

**INTRODUCCION A LA  
INFERENCIA ESTADISTICA:  
MUESTREO Y ESTIMACION  
PUNTUAL Y POR INTERVALOS**

José Luis Vicente Villardón  
Dpto. de Estadística  
Universidad de Salamanca

<b>1.- INTRODUCCION Y MOTIVACION.....</b>	<b>3</b>
<b>2.- INFERENCIA Y MUESTRAS.....</b>	<b>8</b>
<b>3.- MUESTREO .....</b>	<b>13</b>
MUESTREO ALEATORIO SIMPLE (MAS).....	13
MUESTREO SISTEMATICO.....	14
MUESTREO POR CONGLOMERADOS.....	14
MUESTREO ESTRATIFICADO.....	15
<b>4.- ESTADISTICOS Y DISTRIBUCIONES MUESTRALES.....</b>	<b>16</b>
<b>5.- DISTRIBUCIONES MUESTRALES DE LA MEDIA Y LA DESVIACION TIPICA. .....</b>	<b>18</b>
<b>6.- EL TEOREMA CENTRAL DEL LIMITE.....</b>	<b>20</b>
<b>7.- ESTIMADORES Y PROPIEDADES DESEABLES DE LOS ESTIMADORES....</b>	<b>20</b>
<b>8.-METODOS DE ESTIMACION.....</b>	<b>22</b>
<b>9.-ESTIMADORES PUNTUALES DE LOS PARAMETROS DE UNA POBLACION NORMAL.....</b>	<b>25</b>
<b>10.- ESTIMADORES DE LOS PARAMETROS DE LAS DISTRIBUCIONES DISCRETAS MAS USUALES.....</b>	<b>27</b>
<b>11.- EJEMPLO.....</b>	<b>28</b>
<b>12.- ESTIMACION POR INTERVALOS.....</b>	<b>30</b>
INTRODUCCION.....	30
INTERVALO DE CONFIANZA PARA LA MEDIA DE UNA POBLACION NORMAL DE VARIANZA CONOCIDA.....	30
LONGITUD DEL INTERVALO Y ERROR EN LA ESTIMACIÓN.....	32
CALCULO DEL TAMAÑO MUESTRAL PARA ESTIMAR LA MEDIA DE UNA POBLACION CON UNA DETERMINADA PRECISION.....	33
INTERVALO DE CONFIANZA PARA LA MEDIA DE UNA POBLACION NORMAL CON VARIANZA DESCONOCIDA.....	34

## 1.- INTRODUCCION Y MOTIVACION

La principal razón de que el Método Estadístico se haya desarrollado ampliamente en los últimos años dentro de las Ciencias Experimentales es que éstas están sujetas a razonamientos de tipo inductivo que van de lo particular a lo general. Sacaremos conclusiones sobre un grupo de individuos a partir de la información que nos proporciona un subconjunto más o menos amplio de los mismos. De acuerdo con MARTIN ANDRES y LUNA CASTILLO (1990), “El único método científico para validar tales extensiones es el Método Estadístico, pues precisamente esa es la causa de su existencia”.

La expansión del Método Estadístico es tal que, de todas las disciplinas que nuestros alumnos han de estudiar a lo largo de toda la enseñanza secundaria, la Estadística es prácticamente la única que tendrán como asignatura en la mayor parte de las carreras universitarias que puedan elegir en el futuro; desde las típicamente consideradas experimentales, como la Medicina o la Biología, hasta carreras consideradas como de letras como la Psicología, la Sociología o incluso la Geografía. Aquellos que decidan no tomar el camino de la Universidad se encontrarán cada vez más frecuentemente con conceptos procedentes de la Ciencia Estadística como por ejemplo el de error máximo admisible o el de nivel de confianza en cualquier encuesta sociológica de las que habitualmente aparecen en la prensa.

El primer concepto importante que hemos de transmitir a nuestros alumnos es la diferencia existente entre lo que son las estadísticas como meras colecciones de datos y lo que es el Método Estadístico considerado como una disciplina científica con entidad propia.

Es común escuchar la frase “No creo en las estadísticas”, incluso entre profesionales cercanos a la disciplina. Efectivamente las “estadísticas” como posible ayuda a la toma de decisiones dependen de quién y como se hayan tomado los datos y de si las respuestas que dan los encuestados se ajustan a su opinión real. En este sentido los datos pueden ser susceptibles de creencia puesto que uno puede dudar de la intención del encuestado. El Método Estadístico, tal y como está concebido en la actualidad, forma parte del saber científico y es aceptado lo mismo que lo es, por ejemplo, la Teoría de la Relatividad en Física; no es, por tanto, terreno de las creencias y seguirá siendo aceptado como válido hasta que alguien proponga una nueva teoría que lo modifique.

Recapitulando sobre lo expuesto, la Estadística se configura como la tecnología del método científico que proporciona instrumentos para la toma de decisiones cuando estas se

adoptan en ambientes de incertidumbre, siempre que esta incertidumbre pueda ser cuantificada en términos de probabilidad. (MARTIN PLIEGO, 1994).

El procedimiento de toma de decisiones, o de aprendizaje, en el ámbito científico se resume en la figura 1, y consiste básicamente en plantear una hipótesis, contrastarla mediante datos experimentales y modificarla si no puede ser aceptada. Es precisamente en el paso de contraste en el que el Método Estadístico juega un papel fundamental y aunque cualquier científico puede realizar una investigación sin estadística, sin embargo es mucho más fiable si el resultado está basado en métodos estadísticos. No se concibe la investigación aplicada actual sin la utilización de la Estadística en el proceso de inducción.

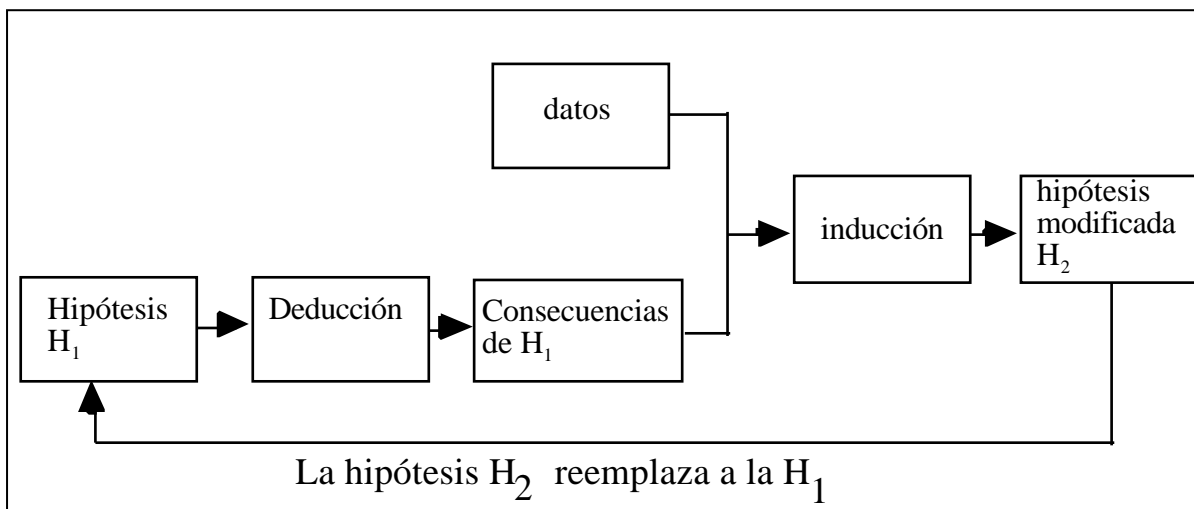


Figura 1: El proceso de aprendizaje.

El cuadro 1 muestra los pasos fundamentales del método científico en relación con el método estadístico.

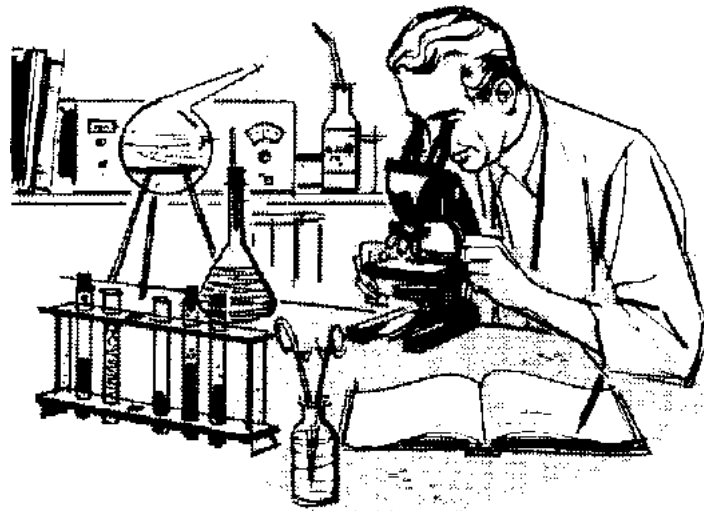


Figura 2: El Método Estadístico es una parte importante de la investigación científica actual.

