Tema 3: POBLACIÓN Y MUESTRAS

Podemos explicar estas dos palabras en una sencilla frase: usamos la información que nos facilita un grupo reducido de personas (muestra) para estimar lo que piensa, hace, opina un grupo mucho más amplio (población).

Un concepto importante para entender todo lo relativo al muestreo es, para empezar, el de tipificación de variables. Puesto que las variables vienen expresadas en unidades distintas, para poder compararlas tendremos que pasarlas a una unidad común. A esta operación se le llama tipificación.

Después de esto, podremos pasar a ver las distintas formas que puede adoptar una distribución de frecuencias, centrándonos en una distribución normal, por se la distribución teórica que va a sustentar toda la teoría del muestreo.

La finalidad es, conocido lo que piensa una muestra, inferir − estimar lo que piensa una población.

* Tipificación − estandarización de las variables

Las variables que construimos en la investigación social vienen expresadas en unidades distintas, y con medias y desviaciones típicas también diferentes, lo que hace imposible su comparación. Para solucionar esto lo que hacemos es la tipificación.

Mediante la tipificación o estandarización de las variables, creamos nuevas variables expresadas en unidades de desviación típica (identificadas por la letra Z), y se realiza dividiendo la diferencia de cada valor con respecto a la media, por su desviación típica.

- Propiedades variables tipificadas

. Su media es igual a cero y la desviación típica igual a 1.

. permite conocer la proporción de personas comprendidas en cualquier intervalo de la distribución (sólo aplicable a distribuciones normales).

*** La distribución normal**

A la hora de resumir variables, se suele calcular alguna medida de tendencia (como la media), otra de dispersión (como la desviación típica) y una más relacionada con la forma de la distribución. De todas las formas que puede tomar una distribución, nos centraremos en la normal.

La distribución normal es una curva de gran interés, se utiliza como histograma ideal con el que comparar los histogramas de nuestros datos.

Al tipo de variable que tiene un número infinito de alternativas de respuesta se le llama continua (edad, altura,peso...). es con estas variables, de naturaleza teórica, con las que tiene sentido pensar una distribución normal, igual de teórica.

- Propiedades de la distribución normal:

. Simétrica: se puede dividir en dos mitades iguales, simétricas.

. Conocidas la media y la desviación típica de una distribución normal, podemos calcular la proporción de casos existente en cualquier intervalo de la distribución.

- **Cálculo de la proporción de casos (áreas de la curva) en una distribución normal.**

De todas las posibles distribuciones normales existentes trabajamos con la distribución normal tipificada − estandarizada, con la variable Z, con media 0 y desviación típica 1.

− Uso de la tabla normal

Para calcular un área de la curva, la proporción de casos o la probabilidad de obtener un caso en un intervalo determinado, que todo es lo mismo, tendríamos que entrar en un problema de integrales. Para evitarlo, existe una tabla que muestra la proporción de casos existentes en cualquier intervalo de la distribución.

En los márgenes de la tabla se incluyen los valores de Z; en la vertical las unidades y el primer decimal y en las horizontales el segundo decimal; en el centro de la tabla se muestra la proporción de casos o lo que es lo mismo, la probabilidad de obtener un caso o el área de la curva para un valor $Z \leq Zi$.

- **Cálculo de los intervalos correspondientes a áreas o proporciones de casos en una curva normal.**

La operación contraria es calcular el intervalo en el que está comprendida una determinada proporción de casos y se hace calculando el valor del percentil o calculando el valor de los intervalos centrales.

*** Otras distribuciones**

La distribución normal, no es la más normal de las distribuciones. En la práctica, raro es encontrar distribuciones normales. Se utiliza como referencia para hablar de otro tipo de distribucines.

- Distribuciones simétricas − asimétricas

Cuando una distribución no se puede partir en dos mitades iguales, es asimétrica. Si la mayoría de los individuos se sitúan en torno a los valores inferiores de la variables, mientras que unos pocos se decantan por el extremo superior de la distribución, tendremos asimetría positiva. En caso contrario, será negativa. En la positiva, la media será superior a la mediana. En la negativa, a la inversa. Para saber si una distribución es simétrica o asimétrica (y de que tipo de asimetría se trata) hay que calcular el coeficiente de simetría.

*** Poblaciones y muestras**

- **Nociones básicas.**

Población es el conjunto de unidades de análisis que queremos investigar, al ser un conjunto tan numeroso no podemos observar a todos sus elementos, así que podemos seleccionar un número menor de unidades para proceder a su estudio con la confianza de que las conclusiones obtenidas puedan ser generalizadas al total de las unidades. A esto se le llama muestra.

A medidas tales como la media, cuando tratemos con muestras, las denominaremos estadísticos, mientras que si tratamos con población, se llamarán parámetros.

− Muestreo aleatorio simple.

Selección de n de N elementos de tal manera que cada uno de ellos por separado, además de cualquier combinación que podamos establecer, tenga la misma probabilidad de ser elegido.

La selección de los individuos mediante este tipo de muestreo se realiza asignando a cada individuo de la población un número; los números se van seleccionando aleatoriamente, de dos maneras posibles:

. eliminando aquellos que son elegidos para que no puedan ser reelegidos (*sin reemplazamiento).*

. sin eliminar los que ya han sido elegidos (*con reemplazamiento*).

− Utilización del muestreo.

Surgen tres problemas:

. *la estimación*: con resultados de la muestra estimar lo que hará la población

. *el contraste, prueba o test de hipótesis*: preguntarnos si la diferencia entre dos muestras aleatorias es estadísticamente significativa

. *diseño de muestras*: para poder estimar, las muestras deben cumplir ciertos criterios, es decir, hay que definir el número de casos y la forma de seleccionarlos

- **Fundamentos del muestreo**

La deducción y el cálculo de probabilidades son los fundamentos del muestreo que hacen posible estimar valores (parámetros) poblacionales y contrastar hipótesis a partir de valores (estadísticos) de las muestras.

Conocido lo que piensa una población (conocidos sus parámetros), el cálculo de probabilidades permite deducir qué es lo más probables que piense una muestra sacada de esa población. A la inversa, conocidos los estadísticos de la muestra se puede inferir−estimar cuáles serán los parámetros de la población.

- **De la población a las muestras**

Al sacar no una sino muchas muestras de una población, de la que conocemos su media y desviación típica, no conoceremos con exactitud la media que va a tener cada una de las muestras, pero sí que podremos calcular la media y desviación típica de todas ellas.

A condición de que las muestras sean grandes, o la población de la que se sacan sea normal, la distribución de las infinitas muestras sacadas de una población es normal, con media y desviación típica conocidas.

- **Estadísticos básicos de una distribución muestral**

. *Población y muestra*

. *Distribución muestral*: es fundamental en toda la estadística teórica o inferencial, y hay que distinguirla claramente de las otras dos distribuciones que se han visto hasta ahora: la distribución poblacional y la distribución de una muestra. La distribución muestral es una distribución teórica, se trata de la distribución del estadístico estudiado que obtendríamos si sacáramos infinitas muestras de una población.

. *Estadísticos de la distribución muestral*

+ El valor esperado (la media) de un estadístico obtenido a partir de muestras aleatorias sacadas de una población, es igual al parámetro de esa población

+ La desviación típica de las medias: será igual a la desviación típica de la población partida por la raíz

cuadrada del número de casos de la muestra. Es válida siempre que las muestras sean grandes o se hayan obtenido con reemplazamiento de las unidades seleccionadas.

· *Fluctuación de las medias*: la media de una muestra sacada de la población fluctuará en torno a la media de la población con una dispersión (desviación típica) conocida. Esto ocurre cuando son circunstancias especiales como S=0 y n=N

· *La red*: puesto que conocemos el grado de fluctuación de las medias, medido a través de la desviación típica. En lugar de decir que la media de la muestra es igual a la media de la población, diríamos que se encuentra en un determinado intervalo, que tiene como valor central la media de la población.

- **Normalidad de la distribución**

Conocidas la media y la desviación típica de una población es posible calcular la media y la desviación típica de la distribución de cualquier estadístico obtenido a partir de un número infinito de muestras sacadas de esa población.

- El teorema central del límite: cuando se sacan infinitas muestras de una población aproximadamente normal, o cuando las muestras son suficientemente grandes, la distribución de cualquiera de sus estadísticos (media, mediana, coeficiente de correlación, coeficiente de regresión,...) tendrá forma normal

- **Factores que influyen en la deducción**

- Error muestral: compuesto por nivel de confianza y error típico (compuesto por S y V n). Según la teoría de las muestras, el problema del error de muestreo está controlado: para deducir un estadístico, cuanto más grande sea la muestra, mejor. Que las muestras sean más o menos grandes es cuestión de dinero.

Si una población tiene una variabilidad nula, siempre se acertará a la hora de deducir la media de cualquier muestra que se extraiga, sin importar el tamaño que tenga: su media será igual a la de la población. Si la variabilidad es grande, será difícil acertar cuando se trate de deducir la media de una muestra extraída de esa población.

La variabilidad influye en el error muestral, cada población tienen su variabilidad, sin que sea algo que dependa del investigador.

DE LA MUESTRA A LA POBLACIÓN

Selección de los estimadores

Al hablar de la estimación estamos justo ante el problema inverso a la deducción, conocida la muestra, ¿qué podemos decir de la población?.

Criterios del buen estimador:

- Insesgado

El sesgo de un estimador es la diferencia que hay entre el valor esperado del estadístico muestral y el parámetro de la población.

