLO

- *Media muestral ($\bar{x}$).*
- *Varianza muestral (S^2).*



Fuente: https://sites.google.com/site/zebrein/data-science/statistical-inference-for-data-science

Los estadísticos, substituyen grandes cantidades de datos por unos pocos valores extraídos de aquellos a través de operaciones simples. Durante este proceso, se pierde parte de la información ofrecida originalmente por todos los datos. Es por esta pérdida de datos, por lo que la estadística ha sido tildada en ocasiones de una falacia.
Por ejemplo, si en un grupo de tres personas una de ellas ingiere tres helados, el parámetro que con más frecuencia se utiliza para resumir datos estadísticos (la media aritmética del número de helados ingeridos por el grupo) sería igual a 1, valor que no parece resumir fielmente la información. Ninguna de las personas de este ejemplo se sentiría identificada con la frase resumen: *"He ingerido un helado de media".*

Hay todo tipo de frases míticas sobre la estadística, tales como:

"La estadística es una ciencia que demuestra que, si mi vecino tiene dos coches y yo ninguno, los dos tenemos uno."

George Bernard Shaw

...hasta nuestro gran sabio más contemporáneo se ha pronunciado:

"¡Oh!, la gente sale con estadísticas para probar cualquier cosa, el 14 por ciento del mundo lo sabe."

Homer Simpson

Fuente: https://giphy.com/gifs/season-13-the-simpsons-13x6-3o6Mbczc8Gva6kjRpS

2.2 HIPÓTESIS NULA (H_0)

En estadística, una hipótesis es una afirmación sobre un parámetro que sucede de la población (como la media o desviación típica) y se representa con H_0.

La hipótesis nula es la afirmación de **que dos (o más) parámetros o fenómenos no tienen relación entre sí.**

Es un punto de partida para la investigación que no se rechaza a menos que los datos de la muestra parezcan evidenciar que la hipótesis nula planteada es falsa.

Técnicamente, **la hipótesis nula es una aplicación del método de reducción al absurdo,** por el cual se supone, en principio, lo contrario de lo que se desea probar, hasta que los datos y pruebas obtenidas demuestran que el punto de partida era falso o absurdo y, por tanto, se rechaza. De esa forma se demuestra lo que se quería probar.

Dado que la hipótesis nula tiene la forma lógica de un enunciado universal, para afirmar que la hipótesis nula es verdadera se requiere estudiar a toda la población.

La hipótesis nula generalmente incluye una proposición simple de condición dicotómica (si/no, igual/distinto) para simplificar, aunque, en ocasiones, puede recoger un conjunto de valores (menor o igual a cero) haciendo la proposición más compleja o compuesta.

Así pues, si los resultados de nuestra muestra no respaldan la hipótesis nula, rechazamos la hipótesis; y la conclusión que aceptamos, y que afirma que existe alguna relación entre las muestras, se llama hipótesis alternativa.

En toda investigación estadística, para probar una hipótesis, es clave seleccionar una muestra representativa de la población de estudio. La hipótesis nula no es una excepción. Si el muestreo no se realiza adecuadamente, basándose en las muestras, es posible aceptar (o rechazar) equivocadamente una hipótesis nula.

EJEMPLO

Imaginemos que estamos haciendo un estudio en una empresa que fabrica más de 1 millón de chucherías al día.

H_0: los productos que estoy analizando no difieren de la especificación en cuanto a su peso.