## MÉTODO CIENTÍFICO

1.- PLANTEAR UNA IDEA (HIPOTESIS)

2.- CONTRASTAR LA IDEA

a) Establecer la población o poblaciones a estudiar.

*b) Decidir el método para la recolección de los datos.*

*c) Suponer un modelo, especificando las distribuciones de las poblaciones en estudio.*

*d) Formular las hipótesis de interés en términos de los parámetros del modelo.*

*e) Calcular el tamaño muestral necesario para conseguir los objetivos tan eficientemente como sea posible. El cálculo requiere el conocimiento de la mínima diferencia en la que el investigador está interesado, así como un estimador de la variabilidad subyacente.*

f) Recoger los datos.

*g) Revisar si el modelo supuesto puede considerarse una aproximación razonable.*

*h) Revisión del análisis si las suposiciones de partida del modelo no son ciertas.*

*i) Analizar los datos.*

j) Escribir las conclusiones en lenguaje simple (no estadístico).

3.- REVISAR LA IDEA SI NO SE ACEPTA A PARTIR DEL PROCEDIMIENTO EXPERIMENTAL.

Cuadro 1: El método científico y su relación con la Estadística. Se han señalado en cursiva los pasos del método directamente relacionados con la Estadística, que van desde la recogida de los datos hasta el análisis de los mismos.

Estudiaremos cada uno de los apartados mencionados aunque no necesariamente en el orden en el que aparecen en el cuadro anterior.

Se plantea ahora un problema que suscita polémica entre los profesionales de las Estadística, el enfoque que debe darse a la explicación de los conceptos fundamentales.

Trataremos de exponer nuestro punto de vista al respecto antes de comenzar con la explicación propiamente dicha. Dos son los enfoques predominantes, si bien pueden considerarse posturas intermedias; el primer bloque estaría formado por aquellos que consideran la Estadística como una especialidad más de las Matemáticas sin características diferenciales claras con respecto al resto de las disciplinas; el segundo bloque estaría formado por aquellos que piensan que la Estadística tiene entidad propia como disciplina científica en la que las Matemáticas han de entenderse simplemente como una herramienta.

Como profesionales de la Estadística Aplicada, nos inclinamos por la segunda de las posibilidades si bien no se debe olvidar el fondo teórico de la disciplina y las herramientas matemáticas básicas, que se entenderán como un medio y no como un fin en sí mismas. Trataremos de explicar esta postura más ampliamente en los párrafos que siguen.

La Estadística como disciplina tiene fundamentalmente un carácter inductivo en contraposición al carácter deductivo de las Matemáticas, el objeto último de la misma es sacar conclusiones sobre una población a partir de la información que proporciona una muestra de la misma, y no el desarrollo de los teoremas propiamente dichos que sería objeto de la denominada Estadística Matemática. Un ejemplo similar sería el de la Física, con un campo propio, y el de los métodos matemáticos aplicados a la Física que forman parte de las Matemáticas.

El objeto de la Estadística Aplicada son los Métodos Estadísticos, los resultados y su aplicación en otras disciplinas científicas; la obtención teórica de dichos métodos utiliza herramientas matemáticas (Cálculo, Álgebra o Geometría) o conceptos de Cálculo de Probabilidades. Siguiendo a WOLFOWITZ (1969)<sup>1</sup>:

*Excepto quizás unos pocos de los más profundos teoremas, y quizás ni siquiera esos, la mayor parte de los teoremas de la Estadística no sobrevivirían en las Matemáticas si el sujeto de la propia estadística (la aplicación) desapareciera. Para sobrevivir al sujeto deben responder más a las necesidades de aplicación.*

*De lo que debemos protegernos es del desarrollo de una teoría que, por una parte, tiene poca o ninguna relación con los problemas reales de la Estadística, y que, por otra parte, cuando se ve como Matemática pura, no es lo suficientemente interesante, por sí misma, ni para sobrevivir.*

También en este sentido TUKEY (1962)<sup>2</sup>, que podría ser considerado como el padre de la aproximación exploratoria del análisis de datos, apunta lo siguiente:

*La máxima más importante a la que el análisis de datos debe prestar atención, y una de las que muchos estadísticos parecen haber olvidado, es ésta: “Mucho mejor una respuesta aproximada a una pregunta correcta, que es a menudo vaga, que una respuesta exacta a la pregunta errónea, que puede hacerse siempre de forma precisa.” El análisis de datos debe progresar aproximando respuestas, en el mejor de los casos, ya que su conocimiento de lo que es realmente el problema será en el mejor de los casos aproximado.*

Todo lo dicho pone de manifiesto que hay distintas formas de entender las cosas probablemente debido a la conjunción de la parte inductiva en la esencia de la disciplina y la parte deductiva en su desarrollo. Es la parte deductiva (matemáticas) la que ha situado a la Estadística, hasta hace pocos años, como una especialidad de la licenciatura de Matemáticas, y es probablemente la parte inductiva la que ha hecho que en esas mismas facultades fuera considerada como la hermana pobre, o cuando menos, como algo extraño y diferente, por los matemáticos tradicionales.

El proceso futuro que seguirá la Estadística como disciplina científica pasará, sin duda, por la separación de las Matemáticas, como lo hizo en su momento la Física, que tiene su propia entidad aunque utilice el método matemático como herramienta. De hecho, ya es posible cursar estudios de Estadística (tanto de primer como de segundo ciclo) en Facultades de Estadística separadas de las de Matemáticas. (Aunque desgraciadamente en la mayoría de los casos siguen controlados por los matemáticos).

Es esta misma disyuntiva es la que ha colocado los conceptos de Estadística necesarios en las Enseñanzas Medias dentro de la asignatura de Matemáticas, y la que ha hecho que muchos de los profesores, con formación matemática tradicional, prefieran relegarla a un segundo plano cuando, en realidad, es la única parte del programa que prácticamente todos los que tomen el camino universitario van a estudiar.

En Facultades Aplicadas (Medicina, Biología, Economía, Psicología, Geografía, Derecho, Biblioteconomía, Traducción y documentación, etc ... ) enseñamos Estadística Aplicada, es decir, los resultados más relevantes que permiten al alumno resolver problemas que se

---

<sup>1</sup> -WOLFOWITZ, J. (1969): 'Reflections on the future of mathematical statistics'. en R. c. Bose *et al.* (eds.) "Essays in Probability and Statistics". University of North Carolina Press. Chapel Hill.

<sup>2</sup> -TUKEY, J.W. (1962): 'The future of Data Analysis'. *Annals of Mathematical Statistics*, 33, 1-67.

encontrará en su ejercicio profesional, aprendiendo el lenguaje y las técnicas básicas que le permitan comprender no sólo las situaciones que se le plantean en el curso sino también posibles situaciones futuras.

No es necesario enseñar la parte deductiva completamente, ya que se trata de usuarios de los métodos, y no es preciso profundizar en aspectos meramente técnicos que pertenecen exclusivamente al mundo de las Matemáticas. De alguna manera, el rigor conceptual para transmitir la filosofía básica de trabajo dentro del método científico, sustituye al rigor matemático en la presentación de resultados ya que los alumnos han de resolver problemas de investigación en su propia rama y no en Matemáticas..

En Facultades de Matemáticas y Estadística el enfoque estará más dirigido al aspecto técnico-matemático, especialmente en las primeras. En las nuevas facultades de Estadística tendrán que aprender que el objeto es la aplicación y que los resultados matemáticos necesarios para el desarrollo deductivo de los "Métodos Estadísticos" son sólo una herramienta y no el objeto en si mismos.

La mayor parte de nuestros alumnos cursará estudios en Facultades Aplicadas por lo que trataremos de centrar nuestra atención en el "Método Estadístico" y no en su deducción técnica, si bien puede realizarse algún ejercicio para aplicar, en este contexto, los conceptos aprendidos en el resto de la asignatura de Matemáticas. Es posible, también utilizar ejercicios en conexión con los profesores de otras asignaturas como Biología, Geografía Económica, etc.

## 2.- INFERENCIA Y MUESTRAS

*La Inferencia Estadística es aquella rama de la Estadística mediante la cual se trata de sacar conclusiones de una población en estudio, a partir de la información que proporciona una muestra representativa de la misma.* También es denominada Estadística Inductiva o Inferencia Inductiva ya que es un procedimiento para generar nuevo conocimiento científico.

*La muestra se obtiene por observación o experimentación.* La necesidad de obtener un subconjunto reducido de la población es obvia si tenemos en cuenta los costes económicos de la experimentación o el hecho de que muchos de los métodos de medida son destructivos.

Toda inferencia inductiva exacta es imposible ya que disponemos de información parcial,

sin embargo es posible realizar inferencias inseguras y medir el grado de inseguridad si el experimento se ha realizado de acuerdo con determinados principios. Uno de los propósitos de la inferencia Estadística es el de conseguir técnicas para hacer inferencias inductivas y medir el grado de incertidumbre de tales inferencias. La medida de la incertidumbre se realiza en términos de probabilidad.

