2016

Ediciones Ay Carumba!

El Barto

[MATEMÁTICA & ESTADÍSTICA (GEOLOGÍA)]

Guía práctica e introductoria –no oficial– para la cátedra de Matemática & Estadística (Geología)

Índice

Unidad I. Estadística. Estadística descriptiva. Tablas. Gráficos

¿Qué es la Estadística?

Estadística es una herramienta que permite realizar inferencias y obtener conclusiones a partir del análisis de datos que, en un principio, no aportan mucha información precisa.

Estadística descriptiva

Es una rama de la estadística relacionada con la descripción de conjuntos de datos específicos, tanto de muestras como de poblaciones.

Áreas de la Estadística

Existen tres:

- *Diseño*: Planeamiento y desarrollo de investigaciones.
- *Descripción*: Resumen y exploración de datos.
- *Inferencia*: Hacer predicciones o generalizaciones acerca de características de una población en base a la información de una muestra de la población.

El *diseño* es una actividad crucial. Consiste en definir el desarrollo de la investigación para dar respuesta a las preguntas que motivaron la misma. La recolección de los datos es muy importante, por lo que ser cuidadoso en la etapa de planificación de la investigación influye en las siguientes etapas. Así, un estudio bien diseñado es simple de analizar y las conclusiones suelen ser obvias, mientras que un experimento pobremente diseñado o con datos mal recolectados puede dar respuestas incorrectas a las preguntas que motivaron la investigación, más allá de lo elaborado que sea el análisis estadístico.

La *descripción* de los datos ayuda a presentar a los mismos de modo tal que sobresalga su estructura. Se los puede organizar en gráficos, que permiten detectar tanto las características sobresalientes como inesperadas, o resumirlos en uno o dos números que pretenden caracterizar el conjunto con la menor distorsión o pérdida de información posible.

La *inferencia* hace referencia a un conjunto de métodos que permiten hacer predicciones sobre la base de información parcial acerca del fenómeno que se esté estudiando.

Los métodos de inferencia permiten proponer el valor de una cantidad desconocida (*estimación*) o decidir entre dos teorías contrapuestas, en las que una de ellas explica mejor los datos observados (*test de hipótesis*).

El objetivo principal de cualquier estudio es aprender sobre las poblaciones. Pero usualmente es más práctico estudiar solo una muestra de cada una de las poblaciones.

Así, se define:

- *Población*: conjunto de todos los datos específicos de interés para el investigador.
- *Muestra*: subconjunto de datos elegidos de la población de interés.
- *Parámetro*: es una medida resumen obtenida a partir de la población.
- *Estadístico*: es una medida resumen obtenida a partir de la muestra.

- *Variable*: característica que cambia o varía con el tiempo y/o para los diferentes individuos u objetos que se consideren.
- *Unidad experimental*: también llamado elemento de muestra, es el objeto sobre el cual se toma una medición. También se podría definir como el objeto sobre el cual se mide una variable.
- *Medición o dato*: se obtiene cuando se mide, en la realidad, una variable sobre una unidad experimental.
- *Datos univariados*: se obtienen cuando se mide una sola variable en una sola unidad experimental.
- *Datos bivariados*: se obtienen cuando se miden dos variables en una sola unidad experimental.

Tipos de variables

Al diseñar una investigación, se intenta estudiar de qué modo una o más variables (*variables independientes*) afectan a una o más variables de interés (*variables dependientes*).

De esta manera, es importante identificar cuántas variables se registraron y cómo fueron registradas, lo que permitirá definir la estrategia de análisis.

A partir de eso, se puede clasificar las variables como:

1) *Variable cualitativa*: mide una cualidad o característica en la unidad experimental. Produce datos que se pueden clasificar de acuerdo con similitudes o diferencias de clase, por lo cual también se la denomina dato categórico.
2) *Variable cuantitativa*: mide una cantidad numérica en cada unidad experimental. Para describir los tipos de valores numéricos que pueden adoptar las variables cuantitativas, éstas se clasifican en:
   a) *Variable discreta*: sólo puede adoptar un número contable o finito de valores.
   b) *Variable continua*: puede adoptar una cantidad infinita de valores que corresponden a puntos en un intervalo lineal.

También se pueden considerar clasificaciones tales como variables nominales, ordinales o porcentuales, entre otras.

Por lo tanto, los métodos a usar para describir los conjuntos de datos dependen del tipo de datos que se haya reunido.

¿Por qué es importante identificar el tipo de variable? Porque va a determinar el método de análisis más apropiado, y que arroje por lo tanto mejores resultados para el estudio que se realice.

Tablas

Luego de recolectar los datos, se pueden consolidar y resumir para mostrar cuáles son los valores de la variable que se midieron y con qué frecuencia apareció cada uno de ellos.

La manera más sencilla de presentar los datos es por medio de una tabla de frecuencias, la cual indica el número de observaciones que caen en cada una de las clases de la variable.

Para construir la tabla de frecuencias, se divide el rango total de los datos en clases o intervalos. Luego, se cuenta el número de observaciones que cae en cada clase y se determina la *frecuencia* en cada una de ellas. Por último, se calculan las *frecuencias relativas*, *frecuencias acumuladas* y *frecuencias acumuladas relativas* para cada intervalo.

Notación de frecuencias

- *Frecuencia (f)*: número de casos observados para cada intervalo.
- *Frecuencia relativa (f_r)*: frecuencia observada, para cada intervalo, dividida por la suma de frecuencias (N):

$$f_i / \Sigma f$$

La suma de frecuencias relativas siempre debe dar 1

- *Frecuencia relativa porcentual (f_r%)*: porcentaje de casos en el intervalo i-ésimo:

$$(f/\Sigma f)\cdot 100$$

- *Frecuencia acumulada (f_{ac})*: suma de las frecuencias desde la primer categoría hasta la categoría i-ésima:

$$f_1 + f_2 + ... + fi$$

El último valor de la suma de frecuencias acumuladas siempre debe ser igual a N

- *Frecuencia acumulada relativa porcentual (f_{ac}%)*: es la suma de las frecuencias acumuladas relativas, desde la primer categoría hasta la última:

$$(f_{ac}/N)\cdot 100$$

El último valor siempre tiene que ser igual a 100. Si grafico estos valores me va a dar lo que se llama *ojiva*

Gráficos

Para variables cualitativas:

- *Gráfico de sectores*: también conocido como gráfico de torta o circular, muestra cómo se distribuyen los datos específicos entre las categorías.
- *Gráfico de barras*: se muestra la misma distribución de mediciones entre las categorías, en donde la altura de la barra determina con qué frecuencia se observa una categoría en particular.

A veces, la información se reúne para una variable cuantitativa medida en diferentes segmentos de la población o para categorías diferentes de clasificación. En estos casos, se pueden usar gráficos de sectores o gráficos de barras para describir los datos, por medio de la cantidad medida en cada categoría en lugar de la frecuencia de ocurrencia de la misma.

La principal ventaja del gráfico de barras sobre la circular es que ésta permite visualizar el comportamiento de las categorías entre sí.

Para variables cuantitativas:

- *Gráfico de líneas*: cuando la variable tiene un registro temporal o intervalos igualmente espaciados, el conjunto de datos forma una serie de tiempo. Los datos se representan con más eficacia en un gráfico de líneas con el tiempo en el eje de las abscisas. Así, se observa una tendencia hacia el futuro.
- *Diagrama de dispersión*: es el tipo de gráfico más simple para interpretar los datos cuantitativos que constan de números que no pueden separarse fácilmente en categorías o intervalos de tiempo.

- *Distribución simétrica*: los lados izquierdo y derecho forman imágenes idénticas, dividiéndose por su valor medio.
- *Distribución sesgada a la derecha*: si una proporción mayor de los datos específicos se localiza a la derecha del valor máximo, generalmente contienen algunas medidas extraordinariamente grandes.
- *Distribución sesgada a la izquierda*: si una proporción mayor de los datos específicos se localiza a la izquierda del valor máximo, contienen datos excepcionalmente pequeños.

Una distribución es unimodal si tiene un máximo, mientras que una distribución bimodal tiene dos máximos y representa (generalmente) una mezcla de dos poblaciones diferentes en el conjunto de datos.

Construcción de un histograma

Primero, se trazan dos ejes (x e y). En el eje x se representan los valores de la variable y en el eje y una medida de frecuencia.

A continuación, en el eje x los límites de los intervalos de clase. Se asocia a cada clase una columna cuya base cubre el intervalo de la misma y cuya altura indica cuantos datos son tomados en un intervalo a través de la frecuencia de la clase. El gráfico se construye con las columnas una a continuación de la otra, sin espacios, a menos que una clase no tenga valores para ser graficados.

Existen fórmulas para determinar el número de clases a usar en el histograma, pero convencionalmente se utilizan entre 5 y 6 clases, ya que permiten un análisis claro del gráfico.

Es importante mencionar que este tipo de gráfico se utiliza para variables cuantitativas continuas, mientras que para variables cuantitativas discretas se utiliza el gráfico de barras (con espaciado horizontal entre barras). También es importante mencionar en el eje y qué frecuencia se eligió y, en el eje x, las unidades de la variable con la que se esté trabajando.

Los histogramas pueden presentar una distribución acampanada, uniforme, asimétrica derecha o asimétrica izquierda.

La marca de clase (MC) es un valor central, utilizado cuando agrupo variables continuas en clases, respetándose el valor de intervalo de clase entre las marcas de clases.

Polígono de frecuencia

El polígono de frecuencias es similar al histograma, aunque su función es dar una imagen aproximada de la curva definida por la distribución de la variable.

Para construirlo, se usan los mismos ejes que en el histograma. En el eje x se indica el punto medio de cada intervalo mientras que, en el eje y, se indica la escala densidad para ese intervalo. Esto va a definir pares (x,y) en el gráfico que se unen con líneas rectas.

Se pueden marcar, además, los puntos medios del intervalo que precede al primero y del que sigue al último.

Ejemplos de gráficos hay al final de la unidad.

Parámetros y estadísticos

Para describir las distribuciones, se definen medidas (o valores) que dan cuenta de la:

- Tendencia central.
- Dispersión, o variabilidad, de los datos.
- Forma.

Estas medidas van a permitir la descripción de la muestra de estudio y de la población.

Así, para los estadísticos (datos de la muestra) se utilizan letras mayúsculas, mientras que para los parámetros (datos de la población) se utilizan letras griegas.

Medidas de tendencia central:

- *Moda*: serie de datos en la que un valor aparece con más frecuencia que cualquier otro.es inestable ya que puede cambiar con el método de redondeo. En datos agrupados, se encuentra en la clase de mayor frecuencia (clase modal).
- *Mediana (Md)*: es el valor medio cuando los valores se disponen según sus magnitudes. En datos agrupados, es la clase cuya frecuencia acumulada supera primero el valor de la mitad más uno de los datos [(n+1)/2].
- *Media aritmética ($\bar{x}$)*: es igual a la suma de los n valores dividido el número total de valores (n).

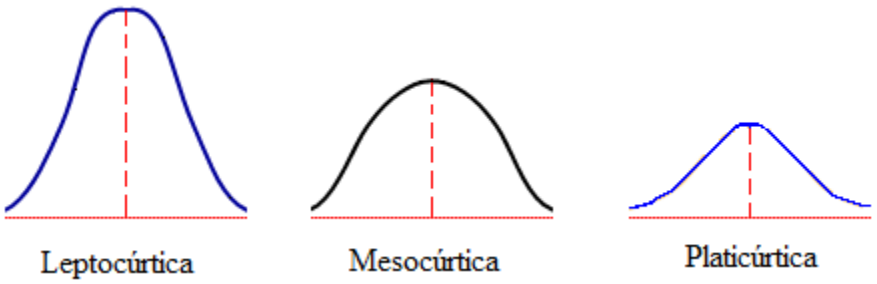
Medidas de dispersión:

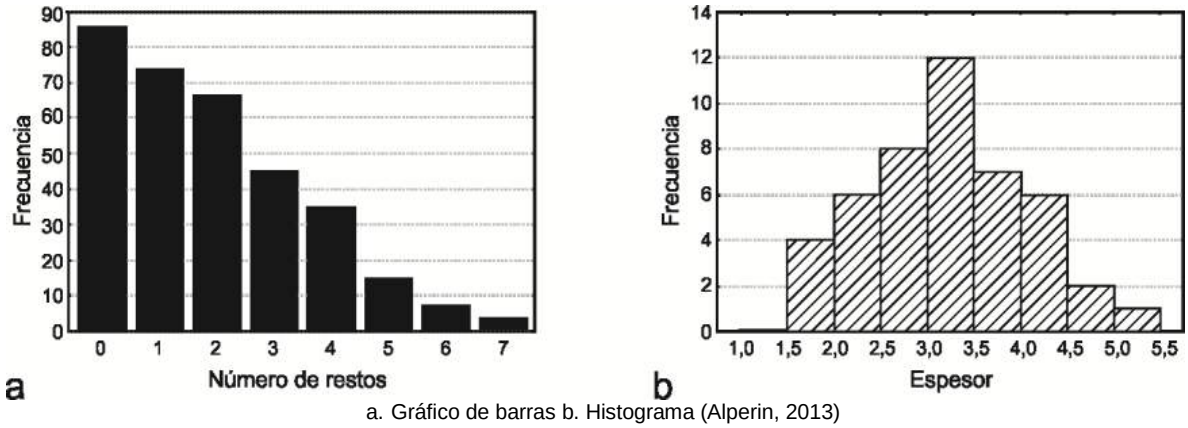
- *Amplitud, rango o recorrido*: para un conjunto de n observaciones, es la diferencia entre el valor máximo y el mínimo.
- *Cuartiles*: sirven para dividir la distribución en cuatro partes iguales. Para ello, se ordenan los datos de mayor a menor, el cuartil inferior va a ocupar la posición [(n+1)/4] en la muestra ordenada, el cuartil central va a coincidir con el valor de la mediana y el cuartil superior la posición [3(n+1)/4] de la muestra ordenada.
  Si la posición resulta ser un número decimal, se promedian los valores a izquierda y derecha del mismo.
- *Varianza (S^2)*: promedio del cuadrado de las desviaciones de los datos con respecto a la media.
- *Desvío estándar (S)*: raíz cuadrada positiva de la varianza.
- *Coeficiente de variación (CV)*: es una medida que da cuenta de la variabilidad relativa de las observaciones. Se calcula como el cociente entre el desvío estándar y la media. Es adimensional, puede ser positivo o negativo como también puede expresárselo de forma porcentual.

Forma:

- *Coeficiente de simetría (CS)*: informa si los datos están equilibrados en torno a la media o si hay más hacia un lado u otro.

- *Coeficiente de curtosis (E)*: mide el grado de achatamiento de un histograma con respecto al modelo teórico normal. Puede suceder que E sea menor a cero, dando un histograma con una distribución más achatada que lo normal (curva platicúrtica); igual a cero, dando un histograma con una distribución normal (curva mesocúrtica); mayor a cero, dando un histograma más puntiagudo que lo normal (curva leptocúrtica).





a. Gráfico de barras b. Histograma (Alperin, 2013)

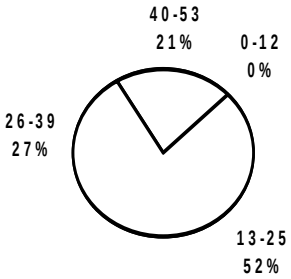


Gráfico de torta

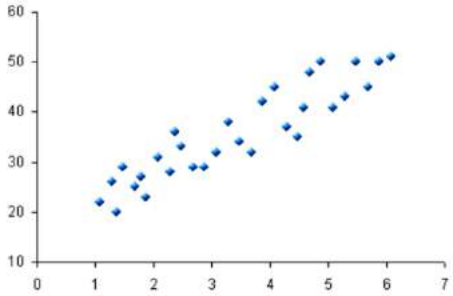


Diagrama de dispersión

Unidad II. Probabilidades. Distribución de probabilidades

La Probabilidad, junto a la Estadística, se encarga del estudio del azar, encarado desde el punto de vista matemático. Así, la Probabilidad propone modelos para los fenómenos aleatorios (aquellos que se pueden predecir con certeza) y estudia sus consecuencias lógicas, mientras que la Estadística ofrece métodos y técnicas que permiten entender los datos a partir de aquellos modelos planteados en la primera.

Al hablar de métodos estadísticos, nos referimos a aquellos planteados en la Estadística Descriptiva (recolección, clasificación y análisis de datos), mientras que las técnicas son aquellas herramientas que permitirán realizar inferencias respecto a los datos obtenidos, y que se citarán más adelante.

La *probabilidad* (P) es una relación entre el número de casos favorables con respecto al número de casos total (definición clásica):

$$P = \frac{\text{nº de casos favorables}}{\text{nº de casos total}}$$

Pudiendo tener sucesos de ocurrencia *equi-* o *inequiprobables*.

Tomemos como ejemplo un dado de 6 caras. Generalmente los mismos vienen equilibrados, con el peso repartido equitativamente en todas sus caras. Elija el número que elija, su probabilidad será siempre 1/6 (ya sea que haya elegido el 1, el 4 ó el 6), puesto que el *número de casos favorables* va a ser 1 (elegí UN solo número) y el número de casos total va a ser siempre 6 (es un dado de 6 caras, por lo tanto puede salir cualquiera de esas 6 caras). Para este tipo de dado, con su peso equilibrado en todas las caras, cada uno de los 6 sucesos (que salga algún número entre 1 y 6) van a ser *equiprobables*. O sea, que todos los números tienen la misma probabilidad de salir (1/6)

Ahora bien, puede pasar que el dado esté 'alterado'. Esto significa que alguna de sus caras (o más de una) tenga un poco de peso más que la otra, por lo que la probabilidad se va a ver alterada y, por más que repitamos el evento de tirar el dado hasta que salga el número que elegí, no voy a obtener los mismos resultados. Para hacerlo un poco más claro, la probabilidad va a dejar de ser 1/6 para todas las caras, y esto hace que cada uno de los sucesos sean *inequiprobables*.

Otra situación a tener en cuenta en este ejemplo es la siguiente: ¿Cuál es la probabilidad de sacar un número par? Bueno, sabemos que el dado tiene 3 caras con números pares (2, 4, 6) por lo que la probabilidad será 3/6 (o sea, ½) y va a ser la misma si me planteo qué probabilidad voy a obtener de sacar un número impar (porque tengo el 1, el 3 y el 5).

¿Y si quiero saber la probabilidad de obtener un número menor a n? Supongamos que n vale 5, por lo tanto, al tirar el dado, puedo sacar cualquier número entre el 1 y el 4. En este caso, el nº de casos favorables va a ser 4 y la probabilidad 4/6 (o sea, 2/3). Tal vez alguno se pregunte '¿Y con el 5 qué pasa? Pasa que me pregunté la probabilidad de obtener un número MENOR a 5, no MENOR O IGUAL. ***¡Ojo! Siempre hay que estar atento a lo que se plantea o pregunta para no obtener errores en los resultados***.

Otra opción es querer saber la probabilidad de obtener un número par menor a 5. Vamos a tener, como opciones, el 2 y el 4. Son dos casos favorables en un total de 4 (2/4, o un medio, que viene a ser lo mismo). El mismo caso, planteado para querer saber la probabilidad de un número impar y menor a 5 da el mismo resultado (½).

Tenemos otra definición para probabilidad, una que es empírica:

$$\lim_{n \to \square} fr = P$$

De momento, en esta unidad vamos a usar más la primera definición (la clásica) y vamos a dejar la empírica solamente planteada. Esto es porque es importante saber que existen ambas, pero a los fines prácticos vamos a usar más la clásica, y para algunos conceptos teóricos (de otros capítulos) se puede llegar a hacer mención de la empírica.

Antes de seguir, es importante definir algunas cosas:

- *Experimento aleatorio*: también llamado simplemente experimento, es cualquiera operación cuyo resultado no puede ser predicho con seguridad de antemano. Ejemplos de esto son el lanzamiento de una moneda, de un dado, la extracción de una carta de una baraja de 52 cartas
- *Espacio muestral*: es el conjunto de todos los resultados posibles asociados a un experimento. Su símbolo es la letra griega Omega (Ω). Puede ser *discreto*, si tiene un número finito de elementos, o *continuo*, si tiene como elementos todos los puntos de algún intervalo real. Ejemplos de esto son el lanzamiento de un dado (discreto) o el tiempo de duración de un tubo fluorescente (continuo).
- *Evento*: también llamado *suceso*, es cualquier subconjunto de un espacio muestral. Podemos tener los llamados *sucesos seguros* (obtener un valor n en un dado de 6 caras) y los *imposibles* (sacar un 7 en un dado de 6 caras).
- *Intersección de sucesos (A∩B)*: es un suceso que está formado por todos los elementos que son, a la vez, de A y de B. se lee 'A y B'.
- *Unión de sucesos (AUB)*: es el suceso que está formado por todos los elementos de A y de B. se lo verifica cuando ocurre A, B o ambos. Se lee 'A o B'.

Axiomas de probabilidad

Siendo Ω el espacio muestral que tengo, y A y B dos eventos cualesquiera del mismo:

- *Axioma 1*: $P(\Omega) = 1$
- *Axioma 2*: $P(A) \geq 0 \quad \forall\ A \subseteq \Omega$
- *Axioma 3*: $P(AUB) = P(A) + P(B)$ si $A \cap B = \emptyset$

Probabilidad condicionada

Este tipo de probabilidad se usa cuando quiero calcular la probabilidad de un evento A cualquiera, teniendo información sobre un evento B. Se expresa como P(A/B), se lee 'probabilidad de A habiendo ocurrido B' y se define de la siguiente manera:

$$P(A/B) = \frac{A \cap B}{P(B)} \qquad \text{con } P(B) \text{ distinto a cero}$$

Uno de los usos más frecuentes de esto es para dar un procedimiento sencillo para designar probabilidades a intersecciones de eventos.

Probabilidad compuesta

También llamada *regla de multiplicación de probabilidades*, proviene de la probabilidad condicionada y establece lo siguiente: la probabilidad de que dos sucesos (suceso de intersección de A y B) se den de forma simultanea es igual a la probabilidad a priori del suceso A multiplicada por la del suceso B condicionada al cumplimiento del suceso A.

La fórmula para calcular este tipo de probabilidades es la siguiente:

$$P(A \cap B) = P(B/A) \cdot P(A)$$

Eventos independientes

Dos eventos son independientes entre sí, si la ocurrencia de uno de ellos no afecta para nada a la ocurrencia del otro. Dicho de otra forma, más matemática, dos eventos, A y B, son independientes si, y sólo si, $P(A \cap B) = P(A) \cdot P(B)$

Hay un teorema que dice: suponiendo que $P(A) \neq 0$ y $P(B) \neq 0$, entonces A y B independientes implica que ellos no son excluyentes y A , B mutuamente excluyentes implica que ellos no son independientes.

Esta parte es un poco complicada de visualizar y entender, pero no imposible. Es cuestión de agarrarle la mano y practicarlo un poco.

Eventos independientes. Extracción con devolución

La definición para este caso sigue siendo la misma: dos eventos, A y B, son independientes si la realización de A no condiciona a la de B:

$$P(A \cap B) = P(A) \cdot P(B)$$

Lo que cambia un poco es la fórmula, pero nada complicado. Y cambia por el tema de la devolución. Digamos que en una caja tenés 3 rocas volcánicas, 2 rocas ígneas y 5 sedimentarias. Eso nos da un total de 10 rocas ¿Bien? Ahora, si queremos saber la probabilidad de sacar una roca ígnea, cualquiera de las que tengo, va a ser 2/10, o sea, 1/5; si quiero saber la probabilidad para las volcánicas, va a ser 3/10 y, para las sedimentarias 5/10 (o sea, ½). Esto es así, manteniendo el total de las rocas (10) porque cada vez que saque una roca de la caja, luego la regreso (por eso es con devolución).

Eventos dependientes. Extracción sin devolución

Acá cambia un poco la definición: dos sucesos, A y B, son dependientes si la ocurrencia de A condiciona a la de B. Puesto en una fórmula, es algo así:

$$P(A \cap B) = P(A) \cdot P(B/A)$$

La fórmula es bastante parecida a la que tenemos un poco arriba, para probabilidad compuesta. El tema de extraer sin devolución es así: ¿viste las rocas que tenemos de ejemplo acá arriba? Bueno, si queremos saber la probabilidad de sacar una roca ígnea, ya sabemos que es 1/5, pero esa roca que sacaste de la caja ahora no la regresás. La dejás aparte, en un costado. Ahora, la probabilidad de una roca volcánica no va a ser más 3/10, porque ya no tenés 10 rocas en la caja (tenés 9 porque dejaste la ígnea en un costado), la probabilidad va a ser 3/9 (o sea, 1/3). Esa misma roca volcánica la dejás junto con la roca ígnea, ahí en el costado. Adentro de la caja nos quedan 8 rocas ahora ¿cierto? La probabilidad de que agarres una roca sedimentaria va a ser, ahora, 5/8. Y así, podés ir armando las probabilidades teniendo siempre en cuenta que cada vez que saques algo, no lo devolvés. Eso te va a ir modificando el valor de los casos totales (el número que va debajo de la división… Que viene de la definición de probabilidad, más arriba).

A veces, **pero no siempre**, es útil usar lo que se conoce como *diagrama de árbol*. Este diagrama sirve para ir armando dicotomías cuando una situación presenta dos opciones, cada una con su probabilidad, haciendo uso de multiplicaciones para obtener los valores de probabilidad, y sumas para ir obteniendo un valor de probabilidad final (llamémoslo de momento así, en la práctica todo esto se ve más fácil).

Variable aleatoria

Una variable aleatoria es una función que asocia un número real a cada elemento del espacio muestral. Se usan letras mayúsculas para designar a las variables aleatorias y letras minúsculas para los valores que adquieren.

Como casi todas las cosas en Estadística, hay un par de conceptos a tener en cuenta:

- Si el espacio muestral contiene un número finito de posibilidades, se lo llama *espacio muestral discreto*.
- Si el espacio muestral contiene un número infinito de posibilidades igual al número de puntos de un segmento de línea, se lo llama *espacio muestral continuo*.

Una variable aleatoria se la llama *discreta* si se puede contar su conjunto de resultados posibles, mientras que se la llama *continua* si se puede tomar en una escala continua.

Distribuciones de probabilidades. Distribución binomial

Es una de las distribuciones de probabilidad más útiles para las variables de tipo discretas. Está relacionada con el experimento aleatorio, el cual puede producir, en cada ensayo o prueba, uno de dos resultados posibles mutuamente excluyentes: ocurrencia de un evento, llamado éxito, y la no ocurrencia de éste, llamado fracaso. Por lo tanto, en binomial vamos a tener dos valores importantes con los cuales trabajar, el éxito y el fracaso.

Se pueden definir varios criterios para este tipo de distribución, a saber:

1. El experimento aleatorio consiste en *n* ensayos o pruebas repetidas, e idénticas, y fijadas antes del experimento. Son pruebas con reemplazamiento o con reposición.
2. Cada uno de los ensayos arroja solo uno de dos resultados posibles resultados: éxito o fracaso.
3. La probabilidad del llamado éxito, *p*, permanece constante para cada ensayo o prueba.
4. Cada prueba o ensayo se repite en idénticas condiciones y es independiente de las demás.
5. El interés recae en hallar la probabilidad de obtener *x* número de éxitos al realizar *n* ensayos del mismo experimento aleatorio.

La distribución binomial sigue una función de probabilidad, la cual es:

$$p(X=k)=\binom{n}{k}p^k \cdot q^{n-k}$$

En donde:

- *n* es el número de pruebas.
- *k* es el número de éxitos.
- *p* es la probabilidad de éxito.
- *q* es la probabilidad de fracaso.

Hay una tabla (de las tantas que hay en Estadística) para calcular las probabilidades cuando usamos Binomial. Eso nos salva de tener que usar la fórmula fea de más arriba (que sirve más que nada para enunciarla si la piden en un examen o si no tenemos la tabla para Binomial).

En resumen:

- Suceso simple binomial o dicotómico
- Parámetros: p, de valor constante, y un n fijo
- Suceso independiente
- E[x]= n.p y V[x]= n.p.q

Y si tengo que calcular p?

- x = μ = E[x] = n.p
- x= n.p
- x/n = p

Distribuciones de probabilidades. Distribución de Poisson

Esta distribución es una de las más importantes distribuciones de variable discreta. Una de sus principales aplicaciones hace referencia a la modelización de situaciones en las que interesa determinar el número de hechos de cierto tipo que pueden ser producidos en un intervalo de tiempo o espacio.

La distribución de Poisson sigue este modelo:

$$P(x=k) = e^{-\lambda} * \frac{\lambda^k}{k!}$$

Siendo:

- λ= n.p (o sea, el número de veces *n* que se realiza el experimento multiplicado por la probabilidad *p* de éxito en cada ensayo)
- k: el número de éxito cuya probabilidad se está calculando.

A no perder la calma… Para eso está la tabla de probabilidades de Poisson. Para usarla, necesitamos conocer λ. Esta fórmula, al igual que la de Binomial, es importante conocerla por si no tenemos tabla a mano (y por si piden enunciarla en un examen parcial o final).

La distribución de Poisson tiene algunas propiedades:

- Esperanza: *E[x]* = λ.
- Varianza: *V[x]* = λ.

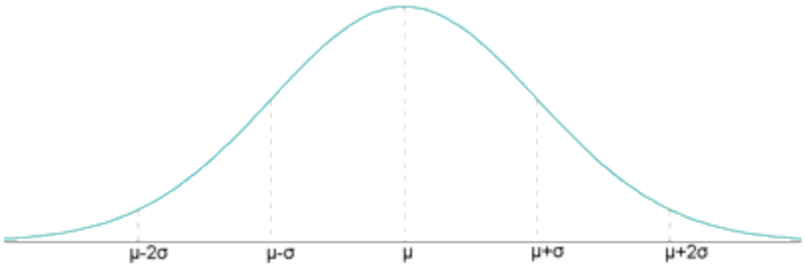
Si prestamos atención, tanto la Esperanza como la Varianza son iguales (iguales a λ). λ (Lambda) es el *parámetro* que se usa en Poisson.

A diferencia de la Distribución Binomial, n no es fijo y tiende a infinito.
A medida que λ aumenta, la distribución tiende a hacerse simétrica y acampanada.

Distribuciones de probabilidades. Distribución Normal (Z)

A partir de acá, tengo un espacio muestral dado por una variable continua. Las variables discretas las dejamos atrás con Poisson.

Una distribución normal de media μ (Mu) y desvío típico σ (Sigma) se designa por N(μ, σ). Su gráfica es la campana de Gauss:



La probabilidad que queremos averiguar es igual al área encerrada bajo la curva. Si abarca toda la curva es igual a 1; de no hacerlo, varía entre 0 y 1.

Distribución Normal estándar

La distribución normal estándar es la que presenta una media de valor cero (μ =0) y por desvío típico la unidad (σ =1).
La probabilidad de la variable x va a depender del área de la curva sombreada en la figura que realicemos. Sí, es conveniente que dibujemos la campana de Gauss para poner los datos y de esa forma calcular mejor la probabilidad. Y para calcularla usamos la tabla correspondiente para esta distribución (Tabla Z).

Como dijimos, la curva normal tiene forma de campana y un solo pico en el centro de la distribución. De esta forma, tanto la media aritmética como la mediana y la moda de la distribución son iguales y se localizan en ese mismo pico. Por tanto, la mitad del área bajo la curva se encuentra a la derecha de este pico y la otra mitad está a la izquierda del mismo. Esto se debe, recordemos, a que la curva es simétrica (y esto es algo muy importante).
La curva normal es asintótica, lo que significa que se acerca cada vez más al eje X pero nunca llega a tocarlo.

Para resolver problemas, se utiliza la *distribución normal estándar*, de forma que todas las distribuciones normales pueden convertirse a la estándar, restando la media de cada observación y dividiendo por la desviación estándar.
Entonces, lo que primero tenemos que hacer es convertir la distribución real en una distribución normal estándar utilizando un valor llamado Z, que será la distancia entre un valor seleccionado, designado X, y la media μ, dividida por la desviación estándar σ.
Así, vamos a tener esta fórmula para utilizar:

$$Z = \frac{X - \mu}{\sigma}$$

Con esta fórmula, la variable aleatoria se distribuye según una normal de media 0 y desviación estándar 1, que es la distribución llamada *normal estándar*. El valor que obtengamos va a ser un valor de Z el cual, si vamos a la debida tabla (Tabla Z, para recordar), va a tener ligado un valor de probabilidad al mismo.

Es importante recordar que, para utilizar Z, tenemos que valernos de alguno de estos datos:

- Un n mayor a 30
- Conocer los parámetros μ y σ.

A veces, **sólo a veces**, se pueden considerar como μ y σ a los valores de media y desvío estándar que ya hallamos calculado previamente. O sea, o nos lo dice el enunciado o en algún punto del ejercicio avisamos que los vamos a considerar como tales.

Distribuciones de probabilidades. Distribución de Student (t)

La *distribución de Student*, t, es otra de las distribuciones con las que vamos a trabajar para las variables cuantitativas continuas. Esta distribución es leptocúrtica, con valores que se concentran en torno a la media, pero a medida que aumentan los grados de libertad (υ) la forma de la distribución *t* tiende a ser similar a la de la distribución normal y para un $\upsilon=\infty$, ambas distribuciones son semejantes.

Esta distribución también cuenta con una tabla para usar, cosa que facilita mucho el resolver los ejercicios. Para usarla, es necesario conocer los grados de libertad (υ) y el valor de α.

¿Qué son υ y α?

Tratando de ser lo más sencillo posible, *α* es el *nivel de significación* y es parte del *nivel de confianza* (1-α) con el que vamos a trabajar varias veces, en especial en la segunda parte del año, y es un valor (generalmente, entre el 1-5%) que nos indica qué tan probable es que el parámetro, o estadístico, caiga dentro del *intervalo de confianza* (o *límite de confianza*, ya lo vamos a ver enseguida).

Los *grados de libertad*, por su parte, se refieren al número de observaciones de una muestra que pueden tomar cualquier valor posible, una vez que ya hayamos calculado, previa e independiente, la estimación de un determinado parámetro en la muestra, o población, de origen. Su importancia radica en que los estimadores siguen distribuciones de frecuencias específicas, cuya forma depende del número de grados de libertad asociados con su estimación. Mientras más grande sea el número de grados de libertad, más estrecha será la distribución de frecuencias y, por lo tanto, mayor será la potencia del estudio para realizar la estimación. Calcular el número de grados de libertad suele ser algo sencillo, pero es diferente según la prueba que usemos. El caso más fácil sería el del cálculo de la media de una muestra, que sería igual a *n-1*. De forma similar, cuando hay dos muestras y dos medias, el número de grados de libertad es *[(n1+n2)-2]*. En general, cuando se calculan dos o más parámetros (esto lo vamos a ver más adelante, en otros temas), los grados de libertad se calculan como *n-k-1*, siendo *k* el número de parámetros a estimar.

Unidad III. Estadística inferencial. Estimación de parámetros. Diseño muestral. Límites de confianza. Prueba de hipótesis.

Estadística inferencial

Requiere que, a partir del universo de estudio que tengamos, realicemos un muestreo del que podamos obtener, valga la redundancia, muestras con las que, a partir de la estadística descriptiva, podamos obtener elementos estadísticos. Es a partir de esos estadísticos que vamos a tener un conocimiento de la muestra y vamos a poder inferir comportamientos de la totalidad de ese universo de estudio (o de la población), manejándonos siempre con niveles de confianza menores al 100%.

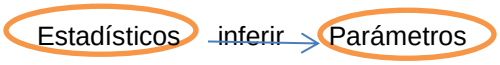
Hay dos grandes campos de trabajo:

- Estimación de parámetros.
- Test, o pruebas, de hipótesis.

En la estimación de parámetros vamos a tener:

- Parámetros, valores referidos a la población (o universo de estudio), simbolizados con letras griegas.
- Estadísticos, valores referidos al muestro que realicemos.

Así:



Estadístico	Parámetro	¿Qué representa?
x	μ	*Media aritmética*
S	σ	*Desvío estándar*
S^2	σ^2	*Varianza*
p	P o Π (pi)	*Proporción*

La estimación puede ser:

- Puntual.
- De intervalos.

En la distribución Normal, tengo dos parámetros (μ y σ), una variable continua y un n>30.

Si desconozco los parámetros, o tengo un n<30 (tal vez ambas cosas), uso la distribución de Student.

Estimación de parámetros

Estimar parámetros consiste en elegir un valor que represente el parámetro poblacional. Existen dos opciones para esto, que se conocen como *estimación puntual* y *estimación por intervalos*.

Primero, vamos a llamar, de forma genérica, a θ (theta) como el parámetro y a $\hat{\theta}$ como el estimador.

Estimación puntual

En la estimación puntual se utiliza el valor de un estadístico de muestra para inferir el parámetro poblacional, por ejemplo la media y la varianza para estimar μ y σ^2, respectivamente.

El estimador, $\hat{\theta}$, es un estimador puntual de θ ya que lo hace en la recta de los números reales y debemos comprender que no cualquier $\hat{\theta}$ es un buen estimador de θ. Para que lo sea, se tienen que cumplir ciertos requisitos:

- Insesgabilidad.

- Consistencia.
- Eficiencia.
- Presentar una varianza mínima.

Definirlos a cada uno sería un tanto tedioso pero, nuevamente, es importante que conozcan al menos que se deben cumplir esas cosas para que la estimación sea de tipo puntual.

Estimación por intervalos

Ya que pocas veces las estimaciones puntuales coinciden con los parámetros poblacionales, es preferible determinar un rango dentro del cual se encuentre el valor del parámetro que se va a estimar.

Vamos a dar una definición matemática de intervalo de confianza. Es así:

$$P[Li \leq \theta \leq Ls] = 1 - \alpha$$

Siendo:

- [Li, Ls]: intervalo que recibe el nombre de intervalo de confianza del 100(1 - α) % para el parámetro desconocido θ.
- Li; Ls: son, respectiva los límites de confianza inferior y superior.
- (1 - α): es el nivel de confianza asociado a este intervalo.

Este concepto de intervalo está así definido en el libro de Marta. Lo mencioné acá porque es necesario y, además, se relaciona con el concepto de *límites de confianza* que ya vamos a ver.

Diseño muestral

Para empezar, es importante que definamos, y tengamos en claro, qué es una *muestra*. Es una parte que representa en algo a la población de estudio que nos interesa. Por lo tanto, debe contener las características que la identifican con esa población. Mientas más grande sea esa población (o universo), más representativa debe ser la muestra y, por ende, de mayor confiabilidad. Ahora bien, el diseño muestral contiene todos los elementos que se contemplan para obtener una muestra representativa, evitando en lo posible errores, que pueden ser de dos tipos:

- *Errores de muestreo*: tamaño de la muestra y/o tipo de muestreo.
- *Errores de no muestreo*: diseño, capacitación y/o recursos.
- *Tamaño de la muestra*: tamaño de la población y/o grado de confianza.

Tipos de diseño muestral

- *Muestreo al azar aleatorio probabilístico*: considera que todos los elementos de la población tienen la misma probabilidad de ser elegidos como parte de la muestra. Ejemplo: las primeras 50 rocas que encontremos en la zona de estudio en la que estemos trabajando.
- *Muestro al azar aleatorio ordenado*: considera que todos los elementos de la población tienen la misma probabilidad de ser elegidos como parte de la muestra, pero con un orden. Ejemplo: cada 5 rocas, elegimos una y descartamos las otras cuatro.
- *Muestreo al azar aleatorio ordenado sistematizado*: todos los elementos tienen la misma probabilidad de ser elegidos, pero con un orden y un sistema.
- *Muestreo por conglomerado o agrupamiento*: todos los elementos tienen la misma probabilidad de ser elegidos siguiendo un orden y un sistema, pero es correspondiente o exclusivo de un tipo de población.
- *Muestreo estratificado*: sigue un orden, sistema y agrupamiento, pero afirma que la población no es homogénea sino heterogénea, por lo tanto se deben considerar en la muestra estratos de la población. Ejemplo: Feldespato/Plagioclasa, grado de alteración.
- *Muestreo combinado*: sigue orden, sistema, agrupamiento, y estratos, pero es el muestreo que nosotros nos proponemos realizar pero con previa justificación.

Límites de confianza

Un *límite de confianza* es un rango de valores que derivan de los estadísticos, y que posiblemente incluya el valor de un parámetro desconocido. Dado que presenta una naturaleza aleatoria, es poco probable que dos muestras de una población en particular generen intervalos de confianza idénticos.

El cálculo de estos límites, al estimar parámetros, nos permite hacer afirmaciones acerca de qué valores podemos esperar para esos parámetros que estamos estimando.

Esos límites van a depender de:

- El parámetro que estemos estimando (media, desvío estándar, etc.).
- El tamaño de la muestra. Mientras mayor sea el n, menor va a ser la diferencia entre el parámetro estimado y el real.
- El nivel de confianza con el que estemos trabajando.

Unidad IV. Prueba de hipótesis.

Prueba de hipótesis

Implica que uno puede plantearse preguntas con respecto a la muestra.

La verificación de la hipótesis conlleva información cuali- y cuantitativa.

El planteo y la prueba de hipótesis nos sirven para conocer el comportamiento semejante o diferente entre las muestras que tengamos, trabajando con niveles de confianza con valores que son, por lo general, entre un 95-99%.

En estadística, trabajamos con una hipótesis geológica, que es el problema que nos planteamos al, por ejemplo, observar una situación, y con un par de hipótesis estadísticas, llamadas hipótesis nula (H_0) y alternativa (H_1). Generalmente, luego de la hipótesis geológica, que se plantea con palabras, nos planteamos la hipótesis nula, que es la que vamos a poner a prueba, y luego la hipótesis alternativa, que es similar a la nula, pero plantea la situación de forma opuesta. El par de hipótesis estadísticas, al contrario de la geológica, se expresa con símbolos y no con palabras.

Entonces, con la hipótesis estadística sometemos a verdadero o falso una situación, mientras que con la prueba de hipótesis realizamos un procedimiento metodológico que nos lleva a aceptar, o rechazar, nuestra hipótesis estadística.

Algo importante **(muy importante**) es que la hipótesis nula es la que siempre (**¡SIEMPRE!**) lleva el signo igual al plantearla. Por ejemplo:

$$H_0: \mu = \mu_0 \; ; \; H_0: \mu \geq \mu_0 \; ; \; H_0: \mu \leq \mu_0$$

Si tenemos el signo igual solo, la prueba que vayamos a hacer va a ser *bilateral* (en la curva marcamos un valor crítico en ambas colas) y el nivel de significancia va a ser $\alpha/2$. Para cualquiera de los otros dos casos, la prueba va a ser *unilateral* (marcamos el valor crítico en una sola cola) y el nivel de significancia va a ser α; dependiendo de si usamos el signo mayor o menor, la prueba va a ser *unilateral derecha* o *unilateral izquierda* (marcamos el valor crítico en la cola de la derecha o en la de la izquierda).

Para interpretar correctamente el valor del test estadístico que utilicemos al someter a prueba la hipótesis nula, es necesario utilizar dos conceptos nuevos: *zona de rechazo* y *valor crítico*. La región de rechazo especifica los valores del test estadístico para los cuales la hipótesis nula es rechazada (por lo tanto, valores para los cuales la hipótesis alternativa es aceptada). Para poder saber si un valor que obtuvimos cae, o no, en la zona de rechazo, es necesario conocer el valor crítico, el cual surge del valor de α con el que estemos trabajando junto con el tipo de test que usemos.

La zona de rechazo identifica los valores del test estadístico que sostienen a la hipótesis alternativa y serían improbables de ser verdadera la hipótesis nula. Por consiguiente, los valores que nos lleven a aceptar, o considerar verdadera, a la hipótesis nula caen en un sector llamado *zona de no rechazo*.

¿Y dónde ubico estas zonas? En la curva que usamos para la distribución Normal, Student, Chi cuadrado, etc. Van a depender mucho del tipo de prueba (si es bi- o unilateral).

Ya definidas las zonas de rechazo y no rechazo, el valor crítico y nivel de significancia (α), vamos a definir estas cosas:

- *Estadístico de prueba*: es una variable aleatoria, con una distribución de probabilidad conocida, y cuyos valores permiten tomar la decisión de aceptar, o rechazar, la hipótesis nula. O sea, es un valor con el cual vamos a comparar el resultado que obtengamos del test.
- *Decisión estadística*: es aceptar, o rechazar, la hipótesis nula a partir de la comparación del estadístico de prueba con el valor calculado.

Errores que se pueden cometer

Son dos:

- *Tipo 1*: aceptar la hipótesis nula como válida cuando las diferencias encontradas son realmente significativas.
- *Tipo 2*: rechazar la hipótesis nula cuando las diferencias no son realmente significativas.

Pruebas paramétricas

Básicamente, si tenemos una variable aleatoria con una determinada distribución, podemos establecer afirmaciones sobre los parámetros de dicha distribución. Es importante saber esto ahora porque tenemos también pruebas no paramétricas, pero esas las vemos en un par de páginas más.

Ejemplo de pruebas paramétricas son Z (normal), t (Student), P (de proporciones).

Pruebas no paramétricas

Con este nombre se incluyen una variedad de pruebas que se emplean en el análisis de variables nominales como de variables que se expresan en escalas ordinales. También se usan para datos medidos en escalas de intervalo o de razón donde la función de distribución de la variable aleatoria no está especificada.

Las pruebas no paramétricas tienen sus ventajas, como también tienen sus desventajas. Se pueden usar cuando se desconoce la distribución de probabilidad la población, brindan respuestas rápidas con pocos cálculos. Por otra parte, si los datos son continuos pero existen dudas del cumplimiento de los supuestos requeridos por las pruebas paramétricas, permiten hacer inferencias sobre los parámetros poblacionales. Aun así, no usan toda la información que está disponible y, al no haber parámetros, es difícil hacer estimaciones cuantitativas; además, son un tanto menos eficientes puesto que, para rechazar la hipótesis nula con el mismo nivel de confianza (que en las pruebas paramétricas) se necesitan muestras mayores.

Ejemplos de pruebas no paramétricas son X^2 (Chi cuadrado), U (Mann-Whitney), H (Kruskal-Wallis), r (Spearman).

Dicho todo esto, les dejo un cuadro de síntesis para dos casos de pruebas de hipótesis. Uno es si tenemos una muestra para estudiar y el otro si tenemos dos muestras. A partir de ahí vemos qué datos tenemos para usar (varianza, media, proporciones, etc.) y después arrancamos con el enunciado del par de hipótesis estadísticas. A continuación del cuadro, explico alguno tests para ver cómo usarlos.

PRUEBAS PARA UNA MUESTRA			
	Hipótesis	**Estadístico de prueba**	**Criterio de rechazo**
Prueba para una media. Varianza Poblacional conocida y/o n > 30. μ_0= un número fijo conocido	Ho: $\mu = \mu_0$ H1: $\mu \neq \mu_0$.	$z_0 = \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$	$\lvert z_0 \rvert > z_{\alpha/2}$ Bilateral
	Ho: $\mu \geq \mu_0$ H1: $\mu < \mu_0$		$z_0 < -z_\alpha$ Unilateral izquierda
	Ho: $\mu \leq \mu_0$ H1: $\mu > \mu_0$		$z_0 > z_\alpha$ Unilateral derecha
Prueba para una media. Varianza Poblacional desconocida y/o n < 30	Ho: $\mu = \mu_0$ H1: $\mu \neq \mu_0$.	$t_0 = \frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{n}}}$	$\lvert t_0 \rvert > t_{\frac{\alpha}{2};\nu}$ Bilateral
	Ho: $\mu \geq \mu_0$ H1: $\mu < \mu_0$		$t_0 < -t_{\alpha;\nu}$ Unilateral izquierda
	Ho: $\mu \leq \mu_0$ H1: $\mu > \mu_0$		$t_0 > t_{\alpha;\nu}$ Unilateral derecha
Prueba de Varianza	Ho: $\sigma^2 = \sigma_0^2$ H1: $\sigma^2 \neq \sigma_0^2$.	$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$	$\chi^2 \leq \chi^2_{1-\alpha/2,n-1}$ ó $\chi^2 \geq \chi^2_{\alpha/2,n-1}$ Bilateral
	Ho: $\sigma^2 \geq \sigma_0^2$ H1: $\sigma^2 < \sigma_0^2$.		$\chi^2 \geq \chi^2_{\alpha/2,n-1}$ Unilateral izquierda
	Ho: $\sigma^2 \leq \sigma_0^2$ H1: $\sigma^2 > \sigma_0^2$.		$\chi^2 \leq \chi^2_{1-\alpha/2,n-1}$ Unilateral derecha

PRUEBAS PARA DOS MUESTRAS			
Prueba para comparar Varianzas	Ho: $\sigma_1^2 = \sigma_2^2$ H1: $\sigma_1^2 \neq \sigma_2^2$.	$F = \frac{S_1^2}{S_2^2}$	$F \leq F_{1-\alpha 2, n1-1, n2-2}$.ó $F \geq F_{\alpha, n1-1, n2-2}$ Bilateral
	Ho: $\sigma_1^2 \geq \sigma_2^2$ H1: $\sigma_1^2 < \sigma_2^2$.		$F \leq F_{\alpha, n1-1, n2-2}$ Unilateral izquierda
	Ho: $\sigma_1^2 \leq \sigma_2^2$ H1: $\sigma_1^2 > \sigma_2^2$.		$F \geq F_{\alpha, n1-1, n2-2}$ Unilateral derecha
Prueba de diferencia de medias. Varianzas Poblacionales conocidas	Ho: $\mu_1 = \mu_2$ H1: $\mu_1 \neq \mu_2$	$z_0 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$\lvert z_0 \rvert > z_{\alpha/2}$ Bilateral
	Ho: $\mu_1 \geq \mu_2$ H1: $\mu_1 < \mu_2$		$z_0 < -z_\alpha$ Unilateral izquierda
	Ho: $\mu_1 \leq \mu_2$ H1: $\mu_1 > \mu_2$		$z_0 > z_\alpha$ Unilateral derecha
Prueba de diferencia de medias. Varianzas Poblacionales desconocidas y estimadas a partir de las varianzas muestrales	Ho: $\mu_1 = \mu_2$ H1: $\mu_1 \neq \mu_2$	$t_0 = \frac{\overline{X}_1 - \overline{X}_2}{S_{\Delta \overline{X}}}$ Recordar que previamente debe calcularse la prueba de homogeneidad de varianzas para utilizar $S_{\Delta \bar{X}}$ que corresponda[1]	$\lvert t_0 \rvert > t_{\frac{\alpha}{2}, \nu}$ Bilateral
	Ho: $\mu_1 \geq \mu_2$ H1: $\mu_1 < \mu_2$		$t_0 < -t_{\alpha;\nu}$ Unilateral izquierda
	Ho: $\mu_1 \leq \mu_2$ H1: $\mu_1 > \mu_2$		$t_0 > t_{\alpha;\nu}$ Unilateral derecha
Prueba de Muestras apareadas	Ho: $\mu_1 = \mu_2$ H1: $\mu_1 \neq \mu_2$	$t_0 = \frac{\overline{X}_d - \mu_{\mu d}}{S_{\mu d}}$	$\lvert t_0 \rvert > t_{\frac{\alpha}{2}, \nu}$ Bilateral
	Ho: $\mu_1 \geq \mu_2$ H1: $\mu_1 < \mu_2$		$t_0 < -t_{\alpha;\nu}$ Unilateral izquierda
	Ho: $\mu_1 \leq \mu_2$ H1: $\mu_1 > \mu_2$		$t_0 > t_{\alpha;\nu}$ Unilateral derecha
Prueba de diferencia Proporciones	Ho: $\pi_1 = \pi_2$ H1: $\pi_1 \neq \pi_2$	$Z_0 = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$	$\lvert z_0 \rvert > z_{\alpha/2}$

			Bilateral
	Ho: $\pi_1 \geq \pi_2$ H1: $\pi_1 < \pi_2$		$z_0 < -z_\alpha$ Unilateral izquierda
	Ho: $\pi_1 \leq \pi_2$ H1: $\pi_1 > \pi_2$		$z_0 > z_\alpha$ Unilateral derecha

Si $\sigma_1^2 = \sigma_2^2$, $S_{\Delta\bar{X}}$ se calcula con $S_{\Delta\bar{X}} = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$, cuando $\sigma_1^2 \neq \sigma_2^2$ $S_{\Delta\bar{X}} = \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$

Pasos para contrastar hipótesis

1. Enunciado del par de hipótesis estadística.
2. Elección del nivel de confianza (1-α).
3. Selección del estadístico de prueba.
4. Determinación de la región crítica.
5. Cálculo del estadístico elegido.
6. Decisión estadística.
7. Conclusiones tomadas a partir del problema y la decisión estadística.

Pruebas de hipótesis de las varianzas de dos poblaciones normales

Cuando se trata de comparar las varianzas, se utiliza la variable $F=S^2_1/S^2_2$, que está relacionada con la distribución F, con n_1-1 y n_2-1 como grados de libertad. Generalmente, se recomienda colocar siempre en el numerador la varianza muestral asociada a la variancia poblacional mayor.

Pruebas de hipótesis para medias poblacionales

Los supuestos que se deben cumplir son que las medias poblacionales sean normales, los desvíos poblacionales conocidos y las muestras independientes. Esto es para desvíos poblacionales conocidos (**Importante!**).

Para el caso en el que los desvíos poblacionales sean desconocidos, los supuestos que se deben cumplir son que los datos deben ser extraídos de dos muestras aleatorias independientes, de tamaño n_1 y n_2 respectivamente. Las varianzas poblacionales, si bien no se conocen, se suponen que son iguales. Primero se debería comprobar la igualdad de dichas varianzas, en particular si los tamaños de las muestras son distintos, con la prueba de F. De ser estadísticamente iguales, aplicamos el estadístico de la tabla. Si los desvíos se suponen diferentes, los supuestos (valga la redundancia) que hay que tener en cuenta son los mismos que mencioné antes, sólo cambia la parte inferior del estadístico de prueba a usar (detallado en la tabla de más arriba).

Pruebas de hipótesis para muestras apareadas

Surge cuando cada observación para un tratamiento está apareada con otra observación para el otro tratamiento. Este par está compuesto por las mismas unidades experimentales observadas dos veces en distintos momentos de la investigación, o por unidades semejantes.

La hipótesis nula que se plantea es que la media de las diferencias es igual/mayor o igual/menor o igual a cero. Cualquiera de esas tres opciones.

Como se establece una hipótesis de un único parámetro poblacional (se podría pensar en una sola muestra), el número de grados de libertad es (n - 1).

La media de las diferencias se obtiene luego de que, para cada par de muestras, hacés la diferencia de cada uno de esos pares. O sea, para cada par de datos que tengas en la tabla, los restás y vas a obtener una nueva columna con el valor que obtengas de cada resta. Por ejemplo:

X	Y	Diferencia
5	3	2
4	2	2
9	2	7
3	3	0
4	5	-1
12	7	5
2	1	1
8	6	2

Cada valor de X está relacionado con un correspondiente valor de Y. Los resto y obtengo un valor que pongo en la tercera columna (*Diferencia*)

Pruebas de hipótesis para proporciones

En este caso, proporciones hace referencia a valores en porcentaje. Por lo tanto, vamos a necesitar de estos valores para poder trabajar.

Por lo general, esos valores de porcentaje ya vienen dados en el enunciado. Muchas veces, puede pasar que esos valores estén comprendidos entre 0% y 100%. Nosotros vamos a tener que pasarlos a sus correspondientes valores comprendidos entre 0 y 1. Por qué? Porque estamos en Estadística, y los valores de porcentaje (**en este caso!**) tienen que estar entre 0 y 1.

Unidad V. Pruebas paramétricas. ANOVA. Correlación. Regresión lineal.

ANOVA

Se utiliza para detectar la existencia de diferencias significativas entre las medias de una determinada variable cuantitativa en tres o más grupos de datos. Grupos, poblaciones, tratamientos… De acuerdo a la bibliografía que consulten, pueden encontrar alguna de estas palabras. Significan lo mismo; por lo general, yo uso 'tratamientos'.

Como vimos antes, se puede utilizar la distribución t para contrastar la hipótesis de que no existen diferencias entre las medias de dos poblaciones (allá por el tema de pruebas de hipótesis, en el práctico). Supongamos que ahora queremos averiguar algo similar, pero teniendo tres o más poblaciones diferentes, podríamos usar el test de t para hacer todas las comparaciones posibles entre pares de medias de las poblaciones con las que estemos trabajando (3, 4 ó más). Si intentamos hacer eso, no sería muy conveniente por las siguientes razones:

- Resultaría tedioso comparar todas las posibles combinaciones de medias.
- Cualquier estadístico basado en parte de la evidencia (como ocurre cuando solo se comparan dos grupos) es menos estable que uno basado en toda la evidencia.
- Si se hacen muchas comparaciones aumenta la probabilidad de que alguna resulte significativa.
- Aumentamos la probabilidad de cometer algún error (de tipo 1 ó de tipo 2).

Por lo tanto, vamos a necesitar usar ANOVA para que nos indique si alguno de los tratamientos presenta una diferencia significativa con respecto a los demás.

ANOVA y su lógica

El contraste de hipótesis del ANOVA se basa en comprobar si las medias de las muestras difieren más de lo que cabe esperar de ser cierta la hipótesis nula. Esta cuestión acerca de las medias se responde analizando las varianzas. Nos centramos en las varianzas dado que, cuando queremos saber si algunas medias difieren entre sí, tenemos que valorar la varianza entre estas medias.

En ANOVA, un estimador de la variabilidad *entre grupos* se compara con la variabilidad *dentro de los grupos*. La variación entre grupos es la variación entre las medias de los diferentes tratamientos debidas al azar (error de muestreo) y al efecto de los tratamientos, si es que existe. Por su parte, la variación dentro de los grupos es la variación debida al azar (error de muestreo) entre individuos a los que se ha dado el mismo tratamiento. Esta comparación de variabilidades se realiza mediante cálculos, los cuales vamos a volcar luego en una tabla, para así llegar a conseguir el valor del estadístico F. Este estadístico, F, es el que vamos a comparar con otro F (el valor crítico, que lo obtenemos de tabla). A partir de esa comparación, es que vamos a aceptar (o rechazar) la hipótesis nula. De rechazarla, se aplica un test de comparación más, llamado *test de Tukey*, para verificar cuál de los tratamientos con los que estemos trabajando sea el que presente la diferencia más significativa.

Supuestos

- Datos distribuidos según una distribución normal.
- Las varianzas de las distintas muestras han de ser iguales.
- Muestras independientes y tomadas al azar.

Planteo de hipótesis

Por lo general, yo suelo plantear el par de hipótesis de la siguiente forma:

$$H_0: \mu_i = \mu_j \qquad H_1: \mu_i \neq \mu_j$$

Siendo $i \neq j$, e (ij) pertenecen a la población

Lo hago así por una cuestión de prolijidad, ya que estamos comparando tres, o más, tratamientos (o poblaciones), que son (i; j). Como no sé cuál de esos tratamientos es el que difiere (el 1, el 2 o… No sé, el 4) es que uso las letras.

A veces, puede pasar que no tengamos la misma cantidad de datos entre los tratamientos. Esto se soluciona mediante un valor que se calcula aparte y luego se agrega a las fórmulas del ANOVA.

Cálculos en ANOVA

Con igual número de tratamientos

Supongamos que tenemos tres pozos de petróleo, con ocho datos cada uno, para estudiar y ver cuál es el que produce más. Sería algo así la tabla:

Pozo 1	Pozo 2	Pozo 3
x_1	y_1	z_1
x_2	y_2	z_2
x_3	y_3	z_3
⋮	⋮	⋮
x_8	y_8	z_8

Los pozos constituyen los tratamientos; por lo tanto, tenemos 3 tratamientos. De cada uno de ellos (los pozos) vamos a calcular la media y la varianza, para luego sacar una media total ($\overline{x_t}$) y una varianza total (S^2_t).

La tabla de ANOVA generalmente tiene esta forma:

	Suma de Cuadrados (SC)	Grados de Libertad (g.l)	Cuadrados Medios (CM)	F
Entre	$SC_E = n[\sum(\overline{x_i} - \overline{x_t})^2]$	K - 1	$SCE/g.l_E$	CM_E/CM_D
Dentro	$SC_D = SC_T - SC_E$	N - K	$SCD/g.l_D$	
Total	$SC_T = SC_E + SC_D$	N - 1		

Hay tres casilleros con líneas horizontales (1 en CM total y 2 en F). Eso es porque no necesitan cálculos ni datos. También hay letras en las fórmulas que llaman la atención, ahí las explico:

- n: número de datos que tengo por tratamiento (en este caso, 8)
- N: es el número total de datos que tengo. Son 8 por cada uno de los 3 tratamientos. Por tanto, N vale 24 para este ejemplo.
- k: es el número de tratamientos (3 en este caso).

La Suma de cuadrados entre es tomar la media que calcularon para el pozo 1, restarlo con la media total y a ese valor lo elevan al cuadrado. Con la media del pozo 2 y del pozo 3 hacen lo mismo. Una vez que hicieron ese paso, suman los 3 resultados. Les queda un valor único al que le tienen que multiplicar n.

Recuerdan que les dije de calcular la varianza total? Es para esto:

$$S^2_t = \frac{(SC_T)}{(g.l_T)}$$

Traducido, la varianza total es igual a la suma de cuadrados total, dividido los grados de libertad total. De ahí, despejan SC_T.

Con igual número de tratamientos

Cuando el n de los tratamientos es diferentes, armamos la tabla de ANOVA de esta forma:

Fuente de variación	Suma de Cuadrados (SC)	Grados de libertad (gl)	Cuadrado Medio (CM)	F
Entre Tratamientos	$\sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n} x_{ij}\right)^2}{n_i} - C$	k-1	$\frac{SCEntre}{k-1}$	$\frac{CMEntre}{CMDentro}$
Dentro de los Tratamientos	$SCTotal - SCEntre$	N-k	$\frac{SCDentro}{N-k}$	
Total	$\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - C$	N-1		

Tabla de ANOVA tomada del libro de Marta Alperin.

Las fórmulas parecen complicadas, pero no tanto:

- Para SC_E, lo primero que hacen es sumar los datos que tengan del primer tratamiento, lo elevan al cuadrado y lo dividen por la cantidad de datos que tiene ese tratamiento. Hacen lo mismo con los demás tratamientos y los resultados los suman. Una vez sumados, le restan el valor de C.
- Para SC_T, elevan al cuadrado cada valor que tengan en el primer tratamiento, le restan el valor de C y los van sumando. Con el tratamiento que siguen hacen lo mismo y con los demás tratamientos también. Una vez que obtuvieron esa sumatoria de 'valores al cuadrado menos C' por tratamiento, suman entre sí esos resultados.

Y C? De dónde lo saco?

Lo sacamos de esta fórmula:

$$C = \frac{\left(\sum_{j=1}^{k}\sum_{i=1}^{n} Xij\right)^2}{N}$$

El procedimiento es similar al de SC_T.

Test de Tukey

Si la hipótesis nula fue rechazada, tenemos que ver (para este ejemplo) qué pozo es el que produce más petróleo.

Para eso, planteamos este par de hipótesis:

$$H_0: \mu_i = \mu_j \;\; ; \;\; H_1: \mu_i \neq \mu_j$$

El estadístico de prueba se calcula así:

$$q_c = \frac{\bar{x}i - \bar{x}j}{SE}$$

Siendo SE:

$SE = \sqrt{(CM_D/n_i)}$ Si tengo el mismo número de datos en los tratamientos

$SE = [\sqrt{(CM_D/2)}] \times [\sqrt{(1/ni) + (1/nj)}]$ Si el número de datos varía para cada tratamiento

Entonces, se hace la diferencia entre medias, se la divide por SE y así obtienen q_c. Ese valor de q_c lo comparan con el de tabla y ven cuál es el pozo que produce más petróleo.

Correlación

La correlación nos permite plantear una relación existente entre dos variables, *x* e *y*, en la que, por lo general, x es la *variable independiente* e y es la *variable dependiente*. También nos sirve para medir la correlación, o la relación, existente entre dos muestras, predecir (o estimar) el comportamiento de la variable dependiente, en relación del conocimiento de la independiente.

Más allá de saber si existe una asociación entre las variables, lo importante es que esta puede ser medible por la "R" Pearson.

Generalmente, pero no siempre, este tema está relacionado con la regresión lineal, y suelen utilizarse casi en paralelo, como si fuesen un solo tema (**¡pero no lo son!**).

Regresión lineal

El análisis de *regresión* se utiliza para fines de predicción. Lo de lineal hace referencia al tipo de regresión con el que vamos a trabajar en clases, ya que existen otros tipos (logarítmica).

Generalmente, existen relaciones entre 2 ó más variables, por ejemplo, concentración de CO_2 y temperatura (ºC), OFe y Eh, etc. Suele ser deseable expresar tales relaciones en forma matemática determinando una ecuación que conecte a las variables.

Para hallar una ecuación que relacione estas variables, primero hay recoger datos que muestran valores correspondientes de las variables que tenemos en consideración. El próximo paso es marcar los puntos (x,y) en un sistema de coordenadas cartesianas, que se denomina *diagrama de dispersión*.

A veces, a partir de este diagrama es posible visualizar una curva que aproxima los datos. Puede tratarse de una recta, una parábola o una ecuación cúbica, por ejemplo. La ecuación con la que más vamos a trabajar es la de la línea recta:

$$y = a.x + b$$

La regresión responde a tres objetivos:

- Estudiar si ambas variables están relacionadas
- Determinar qué tipo de relación, si existe, las une.
- Predecir los valores de una variable a partir de valores conocidos de la otra.

Conocer el grado de relación existente entre ambas variables, permitirá saber si la predicción realizada con el modelo matemático establecido, es buena o mala.

Para medir el grado de relación existente entre la variable independiente y la variable dependiente, lo que más se utiliza es el *coeficiente de correlación lineal* (*r* de Pearson). Hay una fórmula, muy fea, para calcular el valor de r:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}}$$

Por suerte, la calculadora se encarga de calcular el valor de r por nosotros. Así que recomiendo que busquen el manual de la misma (en algún rincón de sus casas o en internet) y vean cómo se calcula.

El valor de *r* se encuentra en el intervalo [-1; 1] y puede pasar alguno de estos casos:

- Si r es igual a 0, entonces no existe correlación entre las variables.
- Si r es igual a -1, entonces la correlación es perfecta y negativa.
- Si r es igual a 1, entonces la correlación es perfecta y positiva.
- Si r se encuentra en el intervalo (-0,5; 0,5), entonces la correlación es mala.
- Si r no se encuentra en el intervalo (-0,5; 0,5), entonces la correlación es buena.

Hay un ANOVA que se usa en Regresión para ver si el test (de Regresión) es significativo. El proceso es muy similar al ANOVA que ya expliqué.

UNIDAD VI. Bondad de ajuste. Tabla de contingencia

El análisis de frecuencias se realiza de dos formas: por *bondad de ajuste* o por *tablas de contingencia*.

En ambos casos, la hipótesis nula a tratar es:

$$H_0: f_o = f_e$$

O sea, esperás que la frecuencia observada sea igual a la frecuencia esperada.

El nivel de significancia (el de alfa) es, en general, 0,05. Como casi siempre (a menos que le pidan otro valor).

En ambos casos, también, hay que armar una tabla para comparar las frecuencias y así poder trabajar.

En bondad de ajuste, los grados de libertad van a depender de qué test estemos trabajando. Va a tomar esta forma:

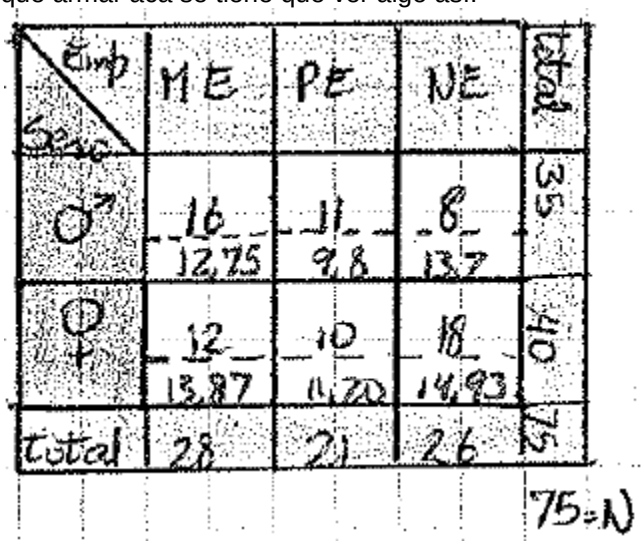
$$\upsilon = k - n - 1$$

Como ven, a la fórmula se agrega la letra k. Y qué valor toma? Depende del test, como mencioné antes. Si el análisis está relacionado a Poisson, k vale 1. Para binomial, vale 1. Para Normal, vale 2.

En tablas de contingencia, tengo dos variables cualitativas y mi duda es si son independientes. Los grados de libertad van a calcularse así:

$$\upsilon = (\text{nº de filas} - 1) \times (\text{nº de columnas} - 1)$$

La tabla que tengo que armar acá se tiene que ver algo así:



Ahora, el estadístico de prueba es el mismo para ambos casos:

$$\chi^2 = \sum[(F_o - F_e) : (F_e)]$$

Esto es: tomar cada frecuencia observada, restarle su respectiva frecuencia esperada y dividirlo por ese mismo valor de frecuencia esperada. Cada valor que obtengamos de esta resta y división, lo sumamos con el otro (por eso el símbolo de sumatoria).

El valor crítico (el que saco de la tabla) es el mismo a usar en ambos casos. Hay que agarrar la tabla de Chi cuadrado y buscar el valor de alfa y de grados de libertad y, a partir de eso, saco el valor crítico.

Una vez que ya tengo el estadístico de prueba y el valor crítico, los comparo, veo si acepto la hipótesis nula y saco conclusiones.

UNIDAD VII. Pruebas no paramétricas. Chi cuadrada. Mann-Whitney (U). Kruskal-Wallis. Spearman (r)

Bueno, una pequeña introducción a esta unidad…

En la unidad anterior vimos pruebas paramétricas. En ellas hay supuestos que se deben cumplir antes de ponernos a realizar cualquiera de esas pruebas. Cada una de ellas tiene sus propios supuestos, por lo tanto los supuestos de, por ejemplo, ANOVA, no van a ser los mismos que los de Regresión lineal.

Si queremos utilizar alguna de esas pruebas paramétricas y no llega a cumplirse al menos uno de los supuestos que plantea, debemos elegir una prueba no paramétrica que se adecue a la que hallamos querido usar.

Chi cuadrada (χ^2)

Es una prueba que solo requiere que los elementos que integran las muestras contengan alguna característica en común.

Sus características son:

- No requiere que las muestras sean de gran tamaño.
- No requiere muestras de igual tamaño.
- Se debe plantear desde el principio la hipótesis nula con su grado de confianza.

Mann-Whitney (U)

El test de Mann-Whitney, también llamado test de U (por su estadístico de prueba), compara las diferencias entre dos medianas, por lo que se basa en rangos en lugar de en los parámetros (media, varianza). Lo usamos cuando los datos no siguen la distribución normal, en lugar del test de Student.

Este test presenta dos requisitos:

- Variable cuantitativa que no cumple los requisitos de normalidad y/o homogeneidad de varianzas, o variable semicuantitativa.
- Muestras independientes y al azar.

Procedimiento de cálculo

- Asignación de rangos a cada dato. Para ello se ordenan todos los datos (juntando los dos grupos) en orden creciente. El rango de cada dato será el número de orden que le corresponde a cada dato. Cuando se repita el mismo valor numérico, el rango que se asigna a esos datos es la media aritmética de los rangos que les corresponderían en función del número de orden que ocupan.
- Se suman los rangos de cada uno de los inventarios (grupos) y se calcula la suma de los rangos de los datos de cada uno de los grupos (R_1 y R_2)
- Se calculan los estadísticos U_1 y U_2 a partir de las siguientes fórmulas:

$$U_1 = n_1 \cdot n_2 + \frac{n_2(n_2+1)}{2} - R_2 \qquad U_2 = n_1 \cdot n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

- Se obtiene el estadístico U_{cal} escogiendo el valor más grande entre U_1 y U_2.
- Se comprueba la significación estadística del estadístico U_{cal} comparando este valor con el valor crítico (U_{crit}) obtenido a partir de las tablas correspondientes.
  - Si $U_{cal} \geq$ Ucrít (α=0.05 o inferior), se rechaza H_0 y, por lo tanto, las medianas son diferentes.
  - Si U_{cal} < Ucrít (α =0.05), se acepta H0 y, por lo tanto, las medianas son iguales.

Kruskal-Wallis (H)

Se basa en rangos en lugar de los parámetros (media, varianza). Se usa cuando los datos no siguen la distribución normal y/o tienen varianzas distintas, sustituyendo así al ANOVA. Cuando el número de grupos que tengo es dos, es idéntico a la U de Mann-Whitney.

El estadístico de prueba se denomina H.

Al igual que Mann-Whitney, esta prueba presenta algunos requisitos:

- Variable cuantitativa que no cumple los requisitos de normalidad y/o homogeneidad de varianzas, o variable semicuantitativa.
- Muestras independientes y al azar.

Procedimiento del cálculo

- La asignación de rangos se realiza de igual manera que en Mann-Whitney.
- Cálculo del estadístico H:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

Siendo k el número de grupos, N el número total de datos y n_i el número de datos en el grupo i-ésimo.

- Puede suceder que tengamos *rangos ligados*. Esto significa que podemos tener dos o más números con el mismo rango, por lo que vamos a aplicar un factor de corrección:

$$H_c = \frac{H}{C} \qquad C = 1 - \frac{\sum_{i=1}^{m} (t_i^3 - t_i)}{N^3 - N}$$

Siendo H_c el estadístico que se utiliza en lugar de H

- El valor crítico del estadístico calculado (H o Hc) se consulta en la tabla de Chi cuadrada si N≥15, o si $k > 5$, para (k-1) grados de libertad. Si N<15 y k<5 se consulta en la tabla específica para H.
  - Si $H_{cal} \geq H_{crit}$ ($\chi^2_{crít}$), se rechaza H_0 y se acepta H_1 (alguna de las medianas es diferente).
  - Si $H_{cal} < H_{crit}$ ($\chi^2_{crít}$), se acepta H_0 y se rechaza H_1 (las medianas son iguales).

Spearman (r)

Bien, esta prueba se usa cuando no se cumplen los supuestos para correlación.
Para calcular el valor de r, hay que realizar los siguientes pasos:

- Ordenar los pares de datos en función del valor de x y asignar rangos a x.
- Repetir la ordenación en función de y. Luego asignar rangos a y.
- Calcular el coeficiente:

$$r_s = 1 - \frac{6 \sum_{i=1}^{i=n} d_i^2}{n^3 - n}$$

- Para comprobar la significación estadística, hay que consultar en la tabla correspondiente el valor crítico de r_s para n pares de datos, para un $\alpha \leq 0.05$ y para el número de colas acorde con la hipótesis. Si r_s cal ≥ $r_{s\ crit}$, se rechaza H_0.