Fuente: https://in.pinterest.com/pin/730498002044977710/

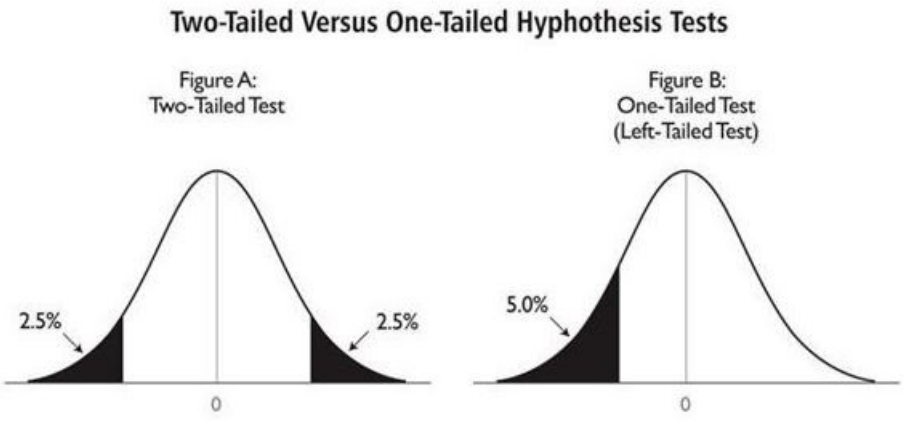
Para comprobar que un producto tiene el peso correcto, tomo una muestra de mil productos. Puede suceder que esos mil productos tengan un peso muy distinto a la media mientras que el resto no lo tengan en realidad (debido, por ejemplo, a que los escogí todos de un mismo lote y de la misma maquina procesadora, por lo que podría pasar que mi muestra no sea representativa. Basándome en esta muestra imaginaria rechazaría la hipótesis nula y afirmaría que los productos no tienen el peso correcto, cuando la realidad es otra muy distinta.

2.3 TEST DE 1 COLA VS 2 COLAS

Un test de hipótesis para una muestra es un tipo de test en el que el área crítica de una distribución es unilateral, de modo que es mayor o menor que un determinado valor en un solo lado de la distribución, pero no en ambos lados, como en el caso del test de hipótesis para dos muestras. **Si la muestra que se prueba cae en el área crítica, se rechazará la hipótesis nula.**

Tal y como podrás deducir, un test de hipótesis para varias muestras es un método en el que el área crítica de una distribución es de dos caras y comprueba si una muestra es mayor o menor que un cierto rango de valores hacia ambas direcciones. Si **la muestra que se prueba cae en cualquiera de las áreas críticas, se rechaza la hipótesis nula.**

Por convención, **el valor crítico (también denominado p-value),** suele situarse en el 5% para el test de hipótesis para una muestra y en el 2.5% hacia ambos lados para el test de hipótesis de varias muestras, lo que significa que cada lado de la distribución se corta en el 2,5%



2.4 TESTS ESTADÍSTICOS

El test estadístico se encarga de **medir lo cerca que está la muestra de la hipótesis nula. S**u valor observado cambia, aleatoriamente, de una muestra aleatoria a otra muestra diferente.

Un estadístico de prueba contiene información sobre los datos que son relevantes para decidir si se rechaza o no la hipótesis nula.

Las diferentes pruebas de hipótesis utilizan diferentes tests estadísticos basados en el modelo de probabilidad asumido en la hipótesis nula.

Tal y como podrás deducir, un test de hipótesis para varias muestras es un método en el que el área crítica de una distribución es de dos caras y comprueba si una muestra es mayor o menor que un cierto rango de valores hacia ambas direcciones. Si la muestra que se prueba cae en cualquiera de las áreas críticas, se rechaza la hipótesis nula.

Los tests más comunes son los siguientes:

Hypothesis test	Test statistic
Z-test	Z-statistic
t-tests	t-statistic
ANOVA	F-statistic
Chi-square tests	Chi-square statistic

En general, **los datos de la muestra deben proporcionar pruebas suficientes para rechazar la hipótesis nula** y concluir que el efecto existe en la población. Lo ideal es que un contraste de hipótesis no rechace la hipótesis nula cuando el efecto no esté presente en la población, y que rechace la hipótesis nula cuando el efecto exista.