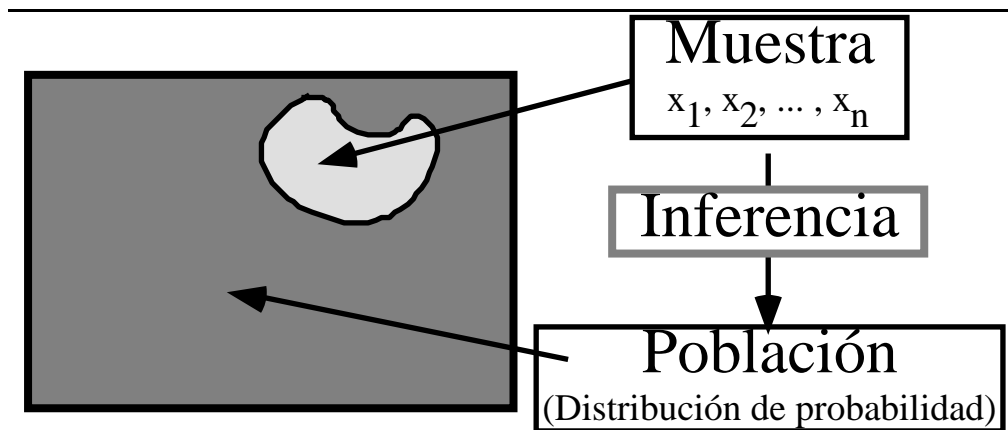


Figura 3: Esquema de Inferencia Estadística.

El primer concepto importante es el de *población*, que es el conjunto de individuos sobre los que se desea información. La población ha de estar perfectamente definida a la hora de comenzar el estudio. (paso 2-a de la descripción del método científico en el Cuadro 1). Por ejemplo, en un ensayo clínico en el que se pretende demostrar la efectividad de un tratamiento han de estar muy claros cuales son los criterios de inclusión de un paciente en la población (muestra) a estudiar.

De la población se extrae un subconjunto que se denomina muestra. La muestra ha de ser representativa de la población, en el sentido de que debe tener una *composición similar en cuanto a la proporción de distintas características*. Por ejemplo, una muestra para un estudio de estaturas no incluirá solamente individuos bajos o altos, sino individuos de ambas clases en proporciones similares a las de la población. La representatividad de la muestra queda garantizada con la elección correcta del método de muestreo, que se estudiarán en el punto siguiente.

Sobre cada uno de los individuos medimos una o varias características que denominamos variables. Así a cada población le corresponde una variable aleatoria que denotaremos con  $X$ . En la teoría de la Estadística quedan identificadas Población y variable aleatoria asociada. Así en toda la teoría de la Inferencia población significará el conjunto de individuos a estudiar, pero también la variable aleatoria asociada a la característica que medimos sobre los individuos.

En general, trataremos con poblaciones infinitas, entendiendo que en la práctica "población infinita" significa lo mismo que "población muy grande" ya que conceptualmente la mayor parte de las poblaciones no pueden ser consideradas infinitas.

En general, supondremos un modelo de distribución de probabilidad para la variable aleatoria en estudio que resuma las características de la misma (apartado 2c del método científico en el Cuadro 1), aunque desconocemos los parámetros que trataremos de estimar a partir de una muestra. Por ejemplo suponemos que  $X$  es  $N(\mu, \sigma^2)$  donde los dos parámetros, o uno de ellos, son desconocidos. En algunos casos no es necesario especificar tales distribuciones y las inferencias se hacen sobre características de la distribución que no son necesariamente parámetros.

La inferencia Estadística puede dividirse en dos apartados **de acuerdo con el conocimiento sobre la distribución en la población.**

**Inferencia Paramétrica:**

Se conoce la forma de la distribución (Normal, Binomial, Poisson, etc .... ) pero se desconocen sus parámetros. Se realizan inferencias sobre los parámetros desconocidos de la distribución conocida.

**Inferencia No Paramétrica:**

Forma y parámetros desconocidos. Se realizan inferencias sobre características que no tienen porque ser parámetros de una distribución conocida (Mediana, Estadísticos de Orden).

**De acuerdo con la forma en que se estudian los parámetros** o características desconocidas, la inferencia puede dividirse en dos apartados:

**Estimación:**

Se intenta dar estimaciones de los parámetros desconocidos sin hacer hipótesis previas sobre posibles valores de los mismos.

**Estimación puntual:** Un único valor para cada parámetro.

**Estimación por intervalos:** Intervalo de valores probables para el parámetro.

**Contraste de Hipótesis:**

Se realizan hipótesis sobre los parámetros desconocidos y se desarrolla un procedimiento para comprobar la verosimilitud de la hipótesis planteada.

Veamos los conceptos con un ejemplo concreto tomado de un estudio de investigación real. El estudio pertenece a otro más amplio llevado a cabo en colaboración por los Departamentos de Química Analítica, Nutrición y Bromatología , y Estadística y Matemática Aplicada.

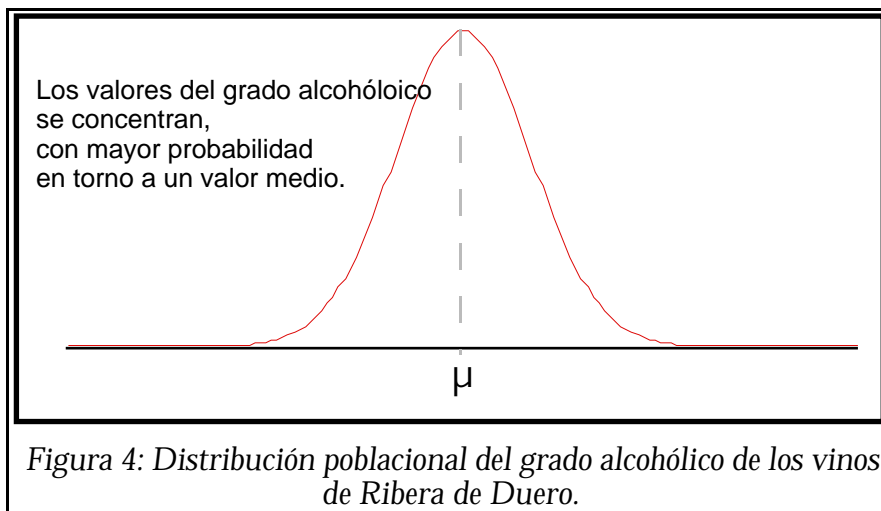
*El objetivo original del trabajo consiste en estudiar los vinos jóvenes embotellados de dos*

denominaciones de origen, Ribera de Duero y Toro, mediante técnicas de laboratorio objetivas, con el fin de buscar las características que los diferencian y evitar los posibles fraudes producidos por el intercambio debido a la proximidad geográfica de ambas denominaciones. Por el momento nos centraremos en una sola variable, el grado alcohólico, y en una sola de las poblaciones, la de Ribera de Duero. Fijaremos además un momento del tiempo, la cosecha del año 1986.

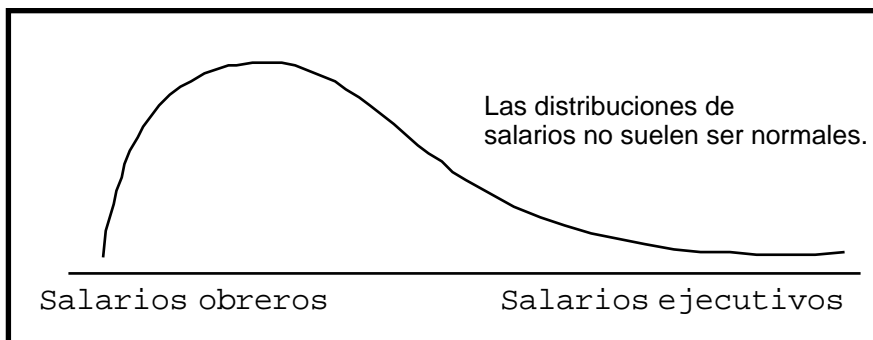
El primer paso de cualquier investigación, la definición clara de la población en estudio, se obtiene de los propios objetivos del mismo. **Estudiaremos vinos jóvenes embotellados de la denominación de origen "Ribera de Duero" en la cosecha de 1986. La variable a medir es el grado alcohólico.**

Seguramente todos hemos observado que en las botellas de vino aparece el grado alcohólico de las mismas, que suele ser entre 12 y 12,5 grados. Es obvio que este valor no es el contenido exacto de cada una de las botellas, sino que se trata de un contenido medio. Supongamos que desconocemos ese contenido medio para la población y deseamos averiguarlo, para lo cual hemos de seleccionar una muestra de la población. La necesidad de seleccionar una muestra es clara ya que el análisis del contenido alcohólico implica la destrucción del individuo, la botella de vino.