Para estimar la desviación típica de la población podemos utilizar la desviación típica de la muestra, al a que

se le resta una unidad en el denominador para hacer que este estimador resulte insesgado.

Definición de sesgo: el sesgo de un estimador es la diferencia que hay entre su valor esperado (su media) y el valor del parámetro poblacional. El sesgo puede ser producto del muestreo o de la medición de los individuos.

- Eficiente

El estimador más eficiente es aquel que tiene mínima variabilidad (desviación típica) muestral − también mínimo error típico. Aunque media y mediana son estimadores insesgados de la media de la población, el primer estadístico es más eficiente que el segundo.

Definición de eficiencia o precisión: la precisión de un estimador es igual a la fluctuación que tiene en torno a la media de la distribución muestral. El azar provoca la fluctuación y el error típico la mide.

El lado práctico del sesgo y la precisión

- El sesgo

El sesgo,como problema relacionado con el muestreo surge siempre que no se respeta el principio de que todos los individuos de la población han de tener la misma probabilidad de ser elegidos.

- La precisión

Siempre que hay muestras hay estimadores que fluctúan (varían) alrededor del parámetro de la población. Esta fluctuación es el error típico, o su cuadrado, la varianza del estimador. Este componente del error de muestreo es difícil de evitar; lo único que se puede hacer es diseñar muestras en las que su valor sea lo menor posible. Por ejemplo, el muestreo estratificado.

- Errores fijo y variable

De los dos componentes del error total de muestreo, el sesgo es un error fijo, mientras que el error típico del estimador es un error variable. El sesgo es un error que se produce sistemáticamente en todas las muestras que sacamos de una población.

Intervalos de confianza

A partir del estadístico que obtenemos en la muestra se puede estimar el parámetro de la población de dos formas distintas: puntual y por intervalo. En el primero se estima que el parámetro de la población tiene el mismo valor que el estadístico de la muestra. En el segundo se dice que el parámetro poblacional estará en un intervalo que tienen como punto central el estadístico en cuestión.

Contraste de hipótesis

Está plagado de múltiples situaciones que pueden dar lugar a distintas hipótesis, con supuestos de dudosa verificación, y además tienen una solución poco satisfactoria.

Tres vías de actuación:

- En un enfoque moderno de los contrastes lo que se hace es, además de formalizar en forma de etapas las explicaciones que hemos dado hasta ahora, calcular un valor−P, que mide la probabilidad de ocurrencia del estadístico, a partir de una muestra sacada aleatoriamente de la población de partida. Basándose en esta probabilidad el investigador saca sus conclusiones.

- En el enfoque clásico, además de formalizar el proceso y cuantificar la probabilidad de ocurrencia, se marca un tope de "rareza", antes incluso de realizar la investigación, llamado región crítica, que sirve para tomar una decisión en cada caso: si la rareza de nuestro estadístico sobrepasa el tope marcado negamos su procedencia de la población de partida; en caso contrario aceptamos dicha procedencia.
- Utilizar los intervalos de confianza como forma de realizar contrastes de hipótesis.

Contraste: medir la probabilidad de que la diferencia entre el parámetro poblacional y el estadístico que se obtienen en una muestra sea fruto del azar.

Los contrastes de hipótesis (contraste de dos colas)

- Formular modelo e hipótesis

Lo primero que hará el técnico es definir claramente el modelo y las hipótesis de su contraste. Los contrastes siempre se hacen para resolver dudas, pero partiendo de algunas bases ciertas. Las dudas son las hipótesis. Las certezas representan el así llamado modelo del contraste. Cada contraste tienen sus certezas y sus dudas.

- Modelo: con relación a la muestra, vamos a dar por supuesto que las personas se han seleccionado mediante un procedimiento aleatorio simple.
- Hipótesis nula y alternativas: para rechazar la hipótesis nula es necesario decidir previamente cuál va a ser la hipótesis alternativa. Si pensamos adoptar la hipótesis alternativa A, rechazaremos la hipótesis nula cuando el estadístico obtenido en la muestra sea significativamente distinto que la proporción postulada en la hipótesis nula. Ello ocurrirá siempre y cuando el estadístico sea mucho mayor o menor que el 10 %.

A este tipo de contraste que rechaza la hipótesis nula cuando el estadístico obtenido en la muestra es muy distinto del parámetro postulado en el modelo, se le denomina contraste, prueba o test de dos colas.

- Cálculos de la distribución muestral y de la probabilidad de obtener nuestro estadístico al azar.

Para ver lo que tienen de normal o raro el resultado obtenido en nuestra muestra tenemos que compararlo con lo que habría ocurrido si hubiéramos sacado muchas muestras de la población modelo.

- Región crítica y nivel de significación del contraste

Conocidas las hipótesis y la distribución muestral hay que decidir cuándo vamos a rechazar nuestra hipótesis nula. Hay que marcar una región crítica en la que, si cayera nuestro estadístico, rechazaríamos.

- Toma de decisión

Con la información de la que disponemos ya podemos decidir si vamos a rechazar nuestra hipótesis nula.

Intervalos de confianza y contrastes de hipótesis (contraste de dos colas)

Una forma alternativa de contrastar una hipótesis es utilizar los intervalos de confianza. Quizá sea la forma más sencilla de proceder.

El lado práctico de los contrastes

Los contrastes tienen varios problemas. El primero de ellos tienen que ver con la diferencia que hay entre la significación estadística y la significación sociológica.

- Significatividad estadística frente a significatividad sociológica

Un contraste puede decir que el resultado de un análisis es estadísticamente significativo sin que por ello podamos decir que este resultado tenga significación sociológica.

- El (mal)uso de los contrastes

No es solución realizar contrastes cuando no se cumplen supuestos exigidos, práctica muy habitual en la investigación social. Si el contraste tienen algún sentido es porque permite cuantificar el riesgo asociado a la toma de unas decisiones que de otra manera habría que adoptar con el único criterio de la intuición.

La aleatoriedad es la única garantía que tenemos para conseguir que la muestra sea representativa. En la encuesta se preguntan cosas que desconocemos a nivel de toda la población y cosas de las que tenemos un conocimiento cierto.

Problemas de muestreo

Poblaciones y muestras pequeñas

En la investigación social, especialmente si se lleva a cabo mediante la técnica de la encuesta, suele ser normal estudiar grandes poblaciones, utilizando grandes muestras.

+ Poblaciones pequeñas y muestreo sin reemplazamiento

- Factor de corrección

El muestreo puede ser con reemplazamiento y sin reemplazamiento. Desde el punto de vista de la eficiencia, los estimadores que se obtienen con el muestreo sin son más eficientes que los obtenidos con el muestreo con.

La importancia de la reducción que se opera en el error típico no depende tanto del tamaño absoluto de la población como de su tamaño relativo: la reducción es más importante cuanto menor sea la diferencia entre población y muestra. Se dan dos situaciones límite:

. Cuando población y muestra son iguales, el factor de corrección tienen un efecto de reducción total, puesto que el error típico se hace igual a cero.

. Cuando la población e smuy grande y la muestra es muy pequeña la reducción del error típico apenas si se nota, pues el valor del factor de corrección es aproximadamente igual a 1.

- Fracción de muestreo

Una forma alternativa del factor de corrección, introduciendo la idea de fracción de muestreo, que es la razón entre el tamaño de la muestra y el tamaño de la población, o el número de individuos de la población a los que representa cada individuo de la muestra. Muestra muy claramente la importancia del tamaño relativo de la población.

+ Muestras pequeñas y desviación típica de la población desconocida: la t de Student.

Siempre que vayamos a estimar el parámetro de una población utilizando intervalos de confianza o queramos contrastar una hipótesis necesitamos conocer la variabilidad de la población.

Si desconocemos la media de la población malamente vamos a conocer su desviación típica. En este caso lo

que hacemos es sustituir la desviación típica de la población por su mejor estimador, la desviación típica de la muestra., pero añadiéndole a la estimación una nueva incertidumbre.

El error que se introduce, por la diferencia que pudiera haber entre ambas cantidades, queda minimizado al dividirlo por la n de una muestra muy grande; sin embargo cuando la muestra es pequeña este error puede tener importancia. En este caso lo que hacemos es construir intervalos de confianza más amplios que tengan en cuenta una nueva fuente de error. Para ello sustituimos los valores Z de la distribución normal por los valores t de una nueva distribución, llamada de Student.

La distribución de la t de Student no es única; existen tantas distribuciones como tamaños de muestra. Los diferentes tamaños de muestra reciben el nombre de grados de libertas y su valor es igual a n−1.

+ Muestras pequeñas de poblaciones no normales

para que el teorema central del lómite sea operativo se necesita que las muestras sean 2suficientemente" grandes o que la población de la que se extraen sea aproximadamente normal. Si las muestras son pequeñas, por debajo de los 30 casos, la distribución muestral deja de tener forma normal para pasar a adoptar la forma de la t de Student.

Muestras con distintas probabilidades de selección de los individuos

Para que la muestra sea representativa de la población la selección de sus elementos ha de hacerse aleatoriamente y dándole a cada uno de ellos la misma probabilidad de ser elegido. Cuando no ocurre así, los estadísticos que se obtengan en la muestra serán estimadores sesgados de sus respectivos parámetros poblacionales.

Ej: Muestreo estratificado no proporcional

La involuntaria desigual representación de ciertos colectivos en la muestra puede ser producto de :

- la no respuesta.
- El mal trabajo de campo.
- El uso de marcos muestrales deficientes.