2.5 TIPOS DE ERRORES

En este punto, ya habrás deducido que todo contraste de hipótesis funciona en base a la muestra que introduzcas como input del proceso, por lo que podríamos llegar a conclusiones distintas si se utiliza otra muestra distinta. Existen dos tipos de errores relacionados con las conclusiones incorrectas acerca de la hipótesis nula.

2.5.1 ERROR TIPO I

Se produce cuando ocurre un rechazo de la hipótesis nula cuando de hecho es verdadera.

Habitualmente, los encontrarás definidos con el término de **falsos positivos.**

Los errores de tipo I pueden ser controlados. El valor de alfa, que está relacionado con el nivel de significación que seleccionamos, tiene una relación directa con los errores de tipo I.

2.5.2 ERROR TIPO II

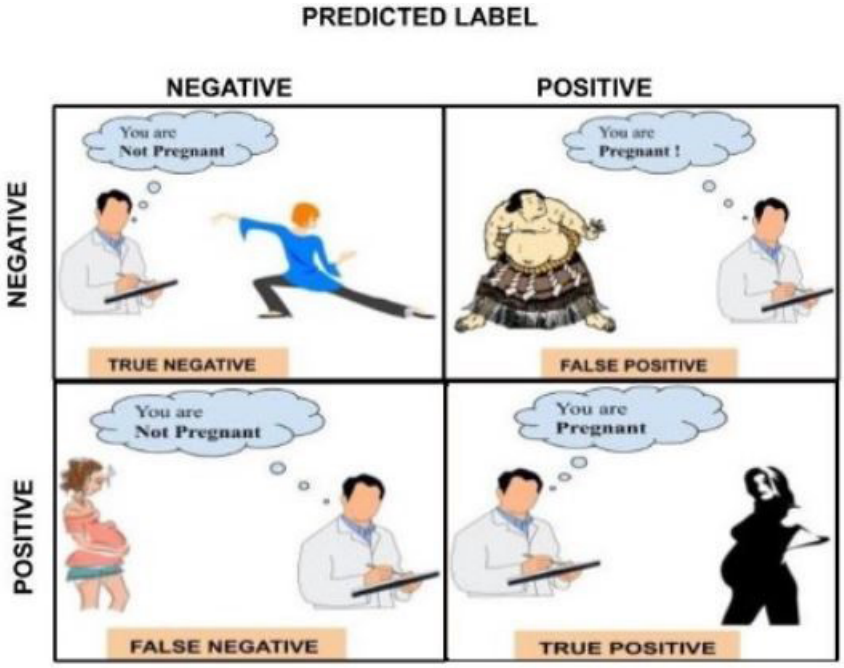
Se produce cuando no se rechaza la hipótesis nula cuando en realidad ésta es falsa.

Habitualmente los encontrarás definidos con el término **falsos negativos.**

A modo resumen, en la siguiente tabla puedes ver dónde se sitúan estos errores dentro de la **matriz de confusión:**

	H_0 es cierta	H_1 es cierta
Se escogió H_0	No hay error	Error de tipo II
Se escogió H_1	Error de tipo I	No hay error

Si prefieres un ligero toque de humor, entonces mejor utiliza este cuadro:



Fuente: https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd

RECUERDA

Este es un concepto clave que deberás tener en cuenta siempre en los modelos de clasificación, especialmente, en aquellos modelos con un **alto imbalanceo de positivos y negativos.**

2.6 NIVEL DE SIGNIFICACIÓN (α)

El nivel de significación es, comúnmente, representado por el símbolo griego α (alfa).

Son comunes los niveles de significación del 0.05, 0.01 y 0.001.