Aunque la población no puede ser infinita supondremos que lo es ya que el número de botellas es muy grande y supondremos que la variable aleatoria sigue una distribución normal. La hipótesis sobre la distribución de probabilidad ha de hacerse a priori, teniendo en cuenta las características conocidas de la población en estudio (hay que tener en cuenta que se trata solamente de un modelo para ajustar la realidad.) El ejemplo parece lógico utilizar una distribución normal ya que es posible suponer que los posibles valores del grado alcohólico se concentran de forma simétrica en torno a un valor medio, y que la probabilidad de encontrar valores decrece a medida que aumenta la distancia a dicho valor medio. (Figura 4).



Si tuviéramos, por ejemplo, la distribución de los salarios de los empleados de una Empresa dedicada a la fabricación de automóviles, en principio no podemos suponer la distribución normal ya la distribución es probablemente asimétrica con una cola hacia los salarios altos determinada por los salarios de los ejecutivos.



En la mayor parte de las investigaciones reales suponemos que las variables o transformaciones de las mismas (logaritmos, etc, ...) tienen distribuciones aproximadamente normales.

El paso siguiente consiste en determinar posibles valores para los parámetros desconocidos, para lo cual hemos de obtener una muestra representativa de la población. La obtención de una muestra representativa se trata en el punto siguiente.

### 3.- MUESTREO

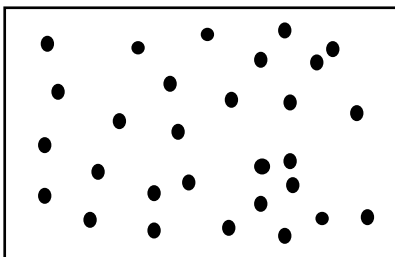
Aunque la teoría que será desarrollada más tarde está referida solamente a muestras aleatorias simples, realizaremos aquí una rápida revisión de posibles métodos para la toma de muestras que podemos encontrarnos en la práctica.

Los pasos a seguir para la recolección de una muestra son los siguientes:

- **Definir la población** en estudio especificando las unidades que la componen, el área geográfica donde se realiza el estudio (si procede) y el periodo de tiempo en el que se realizará el mismo.
- **Definir el marco:** listado o descripción de los elementos que forman la población.
- **Definir la unidad de muestreo:** Ciudades, calles, hogares, individuos, etc ...
- **Definir las variables** a medir o las preguntas que se harán si se trata de una encuesta.
- **Seleccionar el método de muestreo:** Probabilístico o No Probabilístico, aunque son los primeros los que nos permiten la estimación correcta de parámetros.
- **Calcular el tamaño necesario** para obtener una determinada precisión en la estimación. Este punto se verá con más detalle en el apartado dedicado a la estimación por intervalos.
- **Elaborar el plan de muestreo** que guiará el trabajo de campo.

En cuanto al tipo de muestreo, algunas de las características más importantes de los muestreos probabilísticos más usuales se detallan a continuación:

#### ***MUESTREO ALEATORIO SIMPLE (MAS)***



Se trata de un procedimiento de muestreo (sin reemplazamiento), en el que se seleccionan  $n$  unidades de las  $N$  en la población, de forma que cualquier posible muestra del mismo tamaño tiene la misma probabilidad de ser elegidas.

Se realizan  $n$  selecciones independientes de forma que en cada selección los individuos que no han sido elegidos tengan la misma probabilidad de serlo.

El procedimiento habitual consiste en numerar todos los elementos de la población y se seleccionan muestras del tamaño deseado utilizando una tabla de números aleatorios o un programa de ordenador que proporcione números aleatorios.

Recuérdese que "al azar" no significa "de cualquier manera", para que el procedimiento de muestreo sea válido es necesario utilizar correctamente el proceso de generación de números aleatorios.

Entre las ventajas de este procedimiento esta la compensación de valores altos y bajos con lo que la muestra tiene una composición similar a la de la población, es además un procedimiento sencillo y produce estimadores de los parámetros desconocidos próximos a los valores reales de los mismos.

El principal inconveniente de este tipo de muestreo es que necesita un marco adecuado y amplio que no siempre es fácil de conseguir y que no contiene información a priori sobre la población que podría ser útil en la descripción de la misma.

### **MUESTREO SISTEMATICO**

•	•	•
•	•	•
•	•	•

- Se ordenan los individuos de la población y se numeran.

- Se divide la población en tantos grupos como individuos se quieren tener en la muestra. Se selecciona uno al azar en el primer grupo y se elige el que ocupa el mismo lugar en todos los grupos.

-La ventaja principal es que es más sencillo y más barato que el muestreo aleatorio simple, además, se comporta

igual si no hay patrones o periodicidades en los datos.

-La aparición de patrones desconocidos puede llevar a importantes errores en la estimación de los parámetros.

Este tipo de muestreo puede utilizarse, por ejemplo, en encuestas telefónicas programadas mediante ordenador.

### **MUESTREO POR CONGLOMERADOS**


-Se divide la población en grupos de acuerdo con su proximidad geográfica o de otro tipo. (conglomerados).

Cada grupo ha de ser heterogéneo y tener representados todos las características de la población.

Por ejemplo, los conglomerados en un estudio sobre la situación de las mujeres en una determinada zona rural

pueden ser los municipios de la zona.

- Se selecciona una muestra de conglomerados al azar y se toma el conglomerado completo o una muestra del mismo.
- Necesitan menos información previa sobre los individuos particulares.
- Soluciona el problema de los patrones en los datos.
- Si el número de bloques no es muy grande se puede incurrir en errores de estimación si se han incluido conglomerados atípicos.
- Los conglomerados que se realizan teniendo en cuenta proximidad geográfica pueden no tener un significado importante en la población (no responden a una característica real).
- Este tipo de muestreo se utiliza fundamentalmente para reducir los costes de toma de muestras al tomar grupos de individuos completos.

### **MUESTREO ESTRATIFICADO**

• • • • •	• • • •	• • • • • •
• • • •	• • • •	• • • • • •
• • • • •	• • • • •	• • • • •

-Se divide la población en grupos homogéneos (estratos) de acuerdo con las características a estudiar. Por ejemplo, en un estudio de las características socioeconómicas de una ciudad los estratos pueden ser los barrios de la misma, ya que los barrios suelen presentar características diferenciales.

- Se selecciona una muestra aleatoria de cada estrato tratando de que todos los estratos de la población queden representados.
- Permite utilizar información a priori sobre la estructura de la población en relación con las variables a estudiar.
- Obtiene representantes de todos los estratos de la población.
- Diferentes opciones de selección del tamaño de la muestra en los estratos:
  - El mismo número en cada estrato.
  - Proporcional. (La más común)
  - Optima.

### **NOTAS:**

- El problema más importante en la realización de una investigación por muestreo es encontrar el marco adecuado (Lista de los elementos de la población que pueden ser incluidos en la muestra).
- En algunos casos es necesario encontrar una población identificable mayor que la población de interés y que incluya a la misma. Por ejemplo, si queremos realizar una encuesta sobre los trabajadores de la construcción de la ciudad de Salamanca y no disponemos de una lista de los mismos, podemos tomar una lista de los cabezas de

familias trabajadores o de las viviendas ocupadas. El único problema adicional es que la encuesta será más cara.

## 4.- ESTADISTICOS Y DISTRIBUCIONES MUESTRALES

Todo lo que veremos a continuación está pensado para poblaciones infinitas (muy grandes) y con muestreo aleatorio simple. El muestreo aleatorio simple garantiza una muestra representativa de la población y la obtención de observaciones independientes.

Dada una población  $X$ , el proceso de muestreo consiste en obtener, al azar, un valor de la variable  $X$ ,  $x_1$ ; El valor obtenido puede ser cualquiera de los de la población, luego los posibles valores para  $x_1$  son todos los de  $X$ , y por tanto  $x_1$  puede considerarse como una realización particular (observación) de una variable aleatoria  $X_1$  con la misma distribución que  $X$ .

A continuación obtenemos, independientemente de la primera observación, un valor  $x_2$  que puede considerarse como una realización particular de una variable aleatoria  $X_2$  con la misma distribución que  $X$  e independiente de  $X_1$ . Obsérvese que la población no se modifica al extraer uno de sus individuos ya que es infinita. (Si la población es finita podría utilizarse un muestreo con reemplazamiento).

El proceso continúa hasta obtener una muestra de tamaño  $n$ ,  $n$  observaciones  $x_1, x_2, \dots, x_n$  de  $n$  variables aleatorias  $X_1, X_2, \dots, X_n$  independientes e idénticamente distribuidas.

**Definición:** Sea  $X$  una variable aleatoria con f.d.p  $F$ , y sean  $X_1, X_2, \dots, X_n$ ,  $n$  variables aleatorias independientes con la misma f.d.p  $F$  que  $X$ . Se dice que  $X_1, X_2, \dots, X_n$ , son una muestra aleatoria de tamaño  $n$  de  $F$  o bien  $n$  observaciones independientes de  $X$ .

Hemos utilizado letras minúsculas, como en descriptiva, para denotar las observaciones particulares de una muestra, y letras mayúsculas para denotar las variables aleatorias de las que se han tomado. A lo largo de la exposición teórica ambas serán intercambiables y serán utilizadas indistintamente para representar a las correspondientes variables aleatorias.