Elevaciones de la muestra a la población

El Instituto Nacional de Estadística trabaja con muestras, pero da los resultados a nivel de la población. Utiliza elevadores para sacar los números de la población.

Los elevadores son sencillos de calcular, puesto que no son otra cosa que los pesos según el peso de cada individuo.

El tamaño de la muestra

Al hablar de los factores que influyen en el error típico del muestreo hamos visto la importancia que tienen el tamaño de la muestra.

- poblaciones grandes: cuando el tamaño de la población es muy grande, caso de las encuestas sociológicas a la población española, sustituimos el error típico por su varianza.

. Fracciones de muestreo grandes: lo son siempre que el tamaño de la población es muy pequeño, comparado con el tamaño de la muestra.

Existe un estadístico, llamado coeficiente de variación, que sirve para calcular el valor relativo del error típico, lo mismo que servía para calcular el valor relativo de la desviación típica.

Para calcular el tamaño de la muestra en función del coeficiente de variación deseado también procedemos de manera distinta según qué circunstancias tengamos: población o fracción de muestreo grandes.

La potencia de los contrastes

Se trata de evitar que por alta de muestra lleguemos a la conclusión de que un estadístico no es estadísticamente significativo cuando realmente lo es.

Nos permite determinar el tamaño de las muestras sobre una base complementaria a la disminución del error de muestres; obliga a salir de la rutina en la que se ha instalado gran parte de la investigación social, que lleva a contrastar hipótesis sobre una base exclusivamente estadística.

El problema de la potencia está en que si no rechazamos la hipótesis nula llegaremos a la conclusión de que el nuevo sistema no ha tenido efecto, cuando quizá si que lo contenga, sólo que no la rechazamos debido a la baja potencia del contraste.

La potencia viene determinada por:

- el tamaño de la muestra
- nivel de significación.
- La naturaleza de las hipótesis alternativas.
- La misma naturaleza del contraste estadístico: paramétrico frente a no parmétrico.

+ Tamaño del efecto:

Determinar el tamaño del efecto es la parte del cálculo de la potencia de los contrastes en la que encuentran mayor dificultad los investigadores.

Capítulo 4: LA OPERACIONALIZACIÓN DE CONCEPTOS

4.1.− Fundamentos y principios de la operacionalización

Del marco teórico de la investigación extraemos unos conceptos y proposiciones. Los conceptos se traducen en términos operacionales. De ellos se deducen unas variables empíricas o indicadores que posibilitan que contrastemos empíricamente el concepto que estamos analizando.

Según Blalock (1982) hay que diferenciar dos nociones dentro de la operacionalización: la conceptualización y la medición

- la conceptualización: es el proceso teórico por el que se clarifican las ideas. La mayoría de los conceptos constituyen variables latentes, no directamente observables, por lo que hay que concretar de manera precisa la traducción del concepto al indicador o variables empíricas que midan las propiedades latentes enmarcadas en el concepto.
- La medición: es el proceso que vincula las operaciones físicas de medición con las operaciones matemáticas de asignar números a objetos

En toda operacionalización de conceptos teóricos se ha de partir de:

- entre los indicadores y el concepto a medir ha de haber una plena correspondencia, para que su

representatividad sea válida y fiable
- los indicadores pueden materializarse en formas diversas (cuestionario, entrevista,...) dependiendo de la técnica de obtención de datos utilizada
- en la operacionalización se asumen unos márgenes de incertidumbre. La relación entre los indicadores y el concepto siempre es supuesta., hay que intentar reducir el error de medición al mínimo posible

4.2.− La medición de variables: tipologías

Una variable es cualquier cualidad o característica de un objeto que contenga, al menos, dos atributos en los que pueda clasificarse. Por lo tanto, los atributos son los distintos valores o categorías que componen la variable. Por ejemplo, variables como la edad toma el valor numérico de años cumplidos; mientras que la variable sexo toma como valores hombre−mujer.

Por lo tanto, medir una variable, consiste en asignarle valores. Para que la medición sea adecuada hay que cumplir tres requisitos:

- exhaustividad: la variable debe comprender el mayor número de atributos o valores posible
- exclusividad: los atributos de una variable deben ser mutuamente excluyentes
- precisión: realizar el mayor número de distinciones posibles

Hay distintas modalidades de variables según los criterios de clasificación de las mismas:

* Según el nivel de medición (forman una escala acumulativa, cada nivel comparte las propiedades de los niveles que le preceden)

·Variables cualitativas

- *variables nominales*: sus atributos sólo cumplen las condiciones de exhaustividad y exclusividad. Ejemplo: sexo, nacionalidad, grupo sanguíneo,...., cualquier variable que indique una cualidad y no establezca graduación entre sus atributos
- *variables ordinales*: sus atributos cumplen las condiciones de exhaustividad y exclusividad y además se pueden ordenar en el sentido mayor que, menor que aunque no se conoce la magnitud exacta que diferencia un atributo de otro. Las variables ordinales expresan una cualidad, no una cantidad. Ejemplo: clase social, nivel de estudios,...

· Variables cuantitativas

- *variables de intervalo*: en ellas podemos cuantificar la distancia exacta que separa cada atributo de la variable gracias al establecimiento de alguna unidad física de medición Ejemplo: años, horas, centímetros...
- *variables de proporción o razón*: podemos cuantificar la distancia exacta que separa cada atributo de la variable gracias al establecimiento de alguna unidad física de medición y además podemos establecer el cero absoluto. La mayoría de las variables de intervalo son, a su vez, de razón

* Según la escala de medición

- *variables continuas*: aquellas en las que pueden hallarse valores intermedios entre dos valores dados. Ejemplo: edad entre un año y otro hay meses
- *variables discretas*: no existe la posibilidad de hallar valores intermedios entre dos valores dados. Ejemplo: número de mesas en una clase

* Según su función en la investigación

- *variables independientes*, explicativa o predictoras (X): sus atributos influyen en los que adopta una segunda variable. Ejemplo: velocidad, estado del pavimento, consumo de alcohol, condiciones meteorológicas
- *variables dependientes* o criterio (Y): sus atributos dependen de los que adopten las *variables independientes*: ejemplo: accidente de tráfico
- *variables perturbadoras*: variables que median entre las independientes y las dependientes

* Según su nivel de abstracción

- *variables generales*: son tan genéricas y abstractas que no pueden ser directamente observadas. Ejemplo: estatus social
- *variables inmediatas*: expresan alguna dimensión de la variable genérica. Ejemplo: el nivel educativo para la medición de la variable estatus social
- *variables empíricas*: representan aspectos específicos de las dimensiones de una variable genérica. Son directamente medibles. Ejemplo: cursos académicos cumplidos como indicador para la dimensión nivel educativo

4.3.− De los conceptos teóricos a los indicadores e índices

En la operacionalización de los conceptos, tenemos dos momentos. En el primero proporcionamos una definición operativa (que comprenda el significado determinado que se da al concepto) y en el segundo, especificamos los indicadores que representaran a los conceptos.

Por lo tanto, podemos hablar de la delimitación de los conceptos en función a la definición:

- definición nominal: es la que se asigna a un concepto pero carece de las precisiones necesarias para medir los fenómenos a los que hace referencia
- definición operacional: especifica cómo se medirá la ocurrencia de un concepto determinado en una situación concreta.

En la operacionalización del concepto teórico encontramos:

- representación teórica
- especificación del concepto descomponiéndolo en dimensiones
- para cada dimensión seleccionar los indicadores
- sintetizar los indicadores estableciendo índices (medida común que agrupa a varios indicadores de una misma dimensión)

Para el cálculo de un índice se precisa que las distintas medidas se transformen en una escala de medición común. Este proceso de consecución del índice se llama *ponderación*.

A la hora de elaborar un coeficiente de ponderación hay que tener en cuenta:

- representar lo más fielmente la variable que se pondera y las diferencias de sus indicadores
- que el coeficiente sea sencillo, a ser posible un número entero y sencillo
- deben utilizarse los signos (+) y (−) para marcar dos significaciones bien distintas del índice
- los atributos iguales han de ponderarse de igual forma, esto permite la comparación posterior de los índices

4.4.− Cuestiones de validez y fiabilidad en la medición

cuando tenemos los indicadores hay que comprobar hasta qué punto la operacionalización de conceptos que

hemos hecho reúne las condiciones mínimas de validez y fiabilidad.

4.4.1.− La validez de la medición

para que un indicador sea válido ha de proporcionar una representación adecuada del concepto teórico que miden.

- *validez de criterio*: la validez se comprueba comparándola con algún criterio que anteriormente se haya empleado para medir el mismo concepto
    - ♦ *validez concurrente*: cuando se correlaciona la medición nueva con un criterio adoptado de un mismo momento
    - ♦ *validez predictiva*: concierne a un criterio futuro que esté correlacionado con la medida
- *validez de contenido*: concierne al grado en que una medición empírica cubre la variedad de significados incluidos en un concepto
- *validez de constructo o teórica*: cuando se compara una medida particular con aquella que teóricamente habría de esperar a partir de las hipótesis.

4.4.2.− La fiabilidad de la medición

esta característica supone que los resultados logrados en mediciones repetidas del mismo concepto han de ser iguales para que la medición sea fiable.

Para comprobar la fiabilidad podemos:

- aplicar el mismo procedimiento de medición en diferentes momentos
- método test−retest: administrar una misma medida a una misma población en dos períodos de tiempo diferentes
- método alternativo: analizar una misma población en momentos diferentes con distinto instrumento de medición
- método de las dos mitades: no se efectúan dos comprobaciones en períodos diferentes de tiempo, sino al mismo tiempo. Para ello se dividen los ítems totales en dos mitades y se correlacionan los resultados
- método de consistencia interna alpha de Cronbach: se obtiene calculando el promedio de todos los coeficientes de correlación posibles de las dos mitades, midiendo así la consistencia interna de todos los ítems.

Tema 5: <u>LA SELECCIÓN DE LAS UNIDADES DE OBSERVACIÓN:</u>

<u>EL DISEÑO DE LA MUESTRA</u>

1. FUNDAMENTOS Y CLARIFICACIÓN TERMINOLÓGICA

Una de las primeras decisiones a tomar en cualquier investigación es la *especificación* y *acotación* de la población a analizar, que vendrá determinada por cuál sea el problema y los objetivos principales de la investigación. Por población entendemos al conjunto de unidades, para las que se desea obtener cierta información. En la definición y acotación de la misma han de mencionarse características esenciales que la ubiquen en un espacio y tiempo concreto.

Una vez definida la población, se procede al diseño de la muestra, que es la selección de unas unidades concretas de dicha población. Un estudio de casos o un experimento impone menos exigencias en la muestra que una encuesta. Dicha representatividad estará subordinada el tamaño de la muestra y al procedimiento seguido para la selección de las unidades muéstrales. Si a partir de los datos obtenidos en una muestra, quieren

inferirse las características correspondientes a la población, es imperativo diseñar una muestra que constituya una representación a pequeña escala de la población a la que pertenece.

Cualquier diseño muestral comienza con la búsqueda de documentación que ayude a la identificación de la población de estudio. Con el término marco se hace referencia al listado que comprende las unidades de la población. De él se espera que sea un descriptor válido de dicha población, por lo que debe de cumplir varios requisitos mínimos: el marco ha de ser lo mas completo posible, ya que se encuentra limitado a un conjunto de la población; el marco muestral debe estar actualizado, para que las posibilidades de omisiones se restrinjan; se persigue una generalización de los datos muestrales, para que cada representante de la población este igualmente representado en el marco de muestreo, evitándose las duplicidades; tampoco se deben incluir unidades que no corresponden a la población que se analiza, porque estas reduce la probabilidad de la elección de las unidades que sí pertenecen a la población; Debe de contener información suplementaria para localizar las unidades seleccionadas; y ante todo, debe ser fácil de utilizar, porque reduce los costes del diseño de la muestra y contribuye a la reducción de errores.

2. EL TAMAÑO DE LA MUESTRA.

Una de las decisiones preliminares en cualquier diseño muestral es el número de unidades a incluir en la muestra. En esta decisión participan diversos factores como:

1.− *El tiempo y los recursos disponibles*; que se emplean para la materialización del estudio propuesto. En función de cuánta sea la dotación económica y los plazos temporales para cada fase de la investigación, el tamaño de la muestra variará.

2.− *La modalidad de muestreo seleccionada*; esta se halla determinada por los objetivos, el tiempo y los recursos dados para su realización. En general, los diseños muestrales no probabilísticos demandan un tamaño muestral inferior a los diseños probabilísticos.

3.− *La diversidad de los análisis de datos previos*; hay que anticipar la variedad de análisis que se estimen oportunos para la consecución de los objetivos de la investigación. Si el equipo investigador cree oportuno aplicar alguna técnica multivariable, deberá procurar que la muestra analizada incluya un numero elevado de casos. Para la realización del análisis multivariables se precisa una cierta proporcionalidad entre el numero de observaciones y el número de variables incluidas en el estudio.

4.− *La varianza o heterogeneidad poblacional*; Afecta al tamaño de la muestra. Cuanto más heterogénea sea la población, mayor será su varianza poblacional. Por lo tanto necesitaremos un mayor tamaño muestral para que la variedad de sus componentes se halle representada en la muestra. El conocimiento de la homogeneidad poblacional resulta tan primordial en la decisión del tamaño de la muestra, para acceder a dicho conocimiento podemos basarnos en: la experiencia adquirida en estudios que se repiten con periodicidad, (cuando ambas poblaciones coincidan) y en la realización de estudios pilotos previos a la investigación principal, que ayuden al cálculo de las varianzas de las variables de interés.

Cuando se desconoce el valor de la varianza poblacional, se recurre al supuesto mas desfavorable, que es tomar el producto de las probabilidades P y Q como equivalente a la varianza poblacional, presentando ambas probabilidades el valor 0,50. La fórmula común para el cálculo del tamaño muestral en universos infinitos, a un nivel de confianza de 2 sigma es $N=4PQ/E2$; donde E representa el error muestral.

5.− *El margen de error máximo admisible*; cuando se produce un incremento en el tamaño de la muestra, este repercute en una mayor precisión en la estimación de los parámetros poblaciones, es decir, en la reducción del error muestral, mientras que en muestras pequeñas, el error de la muestra aumenta, manteniendo constante la varianza poblacional. A medida que aumenta el volumen del tamaño de la muestra, se produce un decrecimiento en el valor del error muestral.

También se advierte que a partir del 2% de error, se disparan los crecimientos en el tamaño de la muestra para alcanzar una mínima ganancia en la reducción del error muestral.

6.− *El nivel de confianza de la estimación*; expresa el grado de confianza o probabilidad que el investigador tiene en que su estimación se ajuste a la realidad. Hay tres niveles de confianza comunes en la investigación social. Corresponden a áreas bajo la curva normal acotadas por distintos valores de desviación típica, denominada sigma ( ). El mas habitual es 2 , que supone un 95,5% de probabilidad de acertar en la estimación a partir de los datos muestrales.

La distribución normal representa una curva perfectamente simétrica, en forma de campana, que admite valores infinitos. El área total de la curva normal es 1, y en función del valor de Z variará la probabilidad concedida al evento en cuestión.

Todo esto participa en el cálculo del tamaño de una muestra probabilística. La fórmula genérica para una muestra aleatoria sería la siguiente:

- cuándo el universo este compuesto por más de 100.000 unidades: n=Z2S2/E2.
- cuándo el universo este compuesto por 100.000 unidades o menos, se tratará de una población finita: n=Z2S2N/E2(N−1)+Z2S2.

3.− EL ERROR MUESTRAL

Cuando se diseña una muestra el objetivo primordial es conseguir un elevado nivel de adecuación en la selección de la muestra, respecto de la población a la que se pertenece, sto se hace para que la investigación adquiera validez externa. Pero por muy perfecta que sea la muestra, como únicamente se analiza una parte de la población, siempre habrá alguna divergencia entre los valores obtenidos de la muestra y los valores correspondientes en la población. Esa disparidad se denomina *error muestral,* y es el grado de inadecuación que existe entre las estimaciones muestrales y los parámetros poblacionales.

Para el cálculo del error muestral se acude al estadístico llamado error típico, que mide la extensión a la que las estimaciones muestrales se distribuyen alrededor del parámetro poblacional. Se especifica que aproximadamente el 68% de las estimaciones muestrales se hallarán comprendidas entre el ±1 vez el error típico del parámetro poblacional; el 95,5% entre ±2 veces el error típico; y finalmente, el 99,7% entre ±3 veces el error típico. El nivel de confianza en la estimación aumenta conforme se amplia el margen de error. En el cálculo del error típico intervienen los elementos siguientes:

- *El tamaño muestral*, lo que determina el error muestral no es la población que constituye la muestra sino el tamaño *de la muestra. A medida que aumenta el tamaño de la muestra, decrece el error muestral.*
- El nivel de heterogeneidad de una población favorece el error muestral, excepto si se aumenta el tamaño muestral para incluir a todas las distintas variedades que componen el universo. El error muestral se halla mas presente en poblaciones heterogéneas que en universos homogéneos.
- *El nivel de confianza* adoptado, el cual si se aumenta, agranda el tamaño de la muestra, lo cual trae consigo la reducción del error muestral. Incrementos en el tamaño de la muestra conllevan una ampliación del nivel de confianza en la estimación muestral.
- *El tipo de muestreo* realizado, donde el error muestral también se halla afectado por el procedimiento de selección de las unidades muestrales. En general, el muestreo aleatorio estratificado es el que genera un menor error muestral. En cambio, el muestreo aleatorio por conglomerados es el que ocasiona un mayor error muestral. Aunque la agrupación de la muestra en conglomerados presenta la gran ventaja de reducir los costes del trabajo de campo, éste a su vez repercute en una desventaja importante: incrementa el error de la muestra.

Para la muestra aleatoria simple o sistemática, las fórmulas correspondientes al error típico serían las

siguientes:

	Universo infinito	**Universo finito** ("100.000 unidades)
Error típico de la media	E= "S2/n	E= "(S2/n)(N−n/N−1)
Error típico de una proporción	E= "PQ/n	E= "(PQ/n) (N−n/N−1)

Si la muestra fuese aleatoria estratificada proporcional, se introducirían modificaciones en la fórmulas:

- Del error típico de la media: E= " "n S2/n2.
- Del error típico de una proporción: " "nPQ/n2.

Donde P es la proporción de la muestra en el estrato que posee el atributo en cuestión; Q es igual a 1−P; S2 es la estimación de la varianza de la variable de interés para la población en el estrato en cuestión; " es el sumatorio de todos los estratos desde 1 hasta n; y n es el tamaño de la muestra total.

Por último, si la muestra fuese por conglomerados, la fórmula correspondiente el error típico sería la siguiente: E= " (1−(n/M)) (Sb2/m),

Donde M es el número de conglomerados en la población; m es el número de conglomerados seleccionados en la muestra; y Sb2 es la varianza de los valores del conglomerado.

4.− TIPOS DE MUESTREO: DISEÑOS MUESTRALES PROBABILÍSTICOS Y NO PROBABILÍSTICOS.

Las modalidades de muestreo son variadas, aunque podemos agruparlas en, probabilístico y no probabilístico. La elección de un tipo u otro de muestreo vendrá condicionada por la dotación económica, el tiempo programado para su ejecución, la existencia de un marco muestral válido y el grado de precisión que el investigador quiera dar a la indagación.

Muestro probabilístco o aleatorio

Se fundamenta en la aleatorización como criterio esencial de la selección muestral. Ello favorece que cada unidad de la población tenga una unidad igual de probabilidad de participar en la muestra, que la elección de cada unidad muestral sea independiente de las demás.

Este muestreo se adecua más a propósitos de estimación de parámetros y comprobación de hipótesis

Muestreo no probabilístico

La extracción de la muestra se efectúa siguiendo criterios diferentes de la aleatorización. Además repercute en la desigual probabilidad de las unidades de la población para formar parte de la muestra, en la dificultad de calcular el error muestral y en la introducción de sesgos en el proceso de elección muestral.

No obstante el muestreo no probabilístico presenta dos *ventajas* notorias: no precisa de la existencia de un marco de muestreo y su materialización resulta más sencilla y económica que los muestreos probabilísticos.

Este muestreo es más apropiado para la indagación exploratoria, estudios cualitativos y para investigaciones sobre población marginal, de difícil registro y localización.

4.1. Muestreo aleatorio simple.

Constituye el prototipo de muestreo, en referencia al cual se estiman las fórmulas básicas para el cálculo del tamaño y del error muestral. Su realización exige la existencia de una marco muestral que cumpla los

fundamentos y la clarificación terminológica. Una vez localizado, se asigna a cada unidad de la población un número identificador para proceder a la extracción aleatoria de los integrantes de la muestra. La selección muestral debe de garantizar que cada unidad de la población tenga una unidad igual de participar en la muestra, y que la selección muestral sea totalmente aleatoria hasta alcanzar el tamaño muestral fijado.

La elección de las unidades muestrales puede hacerse mediante ordenador (que es el que ejecuta todas las tareas correspondientes). Pero cuando el uso del ordenador no se considere viable, se recurre al procedimiento tradicional: utilizar una tabla de números aleatorios. Estas tablas comprenden múltiples combinaciones de números extraídos al azar. La actuación entonces seria: elegir un punto de partida, ya sea una columna o una fila cualquiera de la tabla (esto ya supone un sesgo); hacer que coincida el número de dígitos de la tabla con el número de dígitos de la población del marco; y que el individuo al que pertenece el número extraído pasará a formar parte de la muestra, salvo que en el marco no se adjunte un medio para su localización.

4.2. Muestreo aleatorio sistemático.

Es imprescindible un listado de la población, pero difiere del muestreo aleatorio simple en que: sólo la primera unidad se elige al azar y los restantes elementos de la muestra se obtienen sumando el coeficiente de elevación, hasta completar el tamaño muestral.

Si no se ha extraído un excedente de unidades muestrales a considerar para las sustituciones en el momento de la selección muestral ha de calcularse un nuevo coeficiente de elevación que permita una nueva selección sistemática de las unidades muestrales no cubiertas en el trabajo de campo.

4.3. Muestreo aleatorio estratificado.

Supone la clasificación de las unidades de población en un número reducido de grupos, en razón de su similaridad. Con esto se persigue que cada estrato tenga representación en la muestra final.

En el estratificado, la muestra se distribuye en diferentes grupos de población, en función de los valores que presente en las variables elegidas para la estratificación. Se hace siguiendo exclusivamente procedimientos aleatorios de selección muestral.

Lynn Lievesley (1991) destacó cuatro puntos básicos para el diseño de un esquema de estratificación:

1. Elección de las variables de estratificación, condicionada a aquellas comprendidas en el marco muestral de referencia.

2. Orden de las variables de estratificación, eligiendo la variable de mayor relevancia para la investigación en el primer estadio y así sucesivamente.

3. Número de variables de estratificación, pudiendo alcanzarse una mayor eficacia siguiendo un esquema de estratificación distinto para las variables incluidas en los diversos estadios de la estratificación.

4. Tamaño de los estratos, dividiendo la población en grupos de igual tamaño para que resulte más adecuada.

Si con la estratificación se persigue el logro de una mayor precisión de la estimación muestral, esta se alcanzará cuando se cumplan dos condiciones esenciales: sean máximas las diferencias entre los estratos y mínimas dentro de cada estrato. Y las variables de estratificación se hallen relacionadas con los objetivos de la investigación.

Las variables de estratificación más empleadas son las variables de sexo y edad, pudiendo añadirse otras variables como la clase social, la ocupación, el nivel de instrucción, etc.

Tras la clasificación de la población en estratos, se procede a afijar la muestra en cada estrato. Por afijación se entiende la distribución del tamaño muestral global entre los estratos diferenciados. Esta distribución se puede cumplir de tres maneras distintas:

- *Afijación simple,* el mismo tamaño de la muestra a cada estrato. Con ello se busca la igual representación de los estratos en la muestra global. Esta equidistribución del tamaño muestral conlleva un inconveniente importante y es que favorece a los estratos de menor volumen de población.
- *Afijación proporcional,* la distribución de la muestra se hace proporcional al peso relativo del estrato en el conjunto de la población. A los estratos que reúnan un mayor número de unidades de población les corresponderá un tamaño muestral superior al de aquellos que representen un porcentaje inferior en la población.
- *Afijación óptima,* donde se añade la variabilidad del estrato respecto a la variable considerada en la estratificación. En conformidad con este último criterio de afijación, les corresponderá un tamaño muestral superior a los estratos de mayor heterogeneidad y peso poblacional.

Las tres variedades de afijación pueden englobarse en dos amplias modalidades de estratificación:

− *Estratificación proporcional,* y se hace de manera que garantice una probabilidad igual de selección para todos los estratos.

− *Estratificación no proporcional,* donde la representación de los estratos en la muestral final no es proporcional a su peso es el conjunto de la población, al haberse dado una probabilidad desigual de selección en cada estrato.

Uno de los inconvenientes fundamentales de la estratificación no proporcional es la necesidad de ponderar la muestra y no se precisa de la ponderación cuando sólo se realizan análisis individuales y comparativos entre los estratos. Por ponderación entendemos el proceso de asignación de pesos a cada estrato, de manera que logre compensarse la desigual probabilidad de selección dada a cada unidad de población que compone el estrato. La ponderación puede efectuarse de varias formas, la más usual consiste en dividir el porcentaje que representa en la muestra.

4.4. Muestreo aleatorio por conglomerados.

Secciona a la población total en grupos como fase previa a la extracción muestral, como ocurre con el muestreo aleatorio estratificado. Si bien se diferencia en aspectos importantes, como que en el muestreo por conglomerados el error muestral disminuye conforme aumenta la heterogeneidad dentro del grupo, que en el muestreo por conglomerados lo que se extrae es una muestra aleatoria de conglomerados y la unidad del muestreo es el conglomerado. Los conglomerados pueden ser de las áreas geográficas que dividen a la población que se analiza, pero también organizaciones y instituciones.

Si a partir de una muestra por conglomerados, se extrae una nueva muestra, con referencia a cada uno de los conglomerados previamente elegidos, y así sucesivamente, se está ante un diseño muestral muy habitual en la investigación social: el muestreo **polietápico por conglomerados.**

El muestreo polietápico por conglomerados representa una extensión del muestreo por conglomerados. En él la unidad de muestreo final no son los conglomerados, sino subdivisiones de estos. Por lo que no se toman cada uno de los integrantes de los conglomerados elegidos aleatoriamente, sino sólo a una parte de ellos, escogidos también de forma aleatoria. La modalidad de muestreo polietápico más sencilla implica la extracción muestral en dos fases, una primera que selecciona las agrupaciones de los miembros de la población de estudio, que son análogas a los conglomerados; y una segunda fase donde se eligen aleatoriamente los miembros de la población a observar, de las unidades de muestreo primarias previamente

seleccionadas.

El muestreo aleatorio por conglomerados se muestra de especial interés cuando resulte difícil compilar una lista exhaustiva de todos los componentes de la población, cuando se quiera reducir la duración y los costes económicos del trabajo de campo en la investigación, y cuando se realicen estudios de ámbito nacional o internacional, que supongan una considerable dispersión de la muestra.

4.5. Muestreo por cuotas.

Una variedad de muestreo no probabilístico que parte de la segmentación de la población de interés en grupos, a partir de variables sociodemográficas relacionada con los objetos de la investigación. Su puesta en práctica necesita la elaboración de una matriz con las característica básicas de la población que se analiza. El propósito es seleccionar una muestra que se ajuste a la distribución de las características fundamentales de la población. Ello garantiza que en la muestra se encuentren representados los distintos grupos de población.

Por otra parte en la elección de las variables intervienen otros factores: la precisión que el investigador desee y la accesibilidad de las variables elegidas. Las cuotas más habituales son las determinadas por la conjunción de las variables sexo y edad. Una vez confeccionada la matriz, se calculan las proporciones relativas para cada celdilla de la matriz, a partir de la proporción que representa cada categoría de las variables seleccionadas en la población total.

Aunque el azar intervenga en las fases iniciales del diseño, la selección de los elementos concretos de la población es totalmente arbitraria. La única condición que se le impone es que la persona se ajuste a las cuotas fijadas por el equipo investigador. Este margen de libertad que se concede al entrevistador representa la principal debilidad porque introduce sesgos ya que el entrevistador es libre de entrevistar a quien quiera o pueda. Además dentro de una cuota se puede escoger a unos individuos con preferencia a otros. Por otra parte el entrevistador puede ubicar a los sujetos en cuotas diferentes a las que realmente pertenecen, en aquellas donde se precisen casos.

El principal inconveniente de este tipo de muestreo es que la muestra finalmente obtenida puede no ser representativa de la población que se analiza, aunque la muestra diseñada coincida con la distribución de la población en los controles de cuotas fijados. Para solventar los sesgos inherentes en el muestreo por cuotas, éste suele complementarse con el muestreo de rutas aleatorias: para cada entrevistador se fija un itinerario aleatorio indicándole en qué puntos concretos ha de realizar cada entrevista, limitado con ello la arbitrariedad del entrevistador.

4.6. Muestreo de rutas aleatorias.

Lo solemos encontrar al final de un diseño muestral complementado tanto a muestreos no probabilísticos como a probabilísticos. Se denomina muestreo de rutas porque establece el camino a seguir en la selección de las unidades muestrales. Las rutas se eligen de forma aleatoria, sobre un mapa del municipio en concreto donde se han de realizar las entrevistas. Una vez que se a elegido de forma aleatoria el comienzo de la ruta, el entrevistador deberá tomar una dirección u otra, siguiendo las normas fijadas por el equipo investigador.

Este procedimiento de selección muestral por rutas aleatorias presenta la gran desventaja de no garantizar que todas las unidades de la población tengan la misma probabilidad de ser elegidas, aunque la designación de rutas sea aleatoria. Para obviar dicha ventaja se aconseja complementar con el muestreo por cuotas.

4.7. Muestreo estratégico.

Es una modalidad de muestreo no probabilístico en el que la selección de las unidades muestrales responde a criterios subjetivos, acordes con los objetivos de la investigación.

Esta variedad de muestreo no probabilístico es habitual en estudios cualitativos y también es frecuente en los experimentos realizados con personas que se ofrecen voluntarias en estudios piloto.

4.8. Muestreo de bola de nieve.

Esta variedad difiere de la anterior en que las unidades muestrales van escogiéndose a partir de las referencias aportadas por los sujetos a los que ya se ha accedido. A su vez los nuevos casos identifican a otros individuos en su misma situación y la muestra va aumentando como una bola de nieve.

Es de gran utilidad cuando se carece de un marco de muestreo que recoja a la población de interés, especialmente en poblaciones que son difíciles de identificar y localizar.

Tema 6: LAS TABLAS DE CONTINGENCIA: relación entre variables nominales (ordinales)

Cuando trabajamos con variables nominales u ordinales y queremos ver la relación entre dichas variables, utilizamos las tablas de contingencia, y a partir de ellas calculamos algún estadístico resumen y/o se realiza un contraste de hipótesis.

*** Cruce de variables nominales.**

Ej: cruce de las variables sexo e identificación de partido, con valores absolutos, porcentajes horizontales, verticales sobre el total. Comparamos la identificación de hombres y mujeres (variable independiente sexo) con cada partido (variables dependiente identificación).

Cálculo de porcentajes: para calcular los porcentajes en una tabla de contingencia, siempre que haya una variable independiente haremos 100 sus marginales y compararemos el porcentaje de sus categorías para cada una de las categorías de la variable dependiente.

Diferencia de porcentajes: actúa como medida de la influencia que tuvo el sexo en la suerte corrida por los pasajeros: cuanto mayor sea la diferencia, mayor será la influencia; y a la inversa, una diferencia pequeña indicará que una variable no tiene influencia en la otra.

*** Ficheros de datos agregados como alternativa a las tablas.**

Construir múltiples tablas de contingencia tiene el inconveniente de que dificulta la comprensión de los resultados obtenidos, debido a que aparecen en diversas tablas.

*** Un contraste: la ji−cuadrado para distribuciones uni y bivariables.**

Si queremos generalizar los resultados obtenidos a toda esta población es necesario realizar contrastes (tests o pruebas) de hipótesis y estimaciones. La ji− cuadrado es el contraste típico que se utiliza en las tablas de contingencia (situación bivariable). También sirve para ver si las frecuencias de las categorías de una sola variable son diferentes a una hipotética distribución de frecuencias (situación univariable).

- *La prueba ji−cuadrado para tablas bivariables.*

La aplicación de la ji−cuadrado a situaciones en las que tenemos dos variables y queremos ver si su relación es estadísticamente significativa o, por el contrario, tan sólo cabe atribuirla al azar, es producto de la muestra que hemos elegido, pero no cabe encontrarla en la población de la que se ha extraído.

+ Cálculo de estadístico ji−cuadrado.

Para ver si la relación es estadísticamente significativa se comparan las frecuencias que se observan en cada casilla con aquellas que se habrían obtenido en el supuesto de que las dos variables fueran independientes. Estas frecuencias "esperadas" se obtienen al multiplicar las probabilidades marginales de las dos categorías que definen cada una de las casillas.

Cuanto mayor sea la diferencia entre valores observados y valores esperados (en el supuesto de la independencia de las variables), mayor será la probabilidad de que la muestra provenga de una población en la que las variables estén relacionadas (no sean independientes). El cálculo de la diferencia se hace mediante el estadístico ji−cuadrado.

+ Contraste del estadístico ji−cuadrado.

Este estadístico tienen una distribución muestral derivada de la normal, que recibe el nombre de ji−cuadrao (cálculo de la distribución muestral). Esta distribución muestral no exigen ningún supuesto sobre la distribución de las variables. Sus valores dependen del tamaño de la tabla, expresado en grados de libertad.

- *La prueba ji−cuadrado para distribuciones univariables.*

La ji−cuadrado también se puede utilizar para estudiar si las frecuencias de una sola variable son diferentes entre sí, o para ver si las frecuencias observadas en la distribución de una de nuestras variables se ajusta a una distribución hipotética previamente fijada. En definitiva, se trata de ver si la distribución es uniforme. El contraste ji−cuadrado con una sola variable tiene interés en problemas en los que aparece el tiempo y su influencia.

Con variables a las que se les supone una distribución normal o aproximadamente normal, no tienen sentido el contraste de la ji−cuadrado para ver la uniformidad de la distribución. Lo mismo que tampoco lo tiene en todas las situaciones en las que no quepa pensar que las frecuencias de las diferentes categorías de la variable vayan a ser las mismas.

· Análisis de los residuos:

La prueba ji−cuadrado sirve para ver si la relación entre un par de variables es estadísticamente significativa. El análisis de los residuos va a utilizar las ideas de la ji−cuadrado para estudiar de una manera más pormenorizada la tabla: en lugar de ver si las dos variables están relacionadas estudiamos la relación entre cada pareja de categorías.

El análisis de residuos (diferencia entre valor observado y valor esperado) es una aplicación de la ji−cuadrado al estudio de las parejas de categorías: observamos las frecuencias obtenidas y las comparamos con las esperadas.

Los residuos ajustados (último número de cada casilla) se interpretan como cualquier valor de una variable estandarizada en una distribución normal: valores superiores a +/− 1,96 difieren 0,0 con una probabilidad superior a 0.95. cuanto mayor sea el valor absoluto del residuo ajustado, mayor será la relación entre la pareja de categorías.

- *Estadísticos de resumen para variables nominales.*

Sirven para ver la intensidad de la relación entre variables.

+ Diferencia de porcentajes: es el mejor estadístico para ver la relación entre variables nominales. La diferencia de porcentajes oscila entre d= 100,0 y d= 0,0. El único problema es que para una sola tabla puede que haya que calcular múltiples diferencias.

· Estadísticos basados en la ji−cuadrado.

La ji−cuadrado tiene el inconveniente de que su valor varía directamente con el número de casos. Debido a esta limitación se construyen una serie de estadísticos, basados en la ji−cuadrado, que tienen como fin controlar el tamaño de la muestra.

Cuando las distribuciones marginales de las tablas son asimétricas, aún cuando los porcentajes de dos tablas sean iguales algunas mediadas de asociación darían resultados diferentes según cuál fuera la tabla analizada. Estadísticos: Phi (tablas dos filas por dos columnas); coeficiente de contingencia (1 filas por dos columnas); V de Cramer (1 fila por J columnas).

Junto a su sensibilidad al tamaño de las tablas y a las distribuciones marginales, las medidas basadas en la ji−cuadrado no tienen una interpretación intuitiva. Incluso cuando van de 0,0 a 1,0 es difícil entender un valor 0,19; parece que la relación es débil, pero no hay una lógica estándar para juzgar su magnitud. Estas medidas se desarrollaron como aproximación al coeficiente de correlación de Pearson y han sido complementadas por otras medidas más comprensibles.

Hablar de variables, cuando tenemos información nominal, no tiene mucho sentido; más procedente es hablar de categorías o grupos de individuos: hombres y mujeres (frente al sexo), solteros y casados (frente a estado civil), tal o cual partido, etc.

· Estadísticos basados en la reducción del error de predicción. Tratan de ver la relación entre las variables intentando predecir cómo se clasifica un individuo en una variable Y a partir de que conocemos su clasificación en otra variable, X.

+ Lambda de Goodman y Kruskal. Contesta la pregunta cuánto mejora nuestra capacidad de predecir la clasificación de un individuo en una variable, Y, el hacho de que sepamos cómo se clasifica en otra variable, X. Lambda puede ser asimétrica o simétrica.

· Estadísticos para variables ordinales con pocas categorías.

+ Gamma: permite comparar relaciones diversas de una manera unívoca. Estadístico basado en el orden relativo de las variables. Se calcula tomando parejas de individuos de diferentes casillas de la tabla y preguntándonos si el orden relativo de estos dos individuos en la primera variable es concordante o discordante con su orden en la segunda variable.

· Estadísticos para variables ordinales con muchos valores o categorías diferentes.

Cuando las variables son intervales se utiliza el coeficiente de correlación de Pearson. Si no se puede asumir este nivel de medida, pues se considera que las variables son ordinales, se puede seguir una doble estrategia:

· transformar los valores de los individuos en rangos (orden de cada valor) y utilizar el coeficiente de correlación de Pearson.

· utilizar el coeficiente de correlación de Spearman.

- Cálculo del estadístico de Pearson: con los datos ordenados por rangos y resolviendo el problema de los empates mediante el procedimiento de la media, ambos coeficientes dan el mismo resultado (Pearson y Spearman).
- El tratamiento de los empates: cuando hay empates entre valores, el coeficiente de Pearson tiene el problema de que su valor depende de cómo lo tratemos; y según el tratamiento que se haga de los empates, el coeficiente casi oscila en un punto. Una recomendación general es utilizar, siempre que se

pueda, el coeficiente de correlación de Spearman. Este coeficiente tiene en cuenta el orden de cada valor de las variables y no el mismo valor. De esta manera asume una relación monótona entre las variables.

· análisis de tablas con tres o más variables.

El punto de partida de una investigación puede ser la constatación de que una variable tienen valores diferentes. El uso de estadísticos univariados permite analizar este hecho.

Las tablas con 3 variables, intentan explicar la explicación bivariada. Cuando introducimos una tercera variable intentamos:

· Descubrir si la relación entre dos variables previamente analizadas, es de tipo causal o, por el contrario, se trata de una relación puramente estadística.

· Conocer la secuencia causal entre dos variables, una independiente y otra dependiente, cuando no se duda de su relación.

· Descubrir relaciones ocultas entre otras dos variables.

· Especificar las condiciones en las que se produce la relación entre dos variables.

· Ver el efecto conjunto de dos variables independientes sobre una dependiente.

- *El control por una tercera variable*:

Calculamos la relación entre dos variables y a continuación repetimos el cruce para cada una de las categorías de la tercera variable. Que la relación entre dos variables sea independiente de la influencia de terceras variables significa que cualquiera que sea la tercera variable que introduzcamos como control la relación se mantendrá firme.

− La estandarización: como forma de controlar la influencia de terceras, cuartas, etc variables sobre la relación entre otras dos. En demografía , estandarizar dos poblaciones significa hacerlas iguales, al menos respecto de una característica (variable).

Tema 7: **COMPARACIÓN DE MEDIAS (proporciones)**

Si estamos trabajando con variables intervales y nominales − ordinales y no queremos perder información, podemos utilizar las siguientes técnicas: las *diferencias de medias* y el *análisis de la varianza*, y una extensión de las ideas subyacentes a estas técnicas llamada *análisis de la segmentación.* Estas técnicas se basan en el cálculo de las medias de la variable dependiente para los grupos que forman las variables independientes y se estudian las diferencias que se observan.

- **La comparación de dos medias (proporciones)**

Es la técnica más elemental de todas. Se utiliza cuando queremos estudiar si dos grupos difieren en una característica o un grupo cambia en una característica con el paso del tiempo. Se distingue entre muestras independientes y dependientes o pareadas.

· Modelo general del contraste de dos medias:

- Modelo e hipótesis de los contrastes de las diferencias. Modelo: se sigue manteniendo la necesidad de que

la distribución de la población sea normal, solamente que ahora tenemos dos subpoblaciones, de éstas se obtienen dos submuestras. A menos que las submuestras sean grandes, ambas subpoblaciones han de ser normales.

Las submuestras que saquemos siempre han de ser aleatorias. Según se asuma la independencia o la dependencia de los casos de las submuestras tendremos contrastes diferentes. Los contrastes de las diferencias de medias añaden un supuesto al modelo de la sola media (proporción). Ahora tenemos dos varianzas poblacionales, correspondientes a las dos subpoblaciones, y, cuando las muestras son independientes, tenemos que decidir entre asumir que son iguales o distintas, pues, si bien su media siempre es igual, la desviación típica de esta distribución variará según sea el caso.

El contraste de las diferencias de medias exige que el nivel de medida de la variables dependiente sea interval, puesto que de o contrario no tendría sentido calcular medias.

Hipótesis: en los contrastes de diferencias también vamos a tener hipótesis nula, y, además, varias hipótesis alternativas. La hipótesis nula siempre será que la diferencia de medias en la población es igual a cero. Hipótesis alternativas:

+ las medias son diferentes.

+ la media del grupo uno es mayor que la del grupo dos.

+ la media del grupo uno es menor que la del grupo 2.

- Distribución muestral:

El mismo teorema que servía para una sola muestra sigue siendo válido para la situación en la que tenemos infinitas parejas de muestras, en cada una de las cuales se calcula un estadístico diferente.

+ Muestras independientes: podríamos demostrar que la distribución de las r diferencias es normal o se aproxima mediante la t de Student, cuando las submuestras son pequeñas, con media o valor esperado de las diferencias igual a la diferencia en las subpoblaciones.

La desviación típica de las diferencias de las medias variará según se asuma que las varianzas de las subpoblaciones sean: distintas (la desviación típica es igual a la suma de las desviaciones típicas de cada uno de los términos de la diferencia) o iguales (se calcula una desviación típica media).

+ Muestras dependientes: la distribución de las diferencias es normal, siendo la media de todas las diferencias o valor esperado de las diferencias igual a la diferencia en las subpoblaciones. El estimador de la desviación típica de esta distribución muestral también tiene un valor conocido y único.

- Valor − P, nivel de significación y región crítica.

Cuando tengamos una única hipotesis alternativa, después de calcular el estadístico de nuestra pareja de submuestras podemos ver la probabilidad de hablerlo obtenido suponiendo que el modelo del contraste fuera cierto.

- La decisión de rechazar o no la hipótesis nula depende de nuestro nivel de exigencia, caso de que trabajemos calculando un valor − P, o del nivel de significación que hayamos fijado en el contraste, en un enfoque clásico.

La ventaja del primer enfoque es que si rechazamos la hipótesis nula lo hacemos conociendo la probabilidad

real que tenemos de cometer un error. El enfoque clásico es que si la probabilidad de obtener al azar nuestro estadístico es 0,049, rechazamos la hipótesis nula; si la probabilidad es 0,051, no la rechazamos. Proceder de esta segunda manera parece demasiado rígido.

- **Muestras independientes:**

+ Contraste no paramétrico para muestras independientes: cuando no estemos en condiciones de garantizar ni la normalidad de la distribución ni la igualdad de las varianzas, siempre es posible recurrir a un contraste no paramétrico. En el caso de las muestras independientes el contraste adecuado es el de Mann − Whiney, o de Wilcoxon: tan solo exige que las observaciones sean una muestra aleatoria, ordenadas de menor a mayor, sin necesidad de que tengan un nivel de medida interval.

Este test plantea como hipótesis nula que los dos grupos provienen de la misma distribución y que, por tanto, las diferencias de medias que se observan entre uno y otro son atribuibles al azar. Utilizando este contraste es más difícil rechazar la hipótesis nula que con el contraste de la t.

- **Muestras dependientes:**

También llamadas pareadas, puesto que están constituídas por parejas de observaciones, normalmente correspondientes al mismo individuo. Dependiendo de que tratemos las muestras como independientes o dependientes, haremos análisis diferentes:

· Independientes: calcularemos las medias de las opiniones sobre las situaciones actual y futura y haremos un contraste para ver si su diferencia es significativa.

·Dependientes: veremos las diferencias de cada pareja de opiniones, calculando posteriormente una diferencia media. En este caso el contraste tienen como fin ver si la diferencia media es distinta de cero. Supone calcular primero las diferencias entre los valores de cada individuo, para estudiar después si la diferencia media es significativamente diferente de cero.

· Supuestos del contraste: este contraste plantea la necesidad de que la distribución de las diferencias sea aproximadamente normal.

Contraste no paramétrico para muestras pareadas. El contraste de la t que utilizamos para estudiar la diferencia de las medias de dos muestras pareadas exige que la distribución de las diferencias entre ambas variables esté normalmente distribuida, o que el tamaño de las muestras de las diferencias sea grande, con el fin de aplicar el teorema central del límite.

El test del signo es una prueba no paramétrica que se utiliza con muestras pareadas para contrastar la hipótesis de que las distribuciones de dos variables son iguales. No exige ningún supuesto sobre la forma de la distribución. La idea del test es que si ambas variables tuvieran la misma distribución, coincidiría el número de diferencias positivas y negativas. Cuanto mayor sea la diferencia entre diferencias positivas y negativas, mayor es la probabilidad de que las distribuciones de ambas variables sean diferentes.

- **Comparación de proporciones:**

(Casi) todo lo que se dice sobre las medias se puede aplicar a las proporciones.

· *Diferencia de porcentajes con una sola variable*.

· *Diferencia de porcentajes con dos variables*: podemos:

·Ver el cruce de ambas variables, mediante una tabla de contingencia, y realizar un contraste de la ji−cuadrado. Esto es útil en tablas de 2x2.

· Hacer igualmente el cruce para ver la diferencia de proporciones y luego realizar un contraste de la diferencia de proporciones.

- **Contraste de la diferencia de proporciones:**

- Modelo e hipótesis del contraste: modelo: dos submuestras aleatorias e independientes. Sólo se adopta el contraste en el que se asumen varianzas iguales. Puesto que la hipótesis nula es que las proporciones poblacionales son iguales, y la varianza de una proporción está basada en esa misma proporción, sería contradictorio planear esta hipótesis con un modelo que postulase la diferencia de varianzas. Hipótesis: la hipótesis nula plantea la igualdad de proporciones en las dos subpoblaciones, mientras que la hipótesis alternativa muestra su diferencia.
- Distribución muestral de la diferencia de proporciones.
- Valor − P, nivel de significación y región crítica: conocida la media y la desviación típica de la distribución muestral, podemos tipificar la diferencia obtenida.
- Toma de decisión.

- **Segmentación de la muestra**

Es una técnica muy útil que no exige mayores conocimientos estadísticos. Es segmentar una variable en subgrupos, para cada uno de los cuales se calcula la media.

+ Relaciones condicionales (interacción): cuando se observa que las medias de las categorías de una variable difieren con el nivel de primera a tercera se dice que existe interacción entre las tres variables. También se dice que la influencia es de tipo condicional, pues las medias de las categorías de una variable cambian según sean sus condiciones. Cuando tratamos los datos como una muestra de la población, hay que realizar contrastes o pruebas que nos permitan ver si las diferencias de medias que se observan entre las categorías son estadísticamente significativas. Tenemos que introducir una nueva prueba, el análisis de la varianza.

- **El análisis de la varianza:**

Es una extensión de las diferencias de medias a situaciones en las que existen más de dos grupos.

+ Análisis de variazna con un factor (oneway). Cuando utilizamos el análisis de la varianza queremos ver el efecto que tienen una o varias variables independientes en otra dependiente. A las variables independientes (nominal u ordinal) se les llama factores, y a sus categorías niveles. Etapas:

- Vemos las medias de valoración para cada grupo.
- Comprobamos si se cumplen los supuestos que justifican la utilización del análisis de la varianza con un solo factor.
- Calculamos un estadístico que resuma la relación entre ambas variables: la F de Snedecor. Si los datos provienen de una muestra aleatoria, contrastamos este estadístico para ver si es estadísticamente significativo.
- Suponiendo que las diferencias sean significativas hemos de comprobar entre qué parejas.

· Estadísticos descriptivos univariables.

· Comprobación de los supuestos y prueba no paramétrica de Kruskal − Wallis.

Tendremos que realizar un contraste de hipótesis que nos permita ver la significatividad estadística de las

diferencias observadas en las tres muestras. El contraste que elegimos es la F. Se supone (modelo del contraste):

· que las submuestras de cada uno de los r niveles de los factores son aleatorias e independientes.

· que sus distribuciones son normales y de igual varianza − supuestos de normalidad y homocedasticidad−.

Como hipótesis nula diremos que las medias poblacionales de las r submuestras son iguales. La hipótesis alternativa postulará su diferencia. El problema se plantea cuando no se cumplen los supuesto de normalidad y homocedasticidad, o la variable criterio no es interval.

Soluciones:

- transformar los datos, tratando de conseguir distribuciones de igual varianza, lo cual suele "normalizar" las variables.
- Utilizar un contraste no paramétrico que no exija ninguno de estos supuestos. En particular podemos utilizar la prueba de Kruskal − Wallis. Esta prueba es una extensión del test de mann − Whitney. El test utiliza sus rangos − por tanto el contraste permite que el nivel de medida de la variable dependiente sea ordinal. Se parte de una ordenación de todos los casos por orden de rango, para ver a continuación el sumatorio de estos rangos para cada uno de los grupos.

    - **Contraste de las medias: idea del contraste:**

- Descomposición de la varianza: esta varianza se puede descomponer en varias partes: una varianza entre las medias de los diferentes grupos − niveles del factor; varianza dentro de cada grupo o nivel del factor. La primera se llama varianza entre grupos o varianza explicada, puesto que es la parte de la varianza de la variable que es atribuible al hecho de que los individuos entrevistados sean de diferentes ideologías. La segunda se llama intragrupos o residual o no explicada, puesto que es la parte de la varianza que no sabemos a qué atribuir.
- Estimación de la varianza: el análisis de la varianza calcula un estadístico, la F, que compara las varianza entre e intragrupos. Cuando las tres medias provengan de una misma distribución, las varianzas entre e intragrupos serán aproximadamente igual y su razón se aproximará a la unidad.

    - **Contrastes y comparaciones múltiples entre medias.**

Una vez que hemos comprobado que existe diferencia entre las medias, tratamos de ver entre qué medias en particular. Es decir, la F del apartado anterior nos dice que las valoraciones medias de los grupos son diferentes. La prueba de Scheffe sirve para hacer comparaciones binarias. Tiene la ventaja de ser aplicable en muestras de tamaño desigual y es bastante robusto frente a desviaciones del supuesto de homocedasticidad.

+ Análisis de la varianza con dos factores (ANOVA): interesa estudiar el efecto de ambos factores, aisladamente y en interacción. Nuevos conceptos:

· Diseños ortogonales: aquel en el que las variables independientes están correlacionadas. El número de casos en cada una de las combinaciones de las categorías de los factores ha de ser el mismo (diseño equilibrado). Se obtienen fácilmente en la investigación experimental. En la no experimental es difícil que se consiga la ortogonalidad de los factores, puesto que las variables independientes suelen estar correlacionadas, además de resultar casi imposible que aparezca el mismo número de casos en cada combinación de sus categorías. La condición de equilibrio se puede obviar siempre y cuando se mantenga la proporcionalidad en las categorías. En estos casos de proporcionalidad es posible utilizar los procedimientos tradicionales del análisis de la varianza, con tal de que se cumplan los supuestos de normalidad y homocedasticidad. Pasos:

- Cálculo de los estadísticos descriptivos básicos.
- Supuestos del análisis de la varianza.
- Contrastes de los efectos de cada uno de los factores y de su interacción.
- Intensidad de la asociación entre los factores y la variable dependiente.
- Si no hay interacción se ofrece un análisis de clasificación múltiple.

· Estadísticos descriptivos básicos: diferencia de medias y relación entre variables.

· Modelo e hipótesis del análisis de la varianza con dos factores: supuestos: las muestras de los grupos tienen que ser aleatorias e independientes, sus distribuciones han de ser normales y de igual varianza. Esto referido a las casillas formadas por las combinaciones de los grupos.

A los supuestos añadimos la condición de que el diseño sea ortogonal (independencia entre los factores) y equilibrado (igual número de casos en cada combinación de los niveles de los factores). Si los factores están correlacionados, parte de la variación explicada por un factor también será explicada por el otro, con lo cual habrá ambigüedad a la hora de decidir qué factor es el responsable de la varianza común explicada.

· Contraste del efecto de cada uno de los factores, por separado, y test de la interacción.

Descomposición de la variación: descomponer la suma de cuadrados total en sus partes constitutivas*:*

- *Variabilidad* factor 1: atribuible a que no todos los individuos son iguales en el primer factor (Factor A):
- *Variabilidad factor 2*: no todos los individuos son iguales en el factor 2 (factor B).
- *Variabilidad de factores*: atribuible al efecto conjunto, diferencial, de los dos factores sobre la variable dependiente. Esta variabilidad se mide viendo las diferencias, al cuadrado, entre las medias de cada combinación de categorías y la media total.
- *Variabilidad residual*: no atribuible a ninguna de las tres causas anteriores. Recibe el nombre de variación (suma de cuadrados) residual o no explicada. Es el error aleatorio en la variable dependiente. Se mide viendo las diferencias , al cuadrado, entre cada observación y la media de la combinación de categorías a las que pertenece.
- *Variabilidad total*: suma de todas las variabilidades anteriores. Mide las diferencias, al cuadrado, de cada individuo con relación a la media, y recibe el nombre de suma de cuadrados total.

· Análisis de la clasificación múltiple. Permite contemplar la información obtenida con los contrastes. Un contraste puede indicar que el efecto de un factor es estadísticamente significativo, sin que por ello sepamos la intensidad de su influencia. Con muestras suficientemente grandes, casi todos los estadísticos que contrastemos serán estadísticamente significativos. Podemos estar interesados en ver la intensidad del efecto de los factores sobre la variable independiente, independientemente de que estos efectos sean estadísticamente significativos.

- **Detector automático de la interacción (AID).**

Realiza un análisis semejante al tratar de la segmentación, sólo que el proceso de subdividir la muestra en subgrupos se realiza automáticamente, siguiendo el criterio de seleccionar las variables independientes de tal manera que maximicen nuestra capacidad para predecir los valores de la variable dependiente.

Dada una serie de variables independientes (predictoras) y otra dependiente, la técnica del SID funciona sobre la base de dicotomizar las variables, para buscar entre todas las variables predictoras aquella que explica mayor varianza dependiente.