Si un contraste de hipótesis proporciona un p-valor inferior a α, **la hipótesis nula es rechazada,** siendo tal resultado denominado **estadísticamente significativo.**

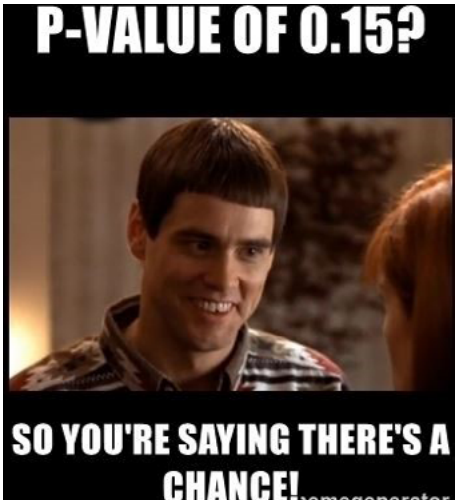
Cuanto menor sea el nivel de significación, más fuerte será la evidencia de que un hecho no se debe a una mera coincidencia (al azar). Es decir:

- Si nuestro p-valor toma el valor 0.051, para un α=0.05: mo rechazamos la hipótesis nula.
- Si nuestro p-valor toma el valor 0.04932, para un α=0.05: rechazamos la hipótesis nula.

Fuente: https://www.pinterest.com/pin/397583473330242821/

2.7 P-VALOR

El P-Valor (conocido también como p, p-valor, valor de p consignado o, directamente, en inglés p-value) se define como la **probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula.**



Fuente: https://memegenerator.net/instance/58348989/lloyd-so-youre-saying-theres-a-chance-p-value-of-015-so-youre-saying-theres-a-chance

En términos simples, el p-valor ayuda a diferenciar resultados que son producto del azar del muestreo, de resultados que son, estadísticamente, significativos.

El P-Valor se utiliza en muchos ámbitos de la estadístcia, desde los contrastes de hipótesis hasta las regresiones pasando por modelos de clasificación realizados con árboles de decisión. **Es muy necesario utilizar y comprender el significado.** A pesar de ser un concepto tan importante, el valor P-valor es un concepto algo resbaladizo que, a veces, se interpreta incorrectamente.

En otras palabras, si has de presentar unos resultados en el que te van a preguntar acerca del p-value, **míratelo con cariño** y tómate tu tiempo, no cometas el mismo error que Jim (la H_0 de su problema estipula que no tiene ninguna oportunidad con la chica...).

2.8 METODOLOGÍA

Ahora que ya conoces todos los conceptos básicos relacionados con los contrastes de hipótesis, veamos la **metodología recomendada** a seguir para hacer un **análisis completo.**

2.8.1 PASO 1: HIPÓTESIS

Imaginemos que una empresa de grandes almacenes está valorando la implementación de un servicio

de compras por Internet para los clientes que, habitualmente, compran en sus tiendas. A tal efecto, quiere realizar la inversión sólo en el caso que se pueda demostrar que merezca la pena y que la cantidad de sus propios clientes que van a utilizar este servicio va a ser superior al 40%.

Así pues, si se rechaza la hipótesis nula H_0, se aceptará la hipótesis alternativa H_1. Esto significa que se implementará el servicio de compras por Internet. Por otra parte, si no se rechaza H_0, entonces no se implementará el nuevo servicio de compras por Internet, a menos que se obtengan pruebas adicionales.

Esta contraste de hipótesis es una prueba de una sola cola, porque la hipótesis alternativa se expresa en forma unidireccional.

La manera más adecuada de formular la hipótesis para este problema es la siguiente:

$$H_0: \pi <= 0.4$$

$$H_1: \pi > 0.4$$

2.8.2 PASO 2: SELECCIONA UN TEST

Para probar la hipótesis nula, es necesario seleccionar un test estadístico apropiado.

Para este ejemplo, sería apropiado el test estadístico z, que sigue la distribución normal estándar:

$$z = (p-\pi)/\sigma_p, \text{ donde } \sigma_p = sqrt(\pi(1-\pi)/n)$$

2.8.3 PASO 3: ELIGE TU α

Es importante definir cuáles serían los errores de Tipo I y Tipo II de nuestro problema concreto.

