Análisis de datos en Psicología

Asignatura Anual

Parte I: Introducción

Lenguaje matemático en Psicología

- Error: *es aquél componente de ignorancia que tiene el campo donde se actúa. El error tiene que ver con el tiempo de desarrollo de esa ciencia. En Psicología el error es bastante importante puesto que es una ciencia débil.*
- Modelo: *son simplicaciones formales de la realidad; en psicología, simplificaciones de la conducta humana. En Psicología hay modelos que están completamente verbalizados (como p.ej. el psicoanálisis). Se distinguen modelos cuantitativos y cualitativos.*
- Estabilidad estructural (Thon): *cuantitativamente los resultados son distintos, no obstante, a nivel general las investigaciones tienden a interpretar estos valores como iguales o equivalentes.*
- El método que utiliza la psicología es el método hipotético: a través de una hipótesis se experimenta y se sacan conclusiones, luego el proceso se replica si es necesario. Si el 100% de resultados apoyan la hipótesis, son Resultados Generales, si la mayoría se acertada, son Resultados Parciales (Estocásticos), si sólo son resultados ciertos en parte son Resultados Existenciales, con los que no se puede trabajar.
- En Psicología siempre se trabaja con datos de tipo estocástico. Los resultados pueden ser generales o parciales, y cualifica la teoría. Este modelo no es aceptado generalmente en Psicología porque no existe concenso entre los psicólogos.
- Metodología:

- Medida (asignación numérica de los hechos observados).
- Fiabilidad (consistencia).
- Validez (en qué medida sirve).

Teoría de la medición

- Existen 4 modelos formales que conviene distinguir, son lo que Stevens denomina Escalas de Medida:

- Escala Nominal. *Sólo funciona la relación igual y distinto. En ésta escala los números son meras etiquetas. Es la más pobre. Ej. el DNI.*
- Escala Ordinal. *Funciona la relación de igual, distinto, mayor y menor. Ej. nº de llegadas en una prueba atlética.*
- Escala de Intervalo. *Funciona de igual manera que la Escala Ordinal, pero con relación de escala numérica. Para distinguirla de la Ordinal, debemos plantearnos si es objetiva la distancia, pues en ese caso es Ordinal. Ej. puntuación en una prueba de inteligencia, en un examen.*
- Escala de Razón. *Es una escala de Intervalo en la que el 0 significa carencia de la propiedad a medir. Ej. número de monedas, número de hijos.*

Obviamente, a mayor fiabilidad en el dato, más posibilidad de tratamiento estadístico. Cuando hay problemas de medida se ha cambiar a una escala de menos rango (siendo el rango el nº de relaciones existentes entre los distintos elementos). No se puede pasar de uno mayor a uno menor. Es importante el concepto de *Mortalidad Experimental* (datos perdidos durante la investigación), la solución a esto es dejar el espacio en blanco o poner un dato absurdo (ej. edad −23). Calcular la tasa de Motalidad Experimental es:

- Nº de observaciones (fila por columna) n.
- Calcular a nivel % la pérdida de datos 100 por errores /n.

Una tasa de Mortalidad Experimental es aceptada si tienen 10% o menos de error. Aunque dentro de Psicología hay excepciones, son casos aislados.

Parte II: Estadística Descriptiva

- La variable en medición. Al conjunto de valores numéricos es lo que denominamos variable, son un conjunto de valores numéricos que tratan de mostrar la asiduidad con que se presenta una característica. Los números de una variable no son valorables en sí mismos, dependen de la escala de medida utilizada. Las variables suelen ser denominadas x, y y z. En álgebra matricial son vectores, vector fila y vector columna. También hay que tener en cuenta el uso de subíndices, etc. También es importante el concepto de Sumatorio. Existen diversos tipos de variables según el nivel de medida:

- Variable Cualitativa o Categórica (escala nominal).
- Variable Cuasicuantitativa (escala ordinal).
- Variable Cuantitativa (escala de intervalo y de razón).

- V.Cuantitativa Discreta.
- V.Cuantitativa Contínua.

- Tabulación de datos. Persigue recoger de forma rápida y sencilla el comportamiento variable. Es construir una tabla / matriz que resuma el comportamiento de una variable.

- Identificar el número de modalidades (valores posibles dentro de una variable, como por ej. medir la clase social tiene 3 números −baja, media y alta−).
- Construir la tabla, teniendo en cuenta el tipo de variable:

- Si son variables cualitativas se usa una tabla en la que han de figurar x, F (Frecuencia absoluta), f (frecuencia relativa) y P (porcentaje) y que su representación gráfica es un Diagrama de Barras. Es importante utilizar bien la escala y no cometer estancamiento estadístico(maquillaje de datos). La y debe partir siempre de 0.
- Si son variables no cualitativas las partes a figurar en la tabla son: x, F, f, P, Fac (Frecuencias Absolutas Acumuladas), fac (Frecuencias Relativas Acumuladas) y Pac (% acumulados). En cuánto a la representación gráfica, se puede usar el Diagrama de Barras, aunque es más correcto usar el Polígono de Frecuencias. Si son variables continuas se usa un Histograma.

- Concepto de Intervalo Compuesto: *es un método de tabulación clásico antes de la aparición de los ordenadores. Dentro de un intervalo compuesto distinguimos un límite inferior real y un límite superior real. Cuanto más números agrupamos en un intervalo mayor es el error.*
- Diagrama de Tallo y Hojas (Stem&Leaf): *es la representación más aceptada siempre que tengamos variables cuantitativas y gran cantidad de datos. Sustituye tanto la representación gráfica como la tabulación de datos.*
- Gráfico Box−Plot: *extensión del Stem&Leaf, su finalidad es intentar determinar en que medida se distribuyen los datos en un comportamiento normal, en qué medida hay puntaciones Outlier (extrañas). Necesitamos la tabla que informa de los % acumulados (o alguna que nos permita llegar a este dato). Existen también Puntuaciones Extremas (tienen un comportamiento mucho más alejado del normal de la variable). Si da un valor negativo y es una escala de razón el valor se sustituye por cero. El valor sugerido será válido sólo si el Rango Normal de la variable (12 en una escala 1−10 sería válido, luego se cogería el 10). En un Box−Plot hay varias partes bien diferenciadas: un rectángulo que se inicia en el 25% de datos y acaba en el 75%, una línea dentro de este en el 50% y*

otros dos segmentos fuera de él, uno en el valor mínimo y otro máximo. Cada uno de los segmentos colocados en el 25 y 75 son las Bisagras de Tukey, y la distancias entre ellos es la Amplitud Intercuartilar. Para determinar la región de rechazo realizamos esta fórmula matemática: Amplitud Intercuartila por 1.5 + Bisagras de Tukey. Las Puntuaciones Extremas se hallan sustituyendo 1.5 por 3.