Otra forma de ver la muestra es como una variable aleatoria multivariante con función de densidad de probabilidad es el producto de las funciones de densidad de cada una de las componentes (ya que son independientes)

$$f(X_1, X_2, \dots, X_n) = f(X_1) f(X_2) \dots f(X_n)$$

donde las funciones de densidad son iguales a la de X. Esta forma de entender la muestra supera el ámbito de un curso introductorio.

Una vez obtenida la muestra la describimos en términos de algunas de sus características fundamentales como la media, la desviación típica, etc ... A tales características las solemos denominar estadísticos.

**Definición:** Un estadístico es una función de los valores muestrales que no depende de ningún parámetro poblacional desconocido.

Un estadístico es también una variable aleatoria ya que es una función de variables aleatorias. Por ejemplo la media muestral

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

es una variable aleatoria de la que tenemos una sola observación

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Veámoslo con un ejemplo sencillo

Supongamos que disponemos de una población finita en la que disponemos de 4 individuos que toman los valores {1, 2, 3, 4}.

Supongamos que obtenemos una muestra sin reemplazamiento de tamaño 2. Las distintas posibilidades son

{1, 2} {1, 3} {1, 4} {2, 3} {2, 4} {3, 4}

Obtendremos, dependiendo de la muestra elegida, las siguientes medias respectivamente:

1.5      2      2.5      2.5      3      3.5

Es claro que la media muestral no es un valor fijo sino que puede considerarse también como una variable aleatoria de la que tenemos una sola observación, la media de la muestra concreta seleccionada.

Dicha variable tendrá una distribución de probabilidad asociada. (En este caso una distribución discreta que toma los valores 1.5, 2, 2.5, 3 y 3.5 con probabilidades 1/6, 1/6, 2/6, 1/6, 1/6, respectivamente).

**Definición:** A la distribución de un estadístico calculado a partir de los valores tomados de una muestra se la denomina distribución muestral del estadístico.

En la mayor parte de los casos supondremos que nuestra población tiene distribución normal y que los estadísticos que vamos a utilizar son la media y la desviación típica (o la cuasi desviación típica).

## 5.- DISTRIBUCIONES MUESTRALES DE LA MEDIA Y LA DESVIACION TIPICA.

Sea  $X_1, X_2, \dots, X_n$ , una muestra aleatoria de una población  $X$  en la que  
 $E(X) = \mu$      $Var(X) = \sigma^2$   
 Entonces el valor esperado (media) y la varianza del estadístico "media muestral" son

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

$$Desv(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

La comprobación del resultado es obvia si tenemos en cuenta que la esperanza de la suma de varias variables aleatorias independientes es la suma de las esperanzas, y que la varianza es la suma de las varianzas, y además que si multiplicamos una variable por una constante, la varianza queda multiplicada por la constante al cuadrado. Entonces

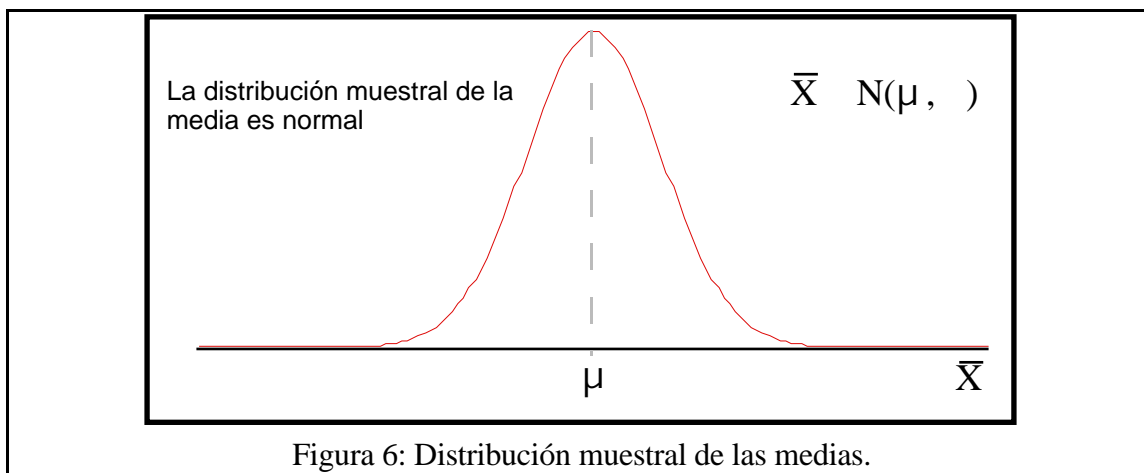
$$E(\bar{X}) = E \left( \frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

$$Var(\bar{X}) = Var \left( \frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n Var(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}$$

Si además, la población es normal, es decir,  $X \sim N(\mu, \sigma^2)$  entonces la media muestral es también normal  $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ .

Basta tener en cuenta las propiedades de la normal que ya se vieron en su momento.

El resultado es importante en estimación ya que, aunque la media poblacional y la media muestral no coincidan, los posibles valores de la media muestral se concentran de forma simétrica alrededor de la media poblacional, además, la dispersión es menor a medida que aumenta el tamaño muestral.



La distribución muestral asociada a varianzas y cuasivarianzas es un poco más compleja y su obtención supera los objetivos del curso, de forma que nos limitaremos a exponerlas.

Sea  $X_1, X_2, \dots, X_n$ , una muestra aleatoria simple de una población  $X \sim N(\mu, \sigma^2)$ , entonces la variable aleatoria

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{2}$$

sigue una ji-cuadrado con  $n-1$  grados de libertad.

Del resultado anterior se deduce que las variables

$$\frac{n S^2}{2} \quad y \quad \frac{(n-1)\hat{S}^2}{2}$$

donde siguen ambas una ji-cuadrado con  $n-1$  grados de libertad.

## 6.- EL TEOREMA CENTRAL DEL LIMITE.

Lo que hemos visto hasta el momento parece bastante restrictivo ya que hemos supuesto, de entrada, que la distribución en la población es normal, pero existen muchos casos en los que no es posible suponer distribución Normal. El siguiente resultado permite trabajar con la normal para la distribución muestral de medias aunque la población no lo sea, y es conocido como Teorema Central del Límite.

Sea  $X_1, X_2, \dots, X_n$ , una muestra aleatoria de una población  $X$  con una distribución de probabilidad no especificada para la que la media es  $E(X) = \mu$  y la varianza  $\text{Var}(X) = \sigma^2$  finita. La media muestral tiene una distribución con media  $\mu$  y varianza  $\sigma^2/n$  que tiende a una distribución normal cuando  $n$  tiende a infinito.

La demostración del resultado excede los límites de un curso introductorio.

La aproximación a la distribución normal es mejor para  $n$  grande ya que se trata de una aproximación y no de una distribución exacta como en el caso de poblaciones normales. En Estadística consideramos  $n$  grande cuando es mayor de 30.

Una consecuencia directa del teorema es que la suma de los valores muestrales sigue una distribución normal de media  $n\mu$  y varianza  $n\sigma^2$ .

El teorema de De Moivre que se explicó en el apartado de la normal puede entenderse también como un caso particular del Teorema Central del Límite.

Sea una población en la que se mide una v.a.  $X$  con distribución binomial  $B(1,p)$ , es decir, toma el valor 1 con probabilidad  $p$  y el valor 0 con probabilidad  $q$ , tiene una media  $p$  y una varianza  $pq$ . Una distribución  $B(n,p)$  puede entenderse como la suma de  $n$  binomiales  $B(1,p)$ , luego aplicando el TCL, si  $n$  es grande la distribución  $B(n,p)$  se puede aproximar por una normal que tiene como media a  $np$  y como varianza  $npq$ .

## 7.- ESTIMADORES Y PROPIEDADES DESEABLES DE LOS ESTIMADORES.

Supongamos ahora que disponemos de una población en la que se mide una variable  $X$  con distribución de forma conocida y parámetros desconocidos, por ejemplo una normal con media y varianzas desconocidas como en el caso práctico que planteábamos

anteriormente.

De la población se extrae una muestra aleatoria simple de tamaño  $n$ ,  $X_1, X_2, \dots, X_n$ . Se trata de calcular, a partir de los valores muestrales, una función de los mismos que proporcione un valor  $\hat{\theta} = u(X_1, \dots, X_n)$  que sustituya al parámetro desconocido de la población  $\theta$ , de forma que ambos sean lo más parecidos en algún sentido. A tal valor obtenido de la muestra se le denomina **estimador**.

Un estimador es también una variable aleatoria. Se trata básicamente de buscar estimadores centrados alrededor del verdadero valor del parámetro y con la menor varianza posible.

Por ejemplo, por simple analogía, si la distribución en la población es normal, la media muestral puede considerarse como un estimador de la media poblacional.

La distancia entre el estimador y el parámetro a estimar puede medirse mediante los que se denomina el error cuadrático medio, que se define como el valor esperado de la diferencia entre el estimador y el verdadero parámetro.

$$ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

El ECM es importante ya que puede escribirse como

$$ECM(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2$$

una es la varianza del estimador y otra el cuadrado del sesgo (concepto que veremos posteriormente).

Consideraremos criterios adicionales para seleccionar estimadores. Las propiedades deseables que ha de tener un estimador para considerarse adecuado son las siguientes:

#### **-Ausencia de sesgo-**

Se dice que un estimador es insesgado (o centrado) si la esperanza del estimador coincide con el parámetro a estimar.  $E(\hat{\theta}) = \theta$ . En caso contrario se dice que es sesgado y a la cantidad  $b(\hat{\theta}) = [\text{Bias}(\hat{\theta})] = E(\hat{\theta}) - \theta$  se la denomina sesgo.

La propiedad es importante ya que los posibles valores del estimador fluctúan alrededor del verdadero parámetro. Por ejemplo, si utilizamos la media muestral como estimador de la media poblacional en una distribución normal, se trata de un estimador insesgado ya que la esperanza de su distribución muestral es la media poblacional  $\mu$ . El hecho de que además, tenga distribución normal, es importante en la práctica, ya que aunque la media muestral y la poblacional no coinciden exactamente, los valores de aquella fluctúan de

forma simétrica alrededor de esta, son valores próximos con probabilidad alta y la dispersión disminuye cuando aumenta el tamaño muestral.

**-Consistencia-**

Se dice que un estimador  $\hat{\theta}$  es consistente si se aproxima cada vez más al verdadero valor del parámetro a medida que se aumenta el tamaño muestral. Más formalmente, un estimador es consistente si  $\Pr\left[|\hat{\theta} - \theta| > \epsilon\right] \rightarrow 0$  cuando  $n \rightarrow \infty$ , para  $\epsilon > 0$ . o dicho de otra forma la distribución del estimador se concentra más alrededor del verdadero parámetro cuando el tamaño muestral aumenta.

La media muestral es un estimador consistente de la media poblacional en una distribución normal, ya que, la varianza de la misma  $\frac{\sigma^2}{n}$  tiende a cero para  $n \rightarrow \infty$ , de forma que la distribución se concentra alrededor del verdadero valor  $\mu$  cuando  $n$  crece.

**-Eficiencia-**

Es claro que un estimador será tanto mejor cuanto menor sea su varianza, ya que se concentra más alrededor del verdadero valor del parámetro. Se dice que un estimador insesgado es eficiente si tiene varianza mínima.

Una cota inferior para la varianza viene dada por la denominada cota de Cramer-Rao.

Sea  $X_1, X_2, \dots, X_n$ . una muestra aleatoria simple de una distribución con densidad  $f(x; \theta)$ . Sujeto a ciertas condiciones de regularidad en la función de densidad, cualquier estimador insesgado verifica que

$$\text{Var}(\hat{\theta}) \geq \frac{1}{nE \left[ \frac{\ln f(X; \theta)}{\theta} \right]^2}$$

A la cantidad  $I_n(\theta) = nE \left[ \frac{\ln f(X; \theta)}{\theta} \right]^2$  se la denomina cantidad de información

de Fisher asociada a una muestra aleatoria simple de tamaño  $n$ .

## 8.-METODOS DE ESTIMACION

### Método de los Momentos

-Consiste en igualar los momentos muestrales y los poblacionales. Prácticamente no se

usa en la investigación actual.

### Método de los Mínimos Cuadrados

-Consiste en minimizar la suma de cuadrados de los errores (diferencias entre valores observados y esperados tras suponer que las observaciones se obtienen como la suma de una parte sistemática o controlada y una parte aleatoria no controlada o fuente de error).

El método es ampliamente utilizado cuando se trabaja con modelos de regresión y técnicas relacionadas.

Ejemplo: Estimación de la media de una población normal.

Cada observación experimental  $x_i$  puede suponerse como la suma de una constante (la media  $\mu$ ) y un error experimental aleatorio ( $\epsilon_i$ )

$$x_i = \mu + \epsilon_i$$

con  $\epsilon_i = x_i - \mu$  con distribución  $N(0, \sigma^2)$ .

El método de los mínimos cuadrados consiste en minimizar la suma de cuadrados de los errores (Diferencias entre valores observados y esperados)

$$D = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (x_i - \mu)^2$$

Derivando con respecto a  $\mu$  e igualando la derivada a cero

$$\begin{aligned} \frac{D}{\mu} &= \sum_{i=1}^n 2(x_i - \mu)(-1) = 0 \\ &\sum_{i=1}^n (x_i - \mu) = 0 \\ &\sum_{i=1}^n x_i \\ \hat{\mu} &= \frac{\sum_{i=1}^n x_i}{n} = \bar{x} \end{aligned}$$

obtenemos la media muestral como estimador de la poblacional.

### Método de la Máxima Verosimilitud

- Consiste en sustituir los parámetros por aquellos valores que maximizan el logaritmo de la función de verosimilitud de la muestra (función de densidad conjunta de todos los valores muestrales en el supuesto de que son independientes).

Ejemplo: Media y varianza de una población normal

Los valores muestrales  $X_1, \dots, X_n$  se supone que son variables aleatorias independientes y todas con distribución  $N(\mu, \sigma^2)$ . La función de densidad conjunta será el producto de las funciones de densidad de cada una de ellas.

$$L(x_1, \dots, x_n / \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2} =$$

$$= \frac{1}{\sigma^n \sqrt{2\pi}^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2}$$

Tomando logaritmos

$$\ln L = -n \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

Derivando con respecto a  $\mu$  y  $\sigma^2$  y resolviendo el sistema se obtienen como estimadores para la media y la varianza

$$\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

### Propiedades de los estimadores Máximo-verosímiles

Los estimadores máximo-verosímiles juegan un papel importante en Estadística debido a que se obtienen mediante un método simple y tienen buenas propiedades con respecto a sesgo eficiencia y consistencia.

Bajo ciertas condiciones de regularidad se verifica:

- Si existe un estimador insesgado y de varianza mínima, cuya varianza alcance la cota de Cramer-Rao, este estimador es máximo verosímil y es la única solución de la ecuación de verosimilitud.
- Si el estimador es sesgado, su sesgo tiende a cero al aumentar el tamaño de la muestra,

además es asintóticamente eficiente (Eficiente para n grande).

- Existe una solución de la ecuación de verosimilitud que proporciona un estimador consistente y asintóticamente normal.  $N\left(\mu, \sqrt{\frac{1}{I_n(\mu)}}\right)$ . Donde  $\frac{1}{I_n(\mu)}$  es la varianza mínima o cota de Cramer-Rao.

## 9.-ESTIMADORES PUNTUALES DE LOS PARAMETROS DE UNA POBLACION NORMAL

Sea una muestra aleatoria simple,  $X_1, X_2, \dots, X_n$  de una población con distribución  $N(\mu, \sigma^2)$ .

**-Estimador de la media**

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

Se trata de un estimador eficiente (insesgado y de varianza mínima).

La distribución muestral de la media es :

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

La cantidad  $ES = \frac{S}{\sqrt{n}}$  estima a la desviación típica de la media  $\frac{\sigma}{\sqrt{n}}$  y se denomina error estándar de la media, por esta razón se dice que el error estándar de la media mide la variabilidad de la media en el muestreo.

**-Estimador de la Varianza**

Varianza muestral (estimador sesgado).

$$\hat{S}^2 = S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

Cuasi-varianza muestral (estimador insesgado)

$$\hat{\sigma}^2 = \hat{S}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Distribuciones muestrales asociadas

---

$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{2}$	$\frac{2}{n-1}$	$\frac{nS^2}{2}$	$\frac{2}{n-1}$	$\frac{(n-1)\hat{S}^2}{2}$	$\frac{2}{n-1}$
--	-----------------	------------------	-----------------	----------------------------	-----------------

---

## 10.- ESTIMADORES DE LOS PARAMETROS DE LAS DISTRIBUCIONES DISCRETAS MAS USUALES

Se dispone de una muestra de tamaño  $n$  en la que el resultado de la observación es una variable dicotómica (dos posibles resultados). Una variable cualitativa con más de dos resultados puede reducirse a una dicotómica sin más que agrupar algunas de las categorías.

Se trata de estimar la probabilidad  $p$  de éxito en la población.

La variable  $X$ = número de éxitos en las  $n$  pruebas, puede tener distintas distribuciones dependiendo de las condiciones en las que se toma la muestra.

### -BINOMIAL

Si se toman muestras de poblaciones infinitas o se realiza un muestreo con reemplazamiento de una población finita. Se realizan  $n$  pruebas y se contabiliza el número de éxitos en las  $n$  pruebas. El estimador de la proporción de éxito es

$$\hat{p} = \frac{\text{n}^\circ \text{ de éxitos}}{n} = \frac{X}{n}$$

Aproximando  $X$  mediante una distribución normal, la distribución muestral del estimador de la probabilidad de éxito para muestras grandes es

$$\hat{p} = \frac{X}{n} \quad N\left(p, \sqrt{\frac{pq}{n}}\right)$$

### -HIPERGEOMETRICA

Si se toman muestras sin reemplazamiento de una población finita de tamaño  $N$  conocido.

$$\hat{p} = \frac{\text{n}^\circ \text{ de éxitos}}{n} = \frac{X}{n}$$

Aproximando  $X$  mediante una distribución normal, la distribución muestral del estimador de la probabilidad de éxito para muestras grandes es

$$\hat{p} = \frac{X}{n} \quad N\left(p, \sqrt{\frac{pq}{n} \frac{N-n}{N-1}}\right)$$

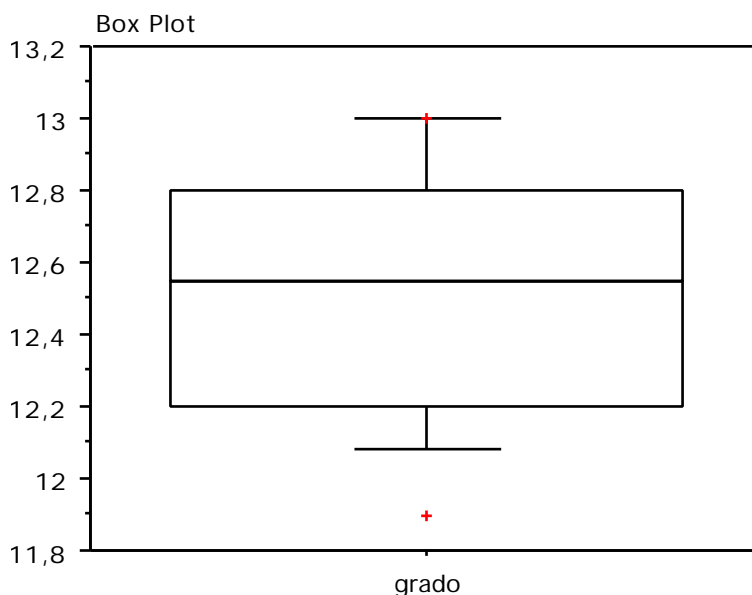
## 11.- EJEMPLO

En el apartado 2 comenzamos a exponer un ejemplo sobre una investigación real para estudiar el grado alcohólico de los vinos jóvenes de la denominación de Ribera de Duero. Habíamos caracterizado la población que queríamos estudiar y los objetivos del estudio. El paso siguiente consiste en tomar una muestra, más tarde trataremos el tema del tamaño de muestra necesario para obtener una determinada precisión. La tabla siguiente contiene una muestra aleatoria simple de 14 observaciones.

RIBERA DE DUERO													
12,8	12,8	12,5	11,9	12,5	12,1	12,2	12,6	13,0	12,4	12,6	12,2	12,8	13,0

Tabla 1: Grado alcohólico de 14 vinos de la denominación de Ribera de Duero.

La primera hipótesis que hicimos era que la población es normal. Si disponemos de una muestra representativa, la distribución de frecuencias de la muestra obtenida debe ser aproximadamente simétrica. Aunque es muy difícil asegurar la posible normalidad con una muestra de sólo 14 observaciones, una primera comprobación puede realizarse mediante un Box-Plot.



El Box plot presenta un aspecto aproximadamente simétrico sin muchas observaciones extremas por lo que, en principio, no hay ninguna razón para suponer una distribución no normal.

Los estimadores de los principales parámetros aparecen en la tabla siguiente junto con otras características útiles para describir la muestra. El estimador de la media de la población es 12.529; este valor probablemente no coincide con la media de la población,;sin embargo, teniendo en cuenta la distribución muestral de medias, es con probabilidad alta, cercano a la misma. La variabilidad de la media en el muestreo se estima mediante el error estándar de la media, que en este caso es 0.09.

El estimador de la varianza es 0.115; el estimador es insesgado ya que ha sido calculado utilizando  $n-1$ .

Otro indicio de que la distribución es aproximadamente simétrica y, por tanto, no muy alejada de la normal, es que la diferencia entre la media y la mediana es muy pequeña.

Descriptive Statistics

	grado
Mean	12,529
Std. Dev.	,338
Std. Error	,090
Count	14
Minimum	11,900
Maximum	13,000
# Missing	0
Variance	,115
Coef. Var.	,027
Median	12,550

## 12.- ESTIMACION POR INTERVALOS

### INTRODUCCION

Dada una muestra aleatoria  $X_1, X_2, \dots, X_n$ , de una población con función de densidad  $f(x; \theta)$  Un intervalo de confianza, de extremos  $L_1$  y  $L_2$ , para el parámetro  $\theta$  de la población es un par ordenado de funciones reales de las  $n$  medidas de la muestra

$$I = [L_1(X_1, \dots, X_n); L_2(X_1, \dots, X_n)]$$

construidas de forma que la probabilidad de que los extremos contengan al verdadero valor del parámetro es un valor prefijado  $1 - \alpha$ . Al número  $1 - \alpha$  se le denomina "nivel de confianza".

El nivel de confianza suele ser 0,95 (95%) ó 0,99 (99%). La interpretación práctica es sencilla, por ejemplo si el nivel de confianza es del 95%, significa que en el 95% de las veces que repitiéramos el experimento, el intervalo de confianza calculado contendría al verdadero valor del parámetro y en el 5% restante el intervalo no contendría el verdadero valor.

Una vez que el intervalo de confianza ha sido particularizado para una muestra concreta, el intervalo obtenido contiene o no contiene al verdadero valor del parámetro, con probabilidad 1, por esa razón, cuando ya tenemos un valor concreto hablamos de confianza y no de probabilidad. Confiamos en que el intervalo que hemos calculado sea del 95% que contiene el verdadero valor.

### INTERVALO DE CONFIANZA PARA LA MEDIA DE UNA POBLACION NORMAL DE VARIANZA CONOCIDA

Supongamos que disponemos de una población en la que tenemos una v.a. con distribución  $N(\mu, \sigma^2)$  con  $\sigma^2$  conocida (de estudios previos, por ejemplo).

Obtenemos una muestra de tamaño  $n$  y deseamos estimar la media  $\mu$  de la población.

El estimador puntual de la misma es la media muestral cuya distribución muestral es conocida

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

la cantidad

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

tendrá distribución normal estándar.

Sobre la distribución  $N(0, 1)$  podremos seleccionar dos puntos simétricos  $-z_{\alpha/2}$  y  $z_{\alpha/2}$ , tales que

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

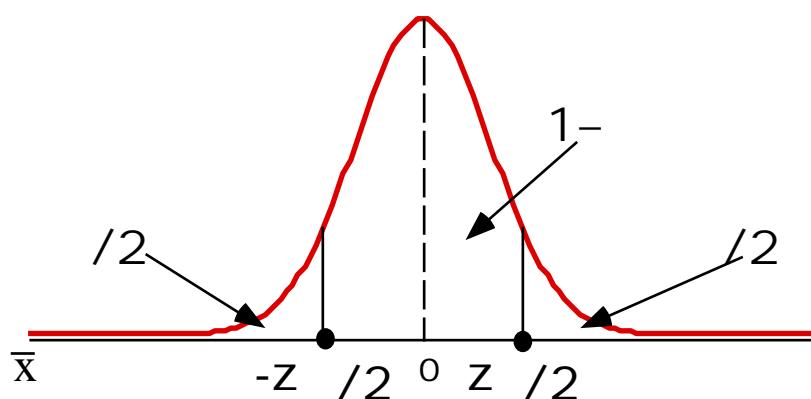


Figura 1: Selección de los puntos críticos para el cálculo del intervalo de confianza.

Sustituyendo  $Z$  por su valor en este caso particular

$$P\left(-z \leq \frac{\bar{X} - \mu}{\sqrt{n}} \leq z\right) = 1 -$$

Despejando la media muestral y la varianza

$$P\left(\bar{X} - z \sqrt{n} \leq \mu \leq \bar{X} + z \sqrt{n}\right) = 1 -$$

que verifica las condiciones de la definición.

Así, el intervalo de confianza para la media puede escribirse como

$$I_{\mu}^{1-\alpha} = \left[ \bar{X} - z \sqrt{n}, \bar{X} + z \sqrt{n} \right] = \bar{X} \pm z \sqrt{n}$$

en la práctica, de todos los posibles valores de  $\bar{X}$  tenemos uno sólo  $\bar{X}$  y por tanto un único intervalo de todos los posibles para distintas muestras

$$I_{\mu}^{1-\alpha} = \bar{X} \pm z \sqrt{n}$$

La importancia del intervalo de confianza para la estimación está en el hecho de que el intervalo contiene información sobre el estimador puntual (valor central del intervalo) y sobre el posible error en la estimación a través de la dispersión y de la distribución muestral del estimador. Obsérvese que el error en la estimación está directamente relacionado con la distribución muestral del estimador y con la varianza poblacional, e inversamente relacionado con el tamaño muestral.

El gráfico siguiente ilustra la interpretación del nivel de confianza para el intervalo de confianza para la media de una distribución normal con varianza conocida. Para los distintos posibles valores de la media, representados mediante su distribución muestral, obtenemos distintos intervalos de confianza. La mayor parte incluye al verdadero valor del parámetro, pero el resto no. Concretamente el 95% lo incluye y el 5% no, si el nivel de confianza es del 95%.

En la práctica disponemos de una única repetición del experimento, y por tanto de un único intervalo de confianza, el señalado en negro en el gráfico, por ejemplo. Confiamos

en que nuestro intervalo sea de la mayoría que con tiene al verdadero valor objetivo aunque no tenemos la seguridad de que sea así, tenemos concretamente un riesgo del 5% de equivocarnos.

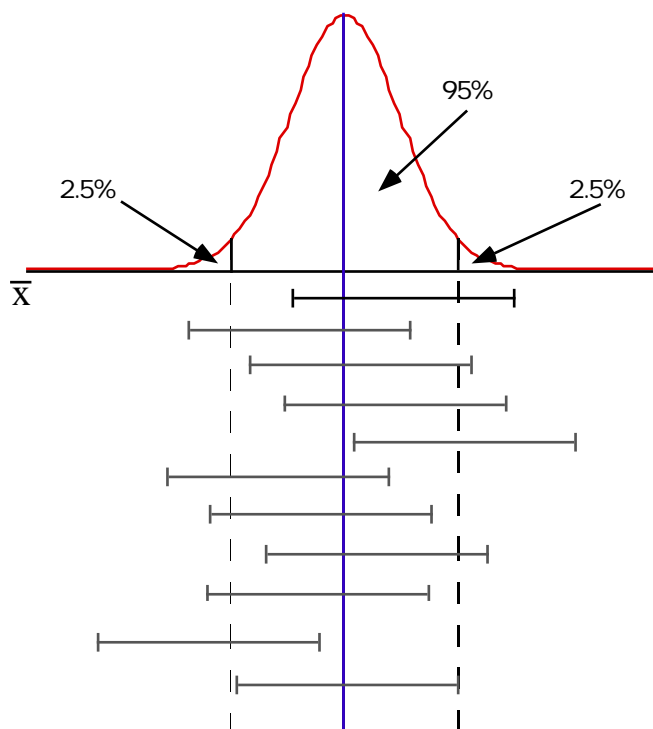


Figura 2: Interpretación del nivel de confianza en el intervalo para la media de una distribución normal.

## LONGITUD DEL INTERVALO Y ERROR EN LA ESTIMACIÓN

En la práctica hemos de tratar de que la longitud del intervalo de confianza sea lo más pequeña posible, es decir, que el error en la estimación sea lo más pequeño posible.

$$\text{long} = 2z \cdot \frac{1}{2} \sqrt{\frac{\sigma^2}{n}}$$

Esto puede conseguirse modificando las distintas cantidades que aparecen en la fórmula: el nivel de confianza, a través del valor crítico, la variabilidad y el tamaño muestral. Estudiaremos cada una por separado

### -NIVEL DE CONFIANZA

La longitud del intervalo de confianza aumenta al aumentar el nivel de confianza ya que el valor crítico de la distribución es mayor. Si consideramos un nivel de confianza del 100%, el intervalo de confianza será  $[- ; + ]$  que, evidentemente contiene al verdadero valor del parámetro pero no es de ninguna utilidad en la práctica. Si disminuimos el nivel de confianza también disminuye la longitud, sin embargo conviene mantenerlo en unos límites razonables que suelen ser del 95% o del 99% en la mayor parte de las aplicaciones.

### -VARIANZA

La longitud del intervalo de confianza disminuye con la varianza, es decir, la estimación será más precisa cuanto menor sea la variabilidad en la población, lo que significa que la población es más homogénea. En la práctica es posible obtener estimaciones más precisas, por ejemplo, restringiendo la población a conjuntos lo más homogéneos posible.

**-TAMAÑO MUESTRAL**

La longitud del intervalo de confianza disminuye al aumentar el tamaño muestral, lo que significa que se obtienen estimaciones más precisas cuanto mayor sea el tamaño muestral. Debido a consideraciones prácticas de coste y tiempo, en general no es posible aumentar indefinidamente el tamaño muestral para obtener estimaciones más precisas, es por ello que en la práctica se selecciona el tamaño muestral necesario para obtener una determinada precisión, establecida a priori.

**CALCULO DEL TAMAÑO MUESTRAL PARA ESTIMAR LA MEDIA DE UNA POBLACION CON UNA DETERMINADA PRECISION**

Supóngase que un investigador está interesado en estimar la media de una población normal de forma que la diferencia existente entre la media muestral que obtendrá del experimento y la media poblacional verdadera, esté por debajo de un error prefijado de antemano.

$$\begin{matrix} |\bar{x} - \mu| & E \\ \bar{x} - E & \mu & \bar{x} + E \end{matrix}$$

Teniendo en cuenta el intervalo de confianza

$$P(\bar{x} - z_{/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{/2} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

podemos escribir

$$E = z_{/2} \frac{\sigma}{\sqrt{n}}$$

Despejando n de la igualdad

$$E^2 = z_{/2}^2 \frac{\sigma^2}{n}$$

$$n = \frac{z_{/2}^2 \sigma^2}{E^2}$$

obtenemos la expresión deseada para el tamaño muestral.

Obsérvese que n ha sido calculado en el supuesto de que la variabilidad es conocida. Si no es así, la variabilidad aproximada puede obtenerse de trabajos bibliográficos o experimentos previos o a partir una muestra piloto con unas pocas observaciones.

Obsérvese que en el cálculo del tamaño muestral se han igualado el error fijado a priori con el error en la estimación obtenido del intervalo de confianza y que este último incluye el nivel de confianza. En este apartado un nivel de confianza del 95%, por ejemplo, implicaría que en el 95% de las veces que repitiéramos el experimento con el tamaño

muestral calculado, obtendríamos un error por debajo del prefijado, mientras que en el 5% restante obtendríamos un error superior.

### **INTERVALO DE CONFIANZA PARA LA MEDIA DE UNA POBLACION NORMAL CON VARIANZA DESCONOCIDA**

La situación práctica más habitual es aquella en la que no se conoce la varianza de la población, que habrá que estimar a partir de los datos muestrales. Utilizaremos la cuasi-varianza muestral como estimador por sus buenas propiedades.

La distribución muestral asociada a la cuasi-varianza es la siguiente:

$$\frac{(n-1)\hat{S}^2}{2} \quad \frac{2}{n-1}$$

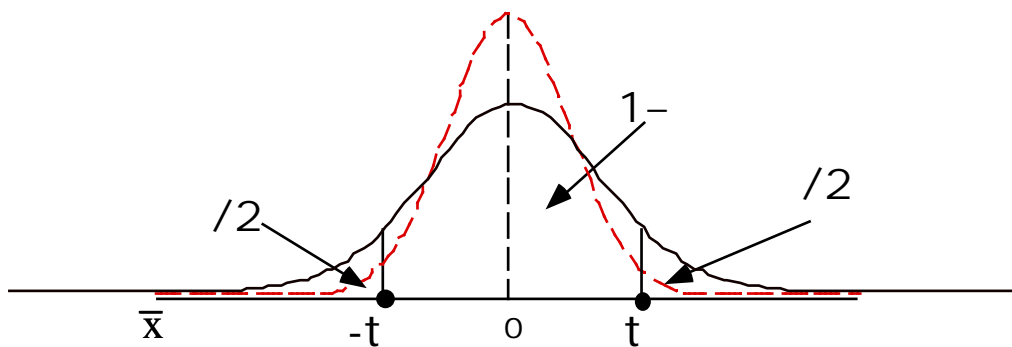
Teniendo en cuenta la distribución normal asociada a las medias y combinándola con la ji-cuadrado, obtenemos una distribución t de Student:

$$t = \frac{N(0,1)}{\sqrt{\frac{2}{n-1}}} = \frac{\frac{\bar{X} - \mu}{\sqrt{n}}}{\sqrt{\frac{(n-1)\hat{S}^2}{2(n-1)}}} = \frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \quad t_{n-1}$$

Siguiendo el mismo proceso que en el caso de la normal el intervalo de confianza resulta

$$I_{\mu}^{1-\alpha} = \bar{X} \pm t_{n-1, \frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}$$

Obsérvese la similitud con el intervalo calculado para la distribución normal, salvo en el valor crítico y en que la varianza ha sido estimada a partir de la muestra.



Desde el punto de vista práctico esto implica que los valores críticos son un poco más grandes y, por tanto el intervalo tiene mayor longitud, este es el precio que debemos pagar a cambio de no conocer la varianza de la población.

Cuando el tamaño muestral es grande, la distribución  $t$  es muy similar a la normal, de forma que pueden intercambiarse los valores críticos correspondientes. El intervalo de confianza para la media en muestras grandes se puede escribir como

$$I_{\mu}^{1-\alpha} = \bar{X} \pm z_{\alpha/2} \frac{\hat{S}}{\sqrt{n}}$$