En nuestro ejemplo, se produciría un error de Tipo I si concluyéramos, basándonos en los datos de la muestra, que la proporción de clientes que prefieren el nuevo plan de servicios es superior a 0.40, cuando en realidad es inferior o igual a 0.40.

El error de Tipo II se produciría si concluyéramos, basándonos en los datos de la muestra, que la proporción de clientes que prefieren el nuevo plan de servicios es menor o igual a 0.40 cuando, en realidad, es mayor de 0.40.

A tal efecto, es necesario equilibrar los dos tipos de errores. Por convención el valor de α suele fijarse en 0,05, aunque, a veces, puede ser inferior si se quiere una solución más robusta (habitualmente al 0.01). Utilizar otros valores para α suele ser menos común. En nuestro ejemplo, consideraremos el valor 0.05 para α.

2.8.4 PASO 4: PREPARA LA MUESTRA

El tamaño de la muestra se determina teniendo en cuenta el α deseado y otras consideraciones cualita-

tivas, así como las limitaciones en términos de presupuesto para poder reunir los datos de la muestra. Para nuestro ejemplo concreto, vamos a imaginarnos que se realizó una encuesta a 30 clientes y 17 indicaron que utilizaban Internet para hacer compras.

Así pues, el valor de la proporción de la muestra es p=17/30=0.567.

El valor de:

$$\sigma_{\bar{p}} = \text{sqrt}(\pi(1-\pi)/n = \text{sqrt}\ ((0{,}40)(0{,}60)/30)=0.089.$$

El estadístico lo podemos calcular del siguiente modo:

$$z=(p-\pi)/\sigma_{\bar{p}}= (0.567-0.40)/0.089 = 1.88$$

2.8.5 PASO 5: VALOR CRÍTICO

Para poder determinar el valor crítico, debemos ir a buscar los valores críticos que toma el estadístico en su propia tabla, en la cuál tendremos en uno de los ejes el valor de α y en otro de los ejes el valor del estadístico.

Utilizando las tablas para el estadístico que hemos escogido para nuestro problema, la probabilidad de obtener un valor z de 1,88 es de 0.96784, es decir:

$$P(z \leq 1.88)=0{,}96784$$

En la siguiente imagen podrás ver cómo es esta tabla:

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670

Fuente: propia

Pero en nuestro caso concreto, estamos interesados en calcular la probabilidad a la derecha de z, ya que

queremos obtener el valor de probabilidad que cae en la región de rechazo o región crítica), es decir:

1-0.96784 = 0.03216

Esta probabilidad es directamente comparable a α.

2.8.6 PASO 6: TOMA LA DECISIÓN

Tal y como hemos visto en el punto anterior, el valor crítico es de 0.03216. Esta es la probabilidad de obtener un valor P de 0.567 (proporción de la muestra = p) cuando π=0,40. Esto es inferior al nivel de significación (α) de 0.05 que habíamos definido en el paso 3. Por lo tanto, la hipótesis nula es rechazada.

En nuestro ejemplo, llegamos a la conclusión de que hay pruebas de que la proporción de usuarios de Internet que compran a través de la red es, significativamente, superior al 0.40. Por lo tanto, la recomendación a los grandes almacenes sería introducir el nuevo servicio de compras por Internet.

2.9 EL TEST AB

Es muy común que, a menudo, te preguntes si el modelo que implementaste o la acción sobre un determinado segmento de clientes después de extraer conclusiones de tu análisis estadístico tuvo algún tipo de impacto, o bien si el escenario que implementaste definitivamente fue mejor que cualquier otro alternativo. El Test AB es una metodología muy interesante que te permitirá poder testar infinidad de modelos y ver si, realmente, funcionan o no.

El test AB, básicamente, es una metodología de comparación de múltiples versiones de una determinada variable, una landing page, un determinado botón... mostrando las diferentes versiones a los clientes o posibles clientes y evaluando la calidad de la interacción mediante alguna métrica.