Medidas de tendencia central

- Se suele distinguir entre población y muestra:

- Población: *conjunto de n elementos que cumplen es una propiedad, es decir, aquéllos sujetos que le interesan al investigador. Investigar una población suele ser algo largo, tedioso y costoso, por eso se usan muestras.*
- Muestra: *subconjunto de la población, que debe cumplir los mismos requisitos que la población. Al procedimiento se le conoce como Procedimiento de Muestra. Al conjunto se le conoce como Estadísticos y los resultados sólo son para generalizar los resultados sobre la propia muestra, los resultados se pueden inducir a la población gracias a los Estadísticos Inferenciales.*

- Estadísticos de tendencia central (Promedios). Solo tienen el problema de que se pierde la información individual:

- Moda (Mo). *Estadístico de tendencia central, único que se puede utilizar en variables cualitativas. Se calcula a partir de una tabla de frecuencias absolutas. La moda es igual a la variable que ocurre con mayor frecuencia. Puede haber una moda (Distribución Unimodal), dos (Dist. Bimodal), tres o incluso más modas (Dist. Multimodal). Cuando todas las variables tienen frecuencia máximo se dice que la variable tiene una Distribución Uniforme.*
- Mediana (Md). *Aquél valor que divide los datos al 50%. La forma más fácil de hallarla es con los porcentajes acumulados. Siempre da un único valor. Se prefiere la Md a la Mo.*
- Media Aritmética (x). *Ponderación general de una serie de puntuaciones a nivel cuantitativo. En principio se aplica a variables cuantitativas, se prefiere a la Md. En los valores numéricos funciona la ley de distancia, también hay que tener en cuenta que los datos han de ser congruentes; en caso de incongruencia podemos o redondear o utilizar un Estadístico de tendencia central más débil (Md o Mo), esto último es lo que más se hace.*
- Media Ponderada (Xw). *Igual que la media pero los valores numéricos tienen distinta importancia a nivel teórico. Esto modifica los valores empíricos. La M. Ponderada es un medio de ponderar la información donde el componente subjetivo modifica de forma grave el resultado final. Se utilizan especialmente en Psicología Industrial y del Aprendizaje.*
- Media Geométrica (Xg). *Xg es igual a la raíz n de xi, teniendo en cuenta que la Media Geométrica deja de ser operativa cuando la expresión da 0, funciona bien siempre que no existan valores nulos. Una forma de evitarlo es transformar los valores al tipo x+1.*
- Media Armónica (Xh). *N partido del sumatorio de 1/xi.*
- Media Cuadrática (Xc2). *Es igual al sumatorio de* X2 partido de N. Sirve para puntuaciones negativas, y una vez se obtiene el resultado hay que hacer la raíz cuadrada puesto que es un resultado elevado al cuadrado.

- Calcular los promedios de Intervalos Compuestos.

- Moda. *Se cogen los valores extremos, se divide entre 2 y se hace la moda.*
- Mediana. *Aplicamos la fórmula Md = L + (n+1/2 − F), todo el paréntesis partido de f y multiplicado por A. Siendo L el límite inferior real del intervalo, A la amplitud real del intervalo, F todos las cosas en los valores inferiores a donde está el intervalo y f todos las cosas en el intervalo.*
- Media. *Se realiza de la manera habitual pero teniendo en cuenta que Xi es el punto medio de un*

intervalo compuesto.
- El grado de error cometido va en relación con la amplitud del intervalo, a mayor tamaño, mayor error. Sólo funciona en intervalos cerrados (intervalos abiertos serían, p.ej. <5 o >10). Los intervalos abiertos no pueden resolverse.

- Transformaciones Lineales. *La Media es susceptible de operar con transformaciones lineales básica, sí y sólo sí está a nivel cualitativo. Ej. de transformaciones lineales básicas es sumar o multiplicar un valor constante por la variable.*
- Reglas para seleccionar que estadístico utilizar:

- En cuantitativas, se una la Moda.
- En cuasicuantitativas, se una la Mediana siempre que sea posible.
- En cuantitativas, se usa la Moda siempre que sea posible (no se usaría, por ejemplo, si las puntuaciones Outlier son muy significativas).

Índices de Posición

- Son estadísticos donde posicionamos al individuo y no al grupo. Todos son métodos inexactos.
- Posicionamiento Empírico. *Va a estar en relación con lo observado en el medio y no con lo teórico.*
- Escala Percentil. *Construcción de una escala donde los individuos se posicionan de acuerdo a cien partes proporcionales; en estas escalas no existe en centil 100, sólo el 99. No admite valores decimales. Para hallarlo se cogen los porcentajes acumulados y se hace correspondencia, teniendo en cuenta que el 100 es 99. Sirve para calcular los perfiles de comportamiento.*
- Escala Decilar. *Similar a la centil pero se constituye en 10 partes. Un centil 10 es igual a un decil 1 y así sucesivamente. El máximo valor es el 9.*
- Cuartiles.*Divide la distribución en cuatro partes (cada cuartil es más o menos 25%)*

Medidas de dispersión

- Nos indican en que medida los sujetos se diferencian unos que otros, existen dos situaciones:

- Homoscedasticidad. *Variaciones pequeñamente diferenciados.*
- Heteroscedasticidad. *Variaciones más amplias.*

- Algoritmos de dispersión en Escala Nominal:

- D de Scott. *Se aplica el algoritmo Ds = 1 / fi 2. El resultado siempre es entre 1 y k, siendo k el número de elemntos de la variable, cuanto más se acerca a 1 más homoscedasticidad existe. No sirve para comparar variables con distinta modalidad (k).*
- Índice de Entropía (utilizado por defecto): *Necesitamos la misma información que en la Ds, debemos aplicar la fórmula H = − fi log2(fi). La ventaja de este método es que permite comparar variables de distinta modalidad (k).*

- Algoritmos de Dispersión en Escala Ordinal (aparte de poder usar los anteriores).

- Rango. *También conocido como Recorrido o Amplitud total, se calcula así AT = M − m, siendo M el valor más alto de la variable y m el más pequeño.*
- Amplitud Semiintercuartílica (ASI). *Se calcula dividiendo entre dos la Amplitud Intercuartílica; se puede usar en las ordinales pero solo ocasional y puntualmente.*

- Algoritmos de Dispersión en Escala de Intervalo (aparte de los anteriores).

- Varianza. *S 2 = (x − media) 2, todo partido de n. Esto es igual a SC / n (Suma de Cuadrados). La varianza es un estimador sesgado, son valores erróneos de lo que sucede en realidad.*
- Cuasivarianza (Se utiliza por defecto). *Es idéntica, tan sólo que se divide entre n−1 y se simboliza como S y encima un ^.*
- El resultado de ambos algoritmos está en un valores cuadrácticos, por eso debemos aplicar la raíz cuadrada al resultado. A esto se lo conoce como Desviación Típica.

- Representación de variables cuantitativas continuas. *Similar al Box−Plot, sólo que éste se basa en las medidas de tendencia central y de dispersión. El centro del diagrama es la media y los extremos la desviación típica. Se pretende visualizar el grado de homoscedasticidad de los sujetos y ver si la media es representativa (lo que sucede cuando los sujetos se diferencian poco entre sí).*
- Transformaciones Lineales. *En el proceso de adición el estadístico de adición no varía, pero en el de multiplicación sí (queda multiplicado por el número).*
- Comparación de la Dispersión. *Cuando comparamos variables del mismo rango es eficaz la mera comparación, pero cuando no se da ese caso hay que usar el estadístico cociente variación (CV). El CV tiene tres soluciones: sesgado, insesgado y robusto, siendo el más eficiente el robusto. Los tres se expresan en %.*

- *CVs = Sx / media , todo ello por 100.*
- *Cvi = nSx / media, todo ello por 100.*
- *CVr = AI / Q3+Q1, todo ello por 100.*

Modelo Integral de Gauss

- Tiene las siguientes propiedades:

- Propiedad de Simetría. *Una integral es simétrica si Mo, Md y media tienen el mismo valor. Si la información se encuentra concentrada en los valores pequeños estamos ante un Modelo Asimétrico Positivo, mientras que si se concentra en los negativos estamos ante un Modelo Asimétrico Negativo.*
- Grado de Concentración de la Información (Apuntamiento ó Kurtosis). *Nos dice en qué medida vienen representados todos los valores. Existen tres modelos: Meskúrtico (todos los valores tienen información), Leptokúrtico (sólo los valores centrales tienen información), Platokúrtico (modelo de tipo uniforme).*

El modelo de Gauss es simétrico y mesokúrtico.

- ¿Cómo comprobar si el modelo es simétrico? *Se aplica el siguiente estadístico, si da entre −1 y +1es una distribución simétrica, si da entre +1 y +2 será una AS+, mientras que si da entre −1 y −1 será una AS−. El estadístico es As = media − Mo / Sx (siendo Sx la desviación típica). El problema del estadístico es que cuando existe más de una Mo el estadístico no funciona y ha de utilizarse otro, el Índice de Dispersión 3, que es [(x − media)3 / n] / Sx3.*
- ¿Cómo calcular la Kurtosis? *Se aplica el índice de dispersión de orden 4, al cuál se le resta 3 para que el caso ideal de como valor 0. [(x − media)4 / n] / Sx4 , todo ello − 3. Si da entre −1 y 1 será Mesokúrtico, si da entre 1 y 2 será Leptokúrtico y si da entre −1 y −2 Platokúrtico.*

- Operaciones con el modelo integral de Gauss. *Se precisan las tablas de la distribución normal y conocer la siguiente fórmula Z = x − media / Sx (desviación típica). Las preguntas posibles son: averiguar el área de un punto dado, averiguar el punto para un área dada, dando N averiguar Z o dado unos valores averiguar la media o la desviación típica.*

Típicas Derivadas

- Son estadísticos en los que lo que se persigue es caracterizar a los sujetos y ni al grupo, en última

instancia nos permite saber si el sujeto está dentro o fuera del grupo. Se basa en el modelo de Gauss y no en una distribución de frecuencias.

- Las puntuaciones directas (x,y) denotan la información que recoge el investigador, es decir, las magnitudes medias en el estudio.
- Puntuaciones diferenciales (x − media).
- Puntuaciones Z (x / Sx = x−media / Sx = Z), que equivalen a la z de Gauss siempre que se ajuste a un modelo normal.

- Una escala típica de derivadas no es más que una puntuación derivada de las puntuaciones Z, sólo puede hacerse si se ajusta al modelo de Gauss. En este curso vamos a ver su ejemplo sobre tres formas de valorar el CI.
- El CI es un constructo hipotético y se puede medir de tres formas:

- Por el método del CI (z * 15) +100
- Por los Estaninos (z * 2) +5
- Por la escala D (z*20) +50

- Si quisiera construir una variable similar a las anteriores pero no poseo una distribución normal hay que ejecutar un maquillaje de datos; lo que se hace es normalizar las Z, se obtienen los porcentajes acumulados y se busca en la integral de Gauss.

Estadística Bivariada

- En este curso solo vamos a ver la ceñida al modelo lineal y dentro de ese, los casos más clásicos. Vamos a suponer que las variables siguen el modelo de Gauss y se ajustan al modelo de línea recta, ya que en otro caso estos algoritmos no servirían.
- y = f(x) y = A + Bxi, siendo A la constante de intercepción y B la pendiente (B = Ay / Ax).
- Covarianza: la fórmula es (x−media) (y −ymedia) / n, siendo n pares de observaciones. Los resultados pueden ser: 0 (ausencia de relación lineal, lo cuál no excluye otro tipo de relación), + o − (no se puede saber con exactitud). Este estadístico es sesgado, se puede conseguir el insesgado sustituyendo n por n−1. La solución de la covarianza por el método matricial es E (L'*L).
- Pearson propuso después una solución matemática a los problemas de la fórmula anterior, el algoritmo conocido como R de Perarson, se diferencia del anterior fundamentalmente por la información de partida, porque en lugar de partir de las puntuaciones diferenciales lo hace de las z. La fórmula es: ^Rxy = (z−x) (z−y) / n−1 = (x−media / Sx) (y −ymedia / Sx) / n−1. El estadístico de Pearson tiene límites claros: 0 (relación al azar), +1 (modelo +) y −1 (modelo −). La resolución matricial es idéntica [(z'*z) * K = R] pero con puntuaciones z en vez de diferenciales. El determinante de R nos permite saber si puede dar 1 (matriz identidad), cuando el determinante de R sea un valor próximo a 0 es que hay variables muy relacionadas en el modelo lineal.
- Técnicas Q. En vez de buscar relaciones entre variables, busca relaciones entre individuos, por ello se traspone la matriz z, ahora la media es la media de cada individuo dentro de la variable que estamos estudiando. Salvo en casos particulares se usan las técnicas R más que las Q.
- Volviendo sobre la Correlación de Pearson, debemos pensar en la interpretación. Para empezar hay que tener claro el índice máximo y mínimo de la correlación (+/−1). El número dice la cuantía de la relación y el signo indica la dirección de esa relación. Si es positiva es relación directa, si es negativa es relación inversa. Para interpretar el grado de relación se usa el Coeficiente de Determinación (V2xy) que nos da la proporción de varianza que x e y comparten. Hay que tener en cuenta que la correlación nunca implica causalidad. Hay que recordar que la relación que buscamos y que puede existir o no es de tipo lineal.
- Factores que influyen en la correlación:

- Los outliers (tanto por hacer creer que no hay una correlación como que sí la hay cuando esto no es cierto).
- La muestra puede no ser representativa de la población (podemos haber cogido una muestra muy restringida, y a más homogeneidad, menos correlación).
- Hay también que tener en cuenta que entre dos variables puede haber alguna relación de una tercera variable que influya. Para evitar esto la solución está en pesar si hay alguna variable de este tipo y realizar sobre ella un control empírico. Si este control no es posible se pueden usar métodos estadísticos. También puede ser que una tercera variable haga creer que no hay relación entre dos variables que si la tienen).

    - Correlación de Spearman. *Se usa cuando las dos variables están en una escala ordinal (variable cuasicuantitativa). Se usará cuando ambas sean cuasicuantitativas o una cuasicuantitativa y la otra ordinal. A veces también se usará con dos cuantitativas por intención del evaluador (aunque no conviene hacerlo). Cuando hay dos variables que son de distinta escala, hay que reducir una de grado, esto se logra dando orden, siendo 1 el valor más bajo. Por último, si varios sujetos tienen la misma puntuación, entonces se les da ambos valores el puesto intermedio. La fórmula es rs = 1 − [6 * "d2 / n (n2−1)], siendo d la diferencia de rango para cada sujeto entre ambas variables, n el número de sujetos que componen la muestra.*
    - Correlación Biserial Puntual. *Cuando una variable es cuantitativa y la otra dicotómica (sólo puede tomar dos valores, como por ejemplo el sexo) se usa otro algoritmo. Hay que distinguir entre variable dicotómica y dicotomizada (una cuantitativa dividida a dos categorías). Las dos fórmulas que se pueden utilizar son Vbp = mediap − media / ^Sx, todo ello por la raíz cuadrada de p/q. La otra fórmula es Vb = mediap − mediaq / ^Sx, todo ello por la raíz cuadrada de p*q. Estos algoritmos son equivalentes y sus símbolos significan: p (proporción de sujetos de la categoría primera), q (proporción de sujetos de la categoría segunda), mediap (media en la variable cuantitativa de los sujetos con proporción p), mediaq (media en la variable cuantitativa de los sujetos con proporción q), media (media aritmética en toda la muestra, sin distinguir) y ^Sx (desviación típica para todos los sujetos.*

<u>Hasta aquí el primer parcial</u>

Regresión Simple

- Se utilizan las relaciones para hacer predicciones, como siempre, bajo modelos lineales (Y=A+BX) y variables cuantitativas.
- Hay que mencionar, claro está, que al usar un modelo lineal hay un ligero desfase de nuestros cálculos respecto a la realidad (error de pronóstico), sólo en caso de una correlación perfecta no habría error. La nueva recta es Y'=A+BX+e (si se conociera e, Y' sería igual a Y. Es un valor teórico, no se le puede dar valor).
- De las infinitas rectas que podemos trazar, ¿cuál seleccionamos? Escogeremos aquélla que comenta menores errores; existen varios criterios, nosotros utilizaremos el criterio de errores cuadráticos mínimos (mínimos cuadráticos), que consiste en hacer ésto: "(Y'−Y)2 / n, y utilizaremos la recta que proporcione el valor más bajo.
- Formas que toman las rectas según trabajemos con un tipo de puntuaciones u otras:

- Directas: Y'=A+BX, siendo A ordenada en el origen y B pendiente de la recta. B= n"XY − "x"Y / n "X2 − ("X)2 o B = Rxy por Sy / Sx.
- Diferenciales: y' = a + bx, siendo b = B y a = 0.
- Típicas: Z'y = + Zx, siendo = Rxy y = 0.

Hay que recordar que la pendiente sirve también como tasa de cambio (p.ej. una B de 1'5 indica que por cada unidad de x hay 1'5 de y) y que estos algoritmos proporcionen la recta con menos errores no quieren decir que los errores sean pocos.

- Valoración. *Hay que observar la nube de puntos en relación a la recta, cuanto más cerca están los puntos de la recta, más acertada será ésta. ¿En qué medida mejoran mis predicciones al usar x además de y respecto de usar y únicamente? Usar y' reduce el error respecto de usar y.*
- Variación total de la variable dependiente: *" (Y − Y') 2 = " (Y' − Ymedia)2 + " (Y − Y') 2. Esto es que la suma de cuadrados es igual a la suma de cuadrados explicada por la regresión más la suma de cuadrados no explicada o error. Si se divide todo por N−1 tenemos tres varianzas: ^S2y = ^S2y' + ^S2e, es decir, Varianza de los Pronósticos = Varianza explicada por equis.*
- Esos algoritmos son los que se utilizan para determinar si una recta explica bien, cuando más cerca estén *^S2y e ^S2y'* mejor explicada estará. Al hacer uno de los siguientes algoritmos se consigue una proporción de varianza explicada:

Sesgado

R2xy = 1 − [^S2y / ^S2y']

Insesgado

^R2xy = R2xy − [p (1−R2xy) / n − p−1)]

^R2xy = 1 − [(1−R2xy) * (n − 1/ n−p−1)]

Siendo p el número de variables independientes, en estos casos, 1, y siendo n el tamaño de la muestra.

- Para obtener la proporción de la varianza no explicada utilizamos el Coeficiente de Alienación (CA). CA = 1 − r2xy, lo que es igual a ^Se2 / ^Sy2.
- A partir de esa operación podemos averiguar:

^Se2 = ^Sy2 (1 − r2xy)

^Sy'2 = ^Sy2 − ^Se2

- Es importante recordar que si hablamos de varianza explicada se refiere a Se2, etc. Si habla de proporción de varianza se refiere al coeficiente de determinación, al CA...
- ¿Bajo qué condiciones puedo aplicar el modelo de regresión lineal?

- *Especificar correctamente el modelo. Que el modelo sea adecuado para lo que queremos, en la hipótesis de partida existe una relación lineal entre las dos variables. Eso se puede realizar con un diagrama de dispersión si en principio tengo un círculo, no lo haríamos. A veces hay soluciones matemáticas para obligarlos a que tengan modelo lineal, mediante logaritmos de la variable en vez de y' a partir de x, y' a paritr de log de x. Algo no muy apropiado en Psicología, pues además de los números hay que interpretar a los sujetos. Cuando en el modelo faltan variables, en Psicología asumirá siempre que utilizamos regresión simple. También lo contrario, porque puede haber un exceso de variables que sean irrelevante. En psicología se utiliza normalmente Regresión Lineal Múltiple.*
- *Las variables están medidas sin error. Uno de los problemas en psicología es la medición de variables. Si medimos mal estamos introduciendo errores y luego si se introduce como puntuación en el modelo matemático dará como resultado muchos más errores. El patrón que debemos encontrar para los errores debe ser unificado, que no siga ninguna correlación, si me saliera algún tipo de relación lineal o servirá, el modelo no cumplirá el supuesto.*

- Igualdad de Varianzas: Homoscedasticidad. *Los distintos valores de x y los valores de los errores y' tienen la misma variabilidad, lo que implica que se dará Homogenenidad Favorable y en la representación gráfica no existirá ningún patrón. Otro caso es la Homogeneidad Desfavorable, en el*

cuál tenemos nubes de puntos con forma de embudo. La dispersión aumentará a medida que aumentan y' y x. En estos casos no podemos utilizar el criterio de mínimos cuadrados, sino que usaríamos mínimos cuadrados ponderados (introduciendo la varianza).

- Independiencia entre los errores. *A lo que nos referimos es que el error que cometamos para el sujeto 1 al hacer un pronóstico no tiene por qué ser él mismo para los sujetos 2 ó 3...esto ocurre para casos estáticos. Hay situaciones dinámicas (cuando se mueve a través del tiempo) en los que es muy fácil que haya correlación en los errores, luego no son independiente y no se puede utilizar la regresión. ¿Cómo determinamos si podemos aplicar la regresión lineal en caso de errores indefinidos? Usaremos el estadístico de Durbin−Watson:*

D = " mi=2 (ei − ei−1)2 / " mi=1 ei2

ei = error de pronóstico para el sujeto i.

ei−1 = error de pronóstico para el sujeto anterior a i.
Toma valores entre 0 y 4, cuando D=2 señala independencia entre los errores, si da <2 correlacionados positivamente, mientras que si da >2 correlacionados negativamente.

Para considerar si el modelo cumple este supuesto consideramos entre 1 y 3 que son independientes, por debajo de 1 y por encima de 3 ya no podríamos usar regresión lineal.

- Que los errores se distribuyan respecto a la curva normal. *Se pueden realizar a ojo mediante un histograma y viendo si sigue la campana de gauss.*

Regresión Múltiple

- Predecimos y a partir de 2 variables independientes, ya no nos sirve la recta en un plano, ahora necesitamos trabajar en modo de plano. Se trabaja siempre con matrices. La expresión matemática es:

Y' = A + B1x1 + B2x2 + ... + Bkxk

Y = A + B1x1 + B2x2 + ... + Bkxk + Error de pronóstico

- La fórmula para los tres tipos de puntuaciones son:

Directas : Y' = A + B1x1 + B2x2 + ... + Bkxk

Diferenciales: y' = a + b1x1 + b2x2 + ... + bkxk

Típicas: z' = + x1 + x2 + ... + kxk

- Trabajando con puntuaciones típicas:

Ejemplo: pronosticar a partir del CI y las horas de estudio una nota (nos dan los datos)

1º Calcular la matriz de correlaciones

2º Invertir y hacer la adjunta

3º Trasponer la adjunta (si es simétrica, ya está traspuesta)

4º Dividir cada elemento de la adjunta traspuesta entre el resultado de la inversión

5º Multiplicar la matriz que resulta por el vector Rxy (es una correlación)

6º El resultado es , cada uno de los resultados es sub1, sub 2, etc.

7º Sustituir en la ecuación de típicas.

- Pasar a otras puntuaciones: se utiliza el método habitual, hay que tener en cuenta que para calcular Y' hay que calcular A, y que A = Ymedia − (B1x1+B2x2+...+Bkxk). La A nos la da la altura a la que se sitúa el plano respecto al eje de coordenadas (el 0,0), la B nos la da la inclinación del plano.
- Problemas: *los cambios cuantitativos no son tantos como los cualitativos. Los valores deben estar en la misma escala para ser comparable, y esa escala son las desviaciones típicas. Si no están en típicas podemos pensar que de las dos o más variables independientes una es la importante cuando en realidad es otra.*
- Formulas que se pueden usar cuando hay 2 variables independientes:

- B1= Ryx1 − Ryx2 Rx1x2 / 1− R2x1x2
- B2= Ryx2 − Ryx1 Rx1x2 / 1− R2x1x2

- Valoración del modelo. *Dado que el que el modelo obtenido sea el mejor no quiere decir que sea bueno debemos valorarlo. En lugar de usar Rxy (Pearson) usaremos correlación múltiple. No es más que la correlación entre una variable y un grupo de variables tomada conjuntamente. Se representa como Ryx1x2..xn (si fueran sólo dos variables sería Pearson), aunque también se puede representar como Ryy. Hay tres formas de hacer la valoración:*

- *Cuando sólo hay dos x: Ryy' = R2yx1 + R2yx2 − 2Ryx1Ryx2Rx1x2 / 1 −R2x1x2.*
- *Si conocemos : Ryy' = Ryx1 + Ryx2 + ... + k RyXy.*
- *Ryy' = Rxy*

Estimadores insesgados ^Ryy': 1 − [(N−1) (1−R2yy') / N−1<−1]. R2yy' = k (1−R2yy') / N−1<−1.

No se debe de valorar sólo el número, sino también el contexto. Algo a tener en cuenta es que si se introduce una nueva variable el coeficiente de determinación siempre aumenta y en el peor de los casos sigue igual. Desde un punto de vista estadístico los modelos funcionan mejor cuanto mayor número de variables, pero no siempre es lo correcto. Existe un truco que es la metodología Stepwise, por la cuál construimos con el ordenador un modelo de forma que llegue un punto en el que introducir variables tenga un efecto tan nimio que no nos sea útil. Hay dos métodos foward (hacia adelante) y backward (hacia atrás).

Correlación Parcial

- Se realiza sobre variables continuas, ya que hay veces que Pearson no capta la realidad con efectividad. Este método se realiza cuando las variables no se pueden controlar pero queremos tenerlas en cuenta, nos permite captar la relación lineal entre dos variables eliminando la influencia que sobre ambas tiene una tercera variable. El algoritmo es: *R12*3 = R12 − R13R23 / " 1−R213 "1−R223.*
- Se puede realizar también una correlación semiparcial, que ve la relación entre dos variables controlando una tercera en una de las dos variables. El algoritmo es: *R1(2−3) = R12 − R13R23 / " 1−R223*

Estadística Inferencial

- Lo que persigue es extrapolar los resultados obtenidos con la estadística descriptiva a la población. La mayoría de lo que vamos a ver se basa en probabilidad ya que el modelo de extrapolación es

probabilístico.

- Experimento Aleatorio: *son los sucesos que podemos observar en un campo científico y en el cuál no es posible determinar con total certeza el suceso o sucesos que pueden ocurrir.*
- Suceso Elemental: *es cada una de las posibilidades que se pueden verificar dentro de un experimento aleatorio.*
- Suceso Compuesto: *es cuando se agrupan de forma arbitraria una serie de experimentos simples. Ej: agrupar sucesos simples en caso de las notas (Sobre, Notable...).*
- Población: *sujetos de estudio.* Muestra: *sujetos que representan a la población.*

- La probabilidad oscilará entre 0 (nunca se produce) y 1 (siempre se produce).

- Representación en función de probabilidad diagrama de barras. Representación en función de distribución polígono de frecuencias.
- Modelos de Probabilidad (existen 3):

- Clásico / Laplaciano. *Se basa en asignar el mismo grado de probabilidad a cada uno de los sucesos. P = casos favorables / casos posibles.*
- Frecuentista a posteriori. *El modelo se calcula a partir de la frecuencia relativa; f = Fi / N P (S) = f.*
- Modelo de Probabilidad subjetiva o Bayesiano. *Cuando la conducta humana no se ajusta a ninguno de los modelos anteriores; el grado de incertidumbre viene dado por fenómenos subjetivos, como por ejemplo el grado de creencia de un sujeto de que va a llover mañana. La probabilidad viene dada por la creencia del sujeto. La parte esencial es el proceso de muestreo, deformación del subjconjunto. Como regla general, cuántos más sujetos tenga la muestra, mayor probabilidad de que sea representativa. Como regla general un 5% de probabilidad es el subjetivo, aunque dependerá del tamaño de la población.*

- Lógica Fuzzy: *en vez de sumar valores (Ej. 0'3+0'2 = 0'5 bajo esta lógica se toma uno de los valores).*
- Variable aleatoria: *se define así toda función que asigna un número real y sólo uno a cada suceso elemental de un espacio muestral. Al referirnos a ella usaremos X y cada resultado concreto con x minúscula y con un subíndice. Dentro de una variable aleatoria pueden ser discretas (espacio muestral finito o infinito pero numerable) y continuas (espacio muestral infinito no numerable).*
- Variables discretas: *dos conceptos:*

- *Función de probabilidad. Aquélla que nos da la probabilidad de que la variable aleatoria tome un valor concreto. Se representa normalmente con f minúscula f(x) P (x=xi). La probabilidad de un valor que no se puede asumir será 0. La suma de todas las funciones de probabilidad ha de ser 1.*
- *Función de distribución. Aquélla que nos da la probabilidad acumulada para un determinado valor de la variable.*

- Valor Esperado o Esperanza Matemática. *E (X) = x f (x). Todos los valores que puede tomar x y la función de probabilidad. Si el valor sale distinto de 0 un juego de azar es injusto. Algunas normas:*

- *E (a) = a.*
- *E (x+y) = E (x) + E (y).*
- *E (x+a) = a + E (x)*
- *E (ax) = a E (x)*
- *E (ax+b) = a E(x) + b*
- *E (a1x1+a2x2) = a1 E (x1) + a2 E (x2).*

- Modelos de Probabilidad Q (Bernoulli). *Llamaremos prueba de Bernoulli a toda realización de un experimento aleatorio en el que sólo son posibles dos resultados que se llamarán éxito y fracaso y que son mutuamente exclusivos. La probabilidad de éxito será p y de fracaso 1−p o q. Fórmula: F(X)*

= P (x < k) = x=0 pxq1−x.

- Distribución Binomial. *Se refiere a n pruebas de Bernoulli independientes tales que la probabilidad de éxito se mantiene constante en todas ellas. El resultado de un experimento no influye en el de los otros. La fórmula no es necesaria puesto que contamos con tablas.*
- En la práctica puede ocurrir que la variable sea continua, pero en la mayoría de los casos aunque lo sea tendremos que trabajar con ella como si fuera discreta. Al trabajar con variables aleatorias continuas denominamos función de densidad de probabilidad a la función de probabilidad. Cuando estamos en variables continuas la probabilidad de que la variable tome un valor concreto es 0. El concepto de función de distribución se mantiene igual.
- Grados de Libertad: *número de elementos de una expresión matemática que pueden escogerse libremente. Número de observaciones que pueden elegirse libremente en un modelo o situación matemática concreta. Normalmente vienen dados por n−k, siendo n el tamaño de muestra y k el número de restricciones que ponemos. Ej. dime 5 números = 5 grados de libertad, pero dime 5 números que sumen 100 son 4 grados de libertad.*
- Modelos de probabilidad para variables continuas:

- Distribución Normal. *Lo que nos da esta tabla es la probabilidad de que la variable adopte un valor o menos. Propiedades de la normal:*

  - *El área bajo la curva es 1.*
  - *Es simétrica.*
  - *Mediana, moda y media coinciden.*
  - *Es asintótica respecto a las absisas (en los extremos se acerca al eje de las x pero no lo toca).*
  - *Hay un punto de inflexión para cada parte y siempre está a distancia de una desviación típica respecto a la media.*
  - *Cualquier combinación lineal de variables normalmente distribuidas da lugar a otra variable normalmente distribuida.*

La curva normal más conocida es la típica y la tabla en éste caso nos da probabilidad.

- Distribución . *Supongamos que tenemos n variables aleatorias distribuidas según la curva normal tipificada y a partir de ellas construimos la siguiente expresión: elevamos cada suma al cuadrado y los vamos sumano y su resultado será la variable y cuya función de densidad de probabilidad tiene unas características que conocemos: n, siendo n los grados de libertad. Ésta distribución se trabaja con tablas dada su dificultad. Las características de las curvas son:*

  - Propiedad aditiva. *Si tengo una variable x distribuida según con n1 grados de libertad y una variable distribuida según con n2 grados de libertad y las sumo, la nueva variable también se distribuye de acuerdo a , pero con n1+n2 grados de libertad.*
  - *no puede tomar valores negativos, siempre entre 0 e infinito.*
  - *es asimétrica; a medida que aumentan los grados de libertad se acerca a la normal. Con 30 o más grados de libertad, se iguala a la normal. Para trabajar con más grados de libertad se aplica: p n = ½ (Zp + "(2*n)−1)2.*

- T de Student (Gosset). *Ésta distribución surge de la combinación de N (0,1) con y da lugar a t = z / "x2/n , siendo n los grados de libertad. Se una la distribución normal tipificada. Las características son:*

  - *Valores entre + y − infinito, aunque suele tomarse entre + y − 3.*
  - *Simétrica en torno a 0, pero más plana y dispersa que la normal, a medida que aumentan los grados de libertad se acerca a la curva normal.*

En la tabla, g nos da los grados de libertad, se nos da la información por debajo y la probailidad es lo que

viene entre 0'60 y 0'995.

- F de Fisher (Snedecor). *Surge de la combinación de dos variables distribuidas de acuerdo a y con n1 y n2 grados de libertad. Es importante el orden. Fn1n2 = / n1 / /n2.* Si un valor no viene en nuestra tabla aplicamos la Propiedad de la Probabilidad Recíproca: si x es una variable con distribución F y m y n grados de libertad, entonces y = 1/x también se distribuye según F pero con n y m grados de libertad. Propiedades:

    - *Asimétrica.*
    - *Siempre toma valores positivos.*
    - *Tiende a hacerse más simétrica y aproximarse a la normal al incrementarse los grados de libertad y converge cuando ambos son infinito.*

    - Distribución de Probabilidad. *Es una función de probabilidad o de densidad de probabilidad definida sobre un conjunto de sucesos exhaustivos y mutuamente exclusivos. Las distribuciones suelen ser de corte técnico, lo que nosotros creemos que pasará; en la práctica suponemos que las variables se ajustarán a algunos de los modelos que hemos ido viendo. Esas distribuciones muestrales van a tener importancia en estadística inferencial porque nos van a permitir tomar decisiones. Éstas distribuciones sirven para los sucesos que ocurren por azar, si encontramos diferencias entre teoría y práctica podemos decir que esto no ha ocurrido por azar.*

Estadística Inferencial

- Suponiendo que queremos hacer una investigación sobre un gran grupo cogemos una muestra y recogemos los datos y luego aplicamos los estadísticos que conocemos, pero todos los valores que obtengamos sólo dan datos sobre la muestra nada más. El paso entre la muestra y la población es de lo que se encarga la estadística inferencial. Cuando uno trabaja en Ciencias Sociales y repite un experimento a diferentes grupos resulta que en cada muestra hay resultados distintos, de forma que difícilmente se da el salto a la Estadística Inferencial. La solución es introducir la probabilidad, sin embargo, al hacer ese salto cabe, obviamente, la posibilidad de error.
- Teoría del Muestreo (Normas a seguir para seleccionar los elementos que van a servir para nuestra muestra). *No es estrictamente estadística inferencial pero sí necesario para ella. Nos permite elegir muestras de la forma adecuada. Conceptos previos son: elemento (unidad básica de la que buscamos información y que es la que nos va a proporcionar los datos para luego analiazrla. Un elemento puede ser humanos, animales, rocas...), población (conjunto de elementos, finito o infinito definido por un conjunto de características que comparten. Es importante definirlo bien. A veces es posible trabajar con todos los sujetos de la población; cuando hacemos eso estamos haciendo un Censo, de hecho, la estadística surge de los censos. En la realidad se utiliza muy pocas veces por sus altos costes y sus métodos invasivos) y muestra (subconjunto de la población que pretende ser representativo, se usa en lugar de los censos y podemos conseguir casi tanta información como con los censos. Una muestra es representativa si tiene las mimas características que la población −círculo vicioso−). Dentro de la teoría de muestreo hay que hacer referencia a:*

- *Representatividad de la muestra. La estadística inferencial sólo sirve si la muestra es representativa y para averiguar si lo es hay que tener mucho cuidado escogiendo la muestra. Hay algunas técnicas que dan sesgos de muestreo que nos pueden llevar a error. Cuando hablamos de sesgo no podemos hablar de mala intención, sino que simplemente de forma involuntaria la muestra acaba con un sesgo. Ejemplos de sesgo son el sesgo de selección y el sesgo de la no respuesta.*
- *Aleatoriedad de la muestra. Hay dos tipos de muestreo:*

    - *Probabilístico (todos los sujetos tienen la misma prob. de ser elegidos).*

- *Aleatorio Simple. Necesitamos conocer la población y poder numerarlos, vamos sacando por medios mecánicos los números.*
- *Aleatorio Sistemático. Necesitamos conocer el listado de elementos que componen la población, extraemos un solo elemento (i) y el resto de componentes surge de sumar una constante k, que se consigue con la fórmula k = N / n , siendo N el tamaño de la población y n el tamaño de la muestra. Tiene problemas tales cómo el que los datos vengan ordenados.*
- *Estratificado. A la hora de hacer la muestra vamos a considerar grupos/categorías que ya existan en la población, como por ejemplo el sexo. Debemos asegurarnos de que todas las categorías estén presentes en la muestra final, y esos estratos tienen que ser tales que sean exclusivos y exhaustivos (no puede haber sujetos en más de un extracto ni sujetos que no tengan ninguno). Dentro de cada extracto se usa un método cualquiera de éstos tres (Afijaciones):*

Af. Simple − Dividir tam. de muestra entre nº de estratos.

Af. Proporcional − Se tiene en cuenta el tam. de estratos.

Af. Óptima − Tiene en cuenta tam. y homogenenidad de los estratos, pero su problema es conocer la homoogeneidad de los estratos.

- *Muestreo por Conglomerados. Es un subconjunto de elementos formado de forma más o menos natural (Ej. departamentos de una facutlad), cuando muestreamos ciudades o similares es muestreo por áreas. Una vez creado el conglomerado se escogen todos los sujetos que forman parte de ese subgrupo. Las ventajas son que no necesitamos conocer todos los individuos de la población, pero sí los conglomerado. Encontrar todos los elementos es complicado y por ello se puede hacer Poretápico, que va de lo general a lo individual. Ej. en vez de buscar profesores buscas institutos y luego se escogen dentro de los institutos elegidos.*

    - *No probabilístico (no tienen la misma prob. Dudas sobre su representatividad de la población, más sencillos de hacer):*

- *Muestreo por cuotas. Se basa en que tenemos un buen conocimiento de los estratos que forman una población y además sabe qué sujetos son adecuados para el tipo de investigación que queremos hacer, sólo que la asignación de sujetos no se hace aleatoriamente. Se define una cuota (un tipo de sujetos) y en la investigación se cogen a los primeros sujetos que cumplen los requisitos. Esto elimina la igualdad de probabilidad.*
- *Muestre opinático ó intencional. Se establece un sujeto tipo y se va a por ello. Esto se hace en sondeos electorales.*
- *Muestreo Casual. Coges al que puedes, y un tipo especial son los voluntarios. Lo malo es que el sujeto tiende a hacer lo que cree que el investigador necesita.*
- *Bola de nieve. Uno contacta con unos pocos sujetos que le ponen en contacto con otros sujetos, creciendo la lista como una bola de nieve. Se usa en cosas como los temas de drogas.*
- *Tamaño de la muestra. El ¿Cuantos? va ligado al nivel de error que vayamos a estar dispuestos a admitir. Cuanta más precisión queramos más sujetos necesitamos. Cosas que influyen son:*
- *Variabilidad de la Población (Varianza poblacional...pero es imposible conocerla, luego hay que basarse en estudios previos).*
- *Tipo de muestreo.*
- *Nivel de confianza en el que queremos trabajar.*
- *¿Qué queremos saber?*

Nivel de Confianza + Nivel de Error = 100. El nivel habitual de trabajo en Psicología es del 95% o del 99%.

- Estadístico: *valor numérico que define una característica de una muestra.*

- Distribución Muestral: *distribución teórica que asigna una probabilidad concreta a cada uno de los valores que puede tomar un estadístico en todas las muestras del mismo tamño que es posible extraer de una determinada población.*
- Desviación Típica Poblacional (n), S Desviación Típica Muestral (n−1). Z = Xmedia − / (/"n).
- Estimación de parámetros. Primero hacemos una muestra y luego estimamos. Un estimador tiene cuatro características que debe cumplir:

- Debe tener carencia de sesgo (ser insesgado), es decir, que su valor esperado coincida con el parámetro que se estima.
- Debe tener consistencia (aumenta su eficacia conforme aumenta el tamaño de la muestra).
- Eficiencia (un estimador es más eficiente cuánto menor es su varianza).
- Que sea eficiente (un estimador es eficiente si la estimación no puede ser mejorada).

E (^S2) = varianza insesgada de la muestra = a la de la población.

E (S2) " varianza sesgada de la muestra " a la de la población.

Estimación Puntual. La estimación puntual consiste en atribuir a un parámetro poblacional (aunque suene redundante) el valor concreto tomado por un estadístico tomado en la nuestra como estimador.

Estimación por Intervalos. Consiste en atribuir al parámetro que se desea estimar un rango de valoes entre los que se espera que se pueda encontrar el verdadero valor del parámetro con una probabilidad alta y conocida. Los límites del intervalo son Xmedia + error típico y Xmedia − error típico.

Nivel de Confianza (1−) probabiilad de que acertemos nuestro pronóstico.

Nivel de Riesgo ( ) probabilidad de cometer un error.

Li = Xmedia − |z | x / "n

Ls = Xmedia + |z | x / "n

- x − / x . Siendo por orden: media muestral, media poblacional y error típico. El erro típico se halla con las fórmulas n−1 / "n ó ^S / "n.
- En estadística, seguir un criterio conservador es aceptar poco riesgo, se considera poco riesgo a partir de 0'05, pero es más prestigioso trabaja con 0'01.
- Fórmulas para Inferir en proporción.

Li = p − z "p(1−p) / n

Li = p + z "p(1−p) / n

- Planteamiento de Hipótesis.

- Hipótesis Nula. *No cambia nada (Ho)*
- Hipótesis Alternativa. *Sí cambia algo (H1).*

El contraste de hipótesis es un proceso de decisión en el que una hipótesis es puesta en relación con los datos empíricos para determinar si es o no compatible con ellos (Teoría de la Decisión Estadística o TDE). Los supuestos de un contraste de hipótesis son afirmaciones que necesitamos establecer para conseguir determinar la distancia de probabilidad sobre la que se basa nuestra decisión de Ho. Tiene tres fases: contraste, comparar datos y determinar si es compatible.

- Estadístico de Contraste: *es un resultado muestral que cumple una doble condición, por un lado proporcionar información empírica relevante sobre la opción propuesta en la Ho y por otro poseer una distancia muestral conocida. Al intervalo de confianza lo llamamos 1 −* , también se le puede llamar zona de aceptación. A se le llama Zona de Riesgo y supone la aceptación de H1).
- Regla de Decisión. *Consiste en rechazar la hipótesis nula si el estadístico de contraste toma un valor perteneciente a la zona crítica o de rechazo, y también mantener la hipótesis si el estadístico de contraste toma un valor perteneciente a la zona de aceptación. Aceptar Ho no implica cambios, lo interesante es que caiga en la zona de rechazo. Cuando decidimos mantener una Ho queremos significar con ello que consideramos que esa hipótesis es compatible con los datos, en cambio cuando la rechazamos consideramos probado que esa hipótesis es falsa.*
- Errores:

- Tipo I. *Es el que se comete cuando se decide rechazar una hipótesis nula que en realidad es verdadera. La probabilidad de cometer el error tipo I es* .
- Tipo II. *Es el que se comete cuando se decide mantener una hipótesis nula que en realidad es falsa. A la probabilidad de cometer el error tipo II se le llama* .

El riesgo se hace pequeño cogiendo un 95% o un 99% y el riesgo se logra haciendo grande el , así que se busca un punto idóneo, que suele ser el 0'05 o 0'01. Para minimizar el riesgo se suele conseguir un N mayor, un tamaño de muestra más grande. Otra forma es que haya mucha desviación típica.

	Ho Verdadera	Ho Falsa
Se acepta Ho	*Correcto*	*Error tipo II*
Se rechaza Ho	*Error tipo I*	*Correcto*

CONTRASTE DE HIPÓTESIS SOBRE UNA MEDIA

- Hipótesis.

Ho: = o Ho: " o

Ho: " o Ho: > o

Ho: " o Ho: < o

- Supuestos
- población de partida normal.
- muestra aleatoria de tamaño n.
- Estadístico de Contraste.

T = Xmedia − / ^S / "n ó Xmedia − / ^S / "n−1.

- Distribución Muestral.

T se distribuye según T −1.

- Zona crítica.

  - Contraste Bilateral. T " 1− /2t n−1 y T " 1− /2t n−1
  - Contraste unilateral derecho. T " 1− t n−1
  - Contraste unilateral izquierdo. T " 1− t n−1

CONTRASTE DE HIPÓTESIS SOBRE UNA PROPORCIÓN

II. Supuestos

- La variable aleatoria es dicotómica o dicotomizada (p+q = 1) en la población es la verdadera proporción de éxitos.
- Muestra aleatoria simple de n observaciones con probabilidad constante de éxito cada ensayo.

III. Estadístico de Contraste.

T = P − / " (Ho (1−Ho) / n).

IV. Distribución Muestral.

Z se distribuye según N (0,1).

V. Zona crítica.

- Contraste Bilateral. T " 1−Z/2t n−1 y T " 1−Z/2t n−1
- Contraste unilateral derecho. T " 1−Zt n−1
- Contraste unilateral izquierdo. T " 1−Zt n−1

Estadística no paramétrica o no normal

- Tenemos que trabajar sobre una tabla de contingencias y tener en cuenta Fo (Frecuencias Observadas, es decir, lo que vemos en la muestra) y Fe (Frecuencias Esperadas, lo que debería ocurrir). Existen dos tipos de estadísticos para hallar distintos datos, ambos usan la tabla de y los grados de libertad se calculan por *número de columnas −1 multiplicado por número de filas −1*.
- Contraste de Hipótesis sobre Independencia. *"" (fo − fe)2 / fe =* . Es el valor de el que se contrasta como antes hacíamos con las z.
- Prueba de Bondad de Ajuste *" (fo − fe)2 / fe =* .