Cada vez que quieras probar múltiples variaciones de un determinado output y quieras realizar experimentos, el Test AB es una opción muy interesante.

Para empezar a realizar un test AB, hay que empezar especificando lo que crees que sucederá, que será hipótesis alternativa, mientras que la hipótesis nula no asumirá ninguna diferencia entre las variantes. Por lo tanto, si rechazas la Hipótesis Nula, estarás validando lo que crees que sucederá.

En resumen, necesitarás:

- **Especificar tu Hipótesis Alternativa:** esto es, lo que crees que sucederá. Por ejemplo: La variante B funcionará un 20% mejor que la variante A.
- **Especificar la Hipótesis Nula:** se basará en la asunción que no existe ninguna diferencia entre la variante B y la variante A.
- **Variable objetivo o independiente:** esto significa decidir lo que tu variable quiere conseguir. Ejemplos de ello pueden ser conseguir que alguien haga clic en la siguiente página, ponga más cosas en el carrito de compra virtual, o realmente cualquier otra cosa que implique múltiples variaciones de una acción Sea cual sea esa llamada a la acción o métrica, eso es lo que usaremos para interpretar el rendimiento de nuestras variaciones.

- **Variables independientes:** Al diseñar cualquier experimento, deberás definir qué variables independientes o explicativas deseas utilizar para predecir la variable dependiente. en el caso de test AB, la variable explicativa de la variación de la variable dependiente es, simplemente, qué versión se muestra para conducir a qué resultado.

Otro aspecto clave a la hora de diseñar un Test AB es el aspecto temporal. Es importante que las variantes se distribuyan a los clientes durante el mismo período de tiempo. El tiempo es un buen ejemplo de algo en lo que hay que estandarizar en lugar de dejar que la estacionalidad juegue un papel en el resultado de su experimento. Cuando lances un experimento, lánzalo para todas y cada una de las variantes.

Una vez definidos todos estos puntos, la siguiente pregunta que nos debemos hacer es: ¿cuántas muestras de cada variante necesito para tener resultados estadísticamente significativos?

Para determinar esto, realizamos algo llamado **análisis de potencia.** La idea del análisis de potencia es que identifica el tamaño de muestra necesario en base a una serie de parámetros; cosas como la potencia estadística, el p valor, el número de variantes, y el tamaño de la diferencia entre la medición de los dos grupos, etc. La razón por la que hacemos esto es para asegurarnos de no hacer un experimento tan largo que una tonelada de nuestros clientes tuviera que ver la peor versión, pero, aún así, lo suficiente para justificar nuestros resultados.

Formalmente:

- **k - número de variantes:** al menos dos y tantas como quieras. Una cosa a tener en cuenta es que cuantas más variantes, más datos se necesitan.
- **n - tamaño de la muestra por grupo:** dejaremos esto como valor nulo, eso es lo que estamos resolviendo.
- **f - diferencia observada entre los grupos que queremos validar:** cuanto mayor sea la diferencia, menor será la muestra requerida y cuanto menor sea la diferencia, mayor será la muestra requerida para validarla.
- **Nivel de significación:** p valor a utilizar. Típicamente, aceptaremos un resultado que sea estadísticamente significativo al 0.05.
- **Poder - poder estadístico:** esto significa, más o menos, que si tu hipótesis es cierta, ¿cuál es la probabilidad de que la aceptes? El estándar es, típicamente, 0,8.

3. ACTIVIDAD GUIADA

Vamos a suponer que trabajas en el departamento de Marketing de una empresa y el responsable de UX quiere testar si realizar un cambio de lugar en uno de los botones de la página web clave (el carrito) repercute en una mayor tasa de clics.

Para este ejercicio, partimos de la baseline de que, actualmente, ya hay un 10% de probabilidades de que la gente pulse el botón del carrito de la compra en el lugar que está situado. Con la modificación propuesta por el desarrollador, esperamos un cambio del dos por ciento en la tasa de clics.

El poder que le asignamos a nuestro test AB es del 80% y vamos a exigir que sea, estadísticamente, significativo al 0.05 para que los resultados sean aceptables.

A tal efecto, utilizaremos la función power.prop.test implementada en el paquete base de R para saber el número de observaciones necesarias a realizar en cada escenario.

```r
# Parametros
baseline  <- 0.1
delta     <- 0.02
power     <- 0.8
sig_level <- 0.05

# Funcion para saber el número de observaciones necesarias
result <- power.prop.test(
  p1 = baseline,
  p2 = baseline + delta,
  power = power,
  sig.level = sig_level,
  alternative = "two.sided"
)
result
```

```
     Two-sample comparison of proportions power calculation

              n = 3840.847
             p1 = 0.1
             p2 = 0.12
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

El resultado nos muestra que necesitamos una muestra de, al menos, 3841 observaciones en cada escenario para detectar si efectivamente el cambio producido es del 2%.

Vamos a suponer que hemos lanzado el test y hemos obtenido el número total de muestras y el número total de aciertos para cada uno de los grupos. Dadas estas variables, podemos utilizarlas para calcular si el cambio ha sido debido a la implementación del escenario o no.

```r
# Parametros
count_control <- 974       # Numero de clicks en el baseline
sizes_control <- 10072     # Numero de observaciones en el baseline
count_experiment <- 1242   # Numero de clicks en el nuevo escenario
sizes_experiment <- 9886   # Numero de observaciones en el nuevo escenario

# Realización del 2-sample test
result <- prop.test( c(count_control, count_experiment),
                     c(sizes_control, sizes_experiment) )
result

# Computamos la probabilidad de cada grupo y el erro estandar
p1 <- count_control / sizes_control
p2 <- count_experiment / sizes_experiment
se <- sqrt( p1 * (1 - p1) / sizes_control + p2 * (1 - p2) / sizes_experiment )

# 95 percent confidence interval's z score
conf_level <- 0.95
zscore <- qnorm( conf_level + (1 - conf_level) / 2 )
conf_int <- abs(p2 - p1) + c(-1, 1) * zscore * se
conf_int
```

```
        2-sample test for equality of proportions with continuity correction

data:  c(count_control, count_experiment) out of c(sizes_control, sizes_experiment)
X-squared = 42.007, df = 1, p-value = 9.097e-11
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.03774653 -0.02011042
sample estimates:
    prop 1     prop 2
0.09670373 0.12563221

[1] 0.02021064 0.03764631
```

El cambio debería ser mayor que el mínimo cambio detectable que deseas detectar, es decir, en nuestro caso, hemos definido que el cambio debe ser del 2%. En los resultados de arriba, puedes ver que nuestro intervalo de confianza se encuentra por encima de ese 2% (está entre 2.02% y 3.7%), por lo que, definitivamente, y según este pequeño test realizado, sí que realizaríamos el cambio de botón en la página web.

IDEAS CLAVE

- El Análisis Multivariante permite ver el efecto de más de dos dimensiones.
- La regresión representa la relación causa-efecto, mientras que el clustering representa la interrelación estructural de las variables.
- Un parámetro representa una descripción resumida de una variable perteneciente a la población, mientras que un estadístico representa una descripción resumida perteneciente a una muestra de la población.
- La hipótesis nula es la afirmación de que dos (o más) parámetros o fenómenos no tienen relación entre sí.
- El test estadístico se encarga de medir lo cerca que está la muestra de la hipótesis nula.
- Si la muestra que se prueba cae en el área crítica, se rechazará la hipótesis nula.
- Los errores de tipo I se conocen, popularmente, como Falsos Positivos, mientras que los errores de tipo II se conocen, popularmente, como Falsos Negativos.
- El nivel de significación marca el punto donde la hipótesis nula es rechazada, siendo tal resultado denominado estadísticamente significativo.
- El p valor representa la probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula