Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)
Marzo de 2021

Índice general

Capítulo 1

Introducción

¿Cuales son las estrategias básicas de toda investigación?

- Explorar
- Identificar
- Clasificar
- Relacionar
- Contrastar
- Explicar
- Predecir

¿Históricamente, a qué necesidades atendía la Estadística?

- Padrones de los estados: registro y descripción de los datos.
- Cálculo de probabilidades: prácticas de los juegos de azar.

¿Qué problemas resuelve la Estadística?

- Descripción de datos
- Análisis de muestras
- Contraste de hipótesis
- Medición de relaciones[1]
- Predicción y Proyección[2,3]

[1] ¿Establece relaciones causa-efecto?
[2] ¿Son lo mismo?
[3] Modelos que predigan mejor los datos

Capítulo 2

Método científico y Método estadístico

2.1. Método científico

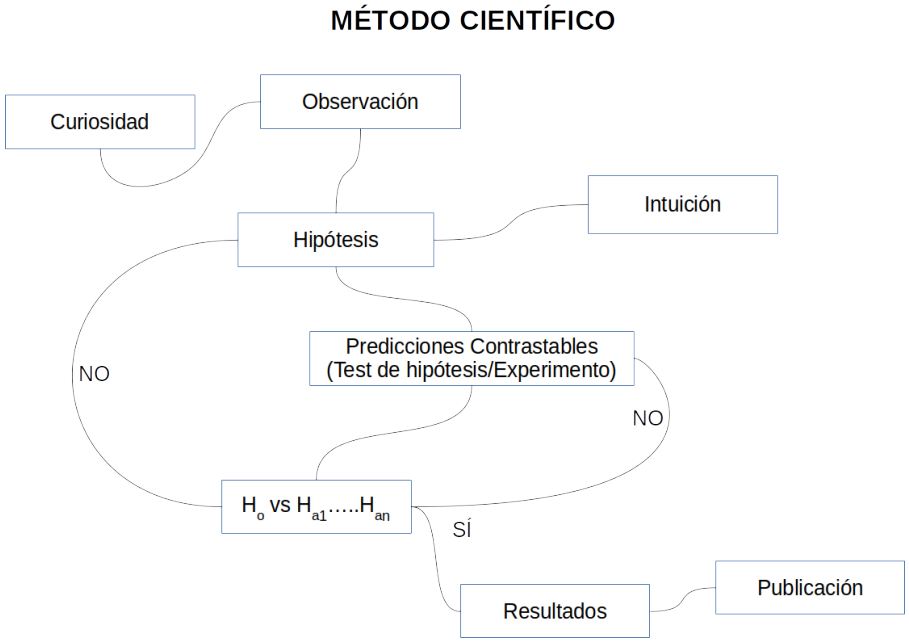


Figura 2.1: Un esquema más sobre el método científico

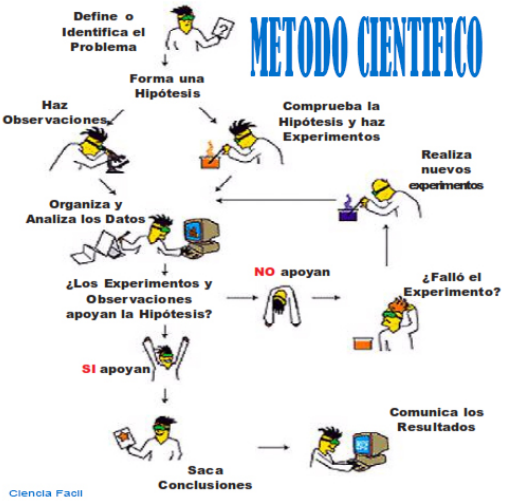


Figura 2.2: Otro más sobre el método científico

2.2. Desarrollo del trabajo estadístico

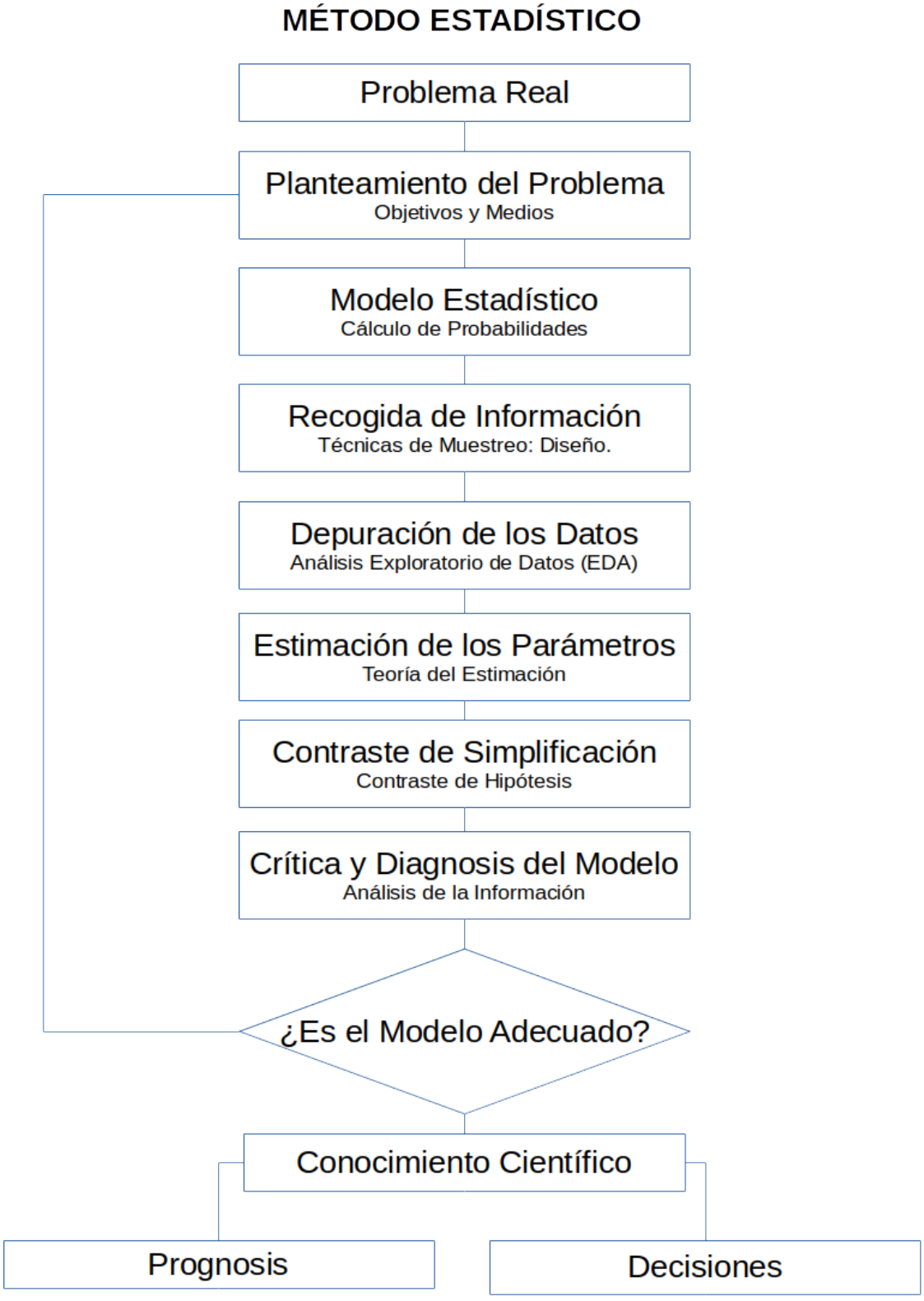


Figura 2.3: Un esquema más sobre el método estadístico

2.2.1. Planteamiento del problema

El primer paso consiste en definir los objetivos del estudio y relacionar estos con valores númericos de variables observables. También hay que tener bien definida la población sobre la que se va a actuar, las variables que mediremos y cómo hacerlo.

2.2.2. El modelo estadístico

El modelo estadístico más utilizado consiste en la descomposición de los valores de la variable en parte sistemática y parte aleatoria.

Se pueden clasificar como:

	Estáticos	**Dinámicos**
Extrapolativos	$Y = \mu + \varepsilon$	$Y = \mu + \vartheta\ Y_{t-1} + \varepsilon_t$
Explicativos	$Y = \mu + \beta\ X + \varepsilon$	$Y = \mu + \beta\ X_t + \vartheta\ Y_{t-1} + \varepsilon_t$

Dependiendo del **momento** en el cual se estudia la variabilidad tendremos dos tipos de modelos:

I. **Estáticos**

No tienen en cuenta el tiempo, son modelos aplicables o aplicados en un momento temporal dado, son los utilizados habitualmente en Biología y Química.

Se pueden clasificar en:

I.1. **Estáticos extrapolativos**: definidos por una variable.

I.2. **Estáticos explicativos**: definido por dos o más variables.

II. **Dinámicos**

Se aplican a situaciones que cambian a lo largo del tiempo, donde este aparece explícitamente.

Se pueden clasificar en :

II.1. **Dinámicos extrapolativos**: como el modelo de Box-Jenkins (B-J), que separa la parte estocástica de la variabilidad. Se usa por ej.: en macroeconomía.

II.2. **Dinámicos explicativos**: igual que B-J pero con dos variables y permite ver su relación, por ej.: Modelos de Regresión Dinámica

Es decir, que dependiendo de la **variabilidad** que tenga el modelo, diremos que son modelos:

- **Extrapolativos**:
  Aquellos que explican la variabilidad tomando únicamente como información sus valores pasados. La parte sistemática del modelo es función de los valores pasados observados.
- **Explicativos**:
  Aquellos que tienen en cuenta el efecto de otras variables. La parte sistemática será la suma de la parte sistemática extrapolativa y del efecto de las variables explicativas.

Otra posible clasificación podría ser:

- **Modelos Descriptivos**:
  Se ocupan de ordenar, clasificar, hacer visibles los datos y también medir relaciones simples entre variables.
- **Modelos Inferenciales**:
  Se hace contraste de hipótesis, que puede ser paramétrica o no-paramétrica.
- **Modelos Multivariados**:
  Se explican simultáneamente el comportamiento de varias variables, haciendo uso intensivo del cálculo matricial. Necesita obligadamente un ordenador debido a su complejidad.

2.2.3. La información

La recogida de información y la forma que la Estadística se aproxima al método científico puede ser de dos formas:

- **Por procedimiento muestral (muestreo):**
  El muestreo consiste en observar *pasivamente* una muestra de las variables y anotar sus valores.
  No se fija ninguna condición. Se utiliza principalmente en modelos extrapolativos. Ej. Modelo de efectos aleatorios (correlación).
- **Por diseño experimental:**
  El diseño de experimentos consiste en fijar o controlar determinadas variables y observar la respuesta de las otras. El diseño experimental debe utilizarse cuando se quiera construir un modelo explicativo, ya que solo tendremos una **base sólida** para juzgar relaciones de "causalidad" entre variables, cuando los datos se obtengan mediante diseño experimental. Ej. Modelo de efectos fijos (regresión).

Por ejemplo:

Regresión / Correlación		
	Diseño Experimental	Muestreo
Pros	Influencia clara entre variables	Puedo hacer predicciones de comportamiento de variables, siempre dentro del rango analizado.
Cons	.- Puedo introducir artefactos. .- No puedo predecir nada a otras condiciones.	No me asegura relación causa-efecto entre variables.

Cuadro 2.1: Ejemplo para Regresión/Correlación del: Diseño experimental vs Muestreo

2.2.4. EDA. Análisis Exploratorio de Datos

En la fase de depuración de la muestra, se espera que alrededor de un 5 % de las observaciones tengan errores de medición, por tanto conviene utilizar técnicas estadísticas simples como las técnicas de análisis exploratorio de los datos (EDA), para así poder identificar los posibles errores, valores faltantes("missing values") o los valores anómalos ("outliers").

2.2.5. Estimación de Parámetros

Los modelos estadísticos dependen de constantes conocidas como **parámetros**[1]. En la fase de estimación se utiliza la información disponible para decidir el valor concreto de estos y cuantificar el posible error de la estimación.

2.2.6. Contraste de Simplificación

Una vez estimados los valores de los parámetros, estudiaremos si el modelo puede simplificarse.

El objetivo es conseguir un modelo lo más simple posible, sin más parámetros de los necesarios. "Entre dos modelos que expliquen unos datos igualmente, es mejor el más sencillo" (navaja de Ockham)[2]. Esta fase es especialmente importante en los modelos explicativos.

Las hipótesis tienen que ser contrastables favorablemente => leyes (ecuaciones matemáticas).

2.2.7. Crítica y Análisis del modelo

Los resultados de las dos etapas anteriores se obtienen suponiendo que el modelo es correcto. Esta fase investiga la compatibilidad entre los valores empíricos y el modelo. Es de especial interés comprobar, que la parte aleatoria no contiene ninguna estructura sistemática.

Si después de esta fase se acepta el modelo como correcto, se usará para tomar decisiones o predecir valores de la variable. En caso contrario, se volverá a la segunda fase y se reformulará el modelo, repitiendo el proceso hasta encontrar un modelo más adecuado.

[1] Tened mucho cuidado con el significado del término parámetro, que está muy contextualizado.
[2] O principio de parsimonia.

Capítulo 3

Diferenciación o Clasificación

La estadística se puede clasificar en relación a varios conceptos:

3.1. Estadística descriptiva vs inferencial

La estadística **descriptiva** realiza una exploración de los datos sin tomar decisiones respecto a los valores de los parámetros en la población, se refiere a la construcción de tablas, gráficos, índices etc. En cambio, la estadística **inferencial** aplica los resultados de la estadística descriptiva sobre la población, estimando los valores de los parámetros poblacionales.

3.2. Estadística paramétrica vs no-paramétrica

La estadística **paramétrica** tiene en cuenta la distribución de los datos, estimando parámetros como la media (μ) y la varianza paramétrica(σ^2).

La estadística **no-paramétrica** no asume ningún tipo de distribución de los datos, es de distribución libre.
Podemos tener una estadística inferencial tanto paramétrica como no-paramétrica, ya que ambas hacen contraste de hipótesis.

Condiciones para aplicar los métodos no-paramétricos:

- Cuando los datos son enumerativos o representan el número de observaciones en una categoría. Ej.: frecuencias en escala de medida débil.
- Nivel de medida ordinal. Comparación de orden.
- Cuando la escala de medida es fuerte, pero **no** nos interesan los parámetros de la distribución poblacional (μ, σ^2)
- No se asume que los parámetros de la población sean por ej.: normales. No nos planteamos que la distribución pueda ser normal o de otro tipo de distribución.

Comparación de técnicas, para la estadística no-paramétrica vs la paramétrica:

- Aparato matemático simple. Los no-paramétricos son más simples.
- Facilidad de aplicación. Son más fáciles.
- Rápidos de aplicar. Igual de rápidos a pesar de que a veces necesitan más tiempo para la ordenación.
- Eficiencia estadística. Son menos eficientes, en principio, que los paramétricos.
- Niveles de significación. Menor significación, en principio, que los paramétricos.
- Aplicación a problemas reales. Es aplicable siempre, mayor aplicabilidad que paramétricos.
- Tamaño de la muestra. Se pueden usar en muestras reducidas.

- Robustez de la prueba. Mayor que los paramétricos, un dato que sea dispar en una técnica no-paramétrica es muchas veces un dato más, pero en cambio en una paramétrica puede verse muy afectada, ej.: los "outliers" o valores anómalos[1].

No-paramétricos	**Paramétricos**	**ERA** (Eficacia Relativa Asintótica)
Prueba de los signos	z o t de Student	0.637
Prueba de Wilcoxon	z o t de Student	0.955 (no detectan el 4 % de las diferencias ≅ a paramétricos)
Prueba de Friedman χ^2	F de Snedecor	0.955
U de Mann-Whitney Kruskal-Wallis	z o t de Student	

Cuadro 3.1: Ejemplo de eficacia entre métodos para y no-paramétricos

3.3. Estadística univariante vs multivariante

Dependiendo del número de variables que se estudian en conjunto, la estadística se divide en **univariante**, si estudia la relación entre los casos de una o dos variables. La estadística **multivariante** entra en juego cuando hay tres o más variables, teniendo en cuenta la interacción entre ellas y los casos.

	Una Muestra	**Dos o más Muestras**
Un conjunto de variables	**I**	**II**
	Análisis de Componentes Principales (ACP) Análisis Factorial	Análisis discriminante/clasificación MANOVA Análisis canónico (poblaciones) Análisis factorial de correspondencias Métodos Bi-Plot MSD (escalamiento multidimensional)
Dos o más conjuntos de variables	**III**	**IV**
	Regresión múltiple Correlación múltiple Correlación canónica Correlación parcial y múltiple	ANCOVA multivariado

Cuadro 3.2: Estadística uni vs multivariante.

En el segundo cuadrante (II) el Análisis Discriminante se usa en taxonomía numérica, que nos da una línea de corte entre especies, o también en medicina para la toma de decisión entre parto o cesárea. El MANOVA es semejante al ANOVA, o sea analiza la varianza multivariante (basándose en la distancia de Mahalanobis). El Análisis Canónico se usa para diferenciar poblaciones, el Análisis Factorial de Correspondencias mejora al ACP.

En el cuadrante III por ej.: la Correlación Canónica se usa en meteorología, donde los conjuntos de variables son por ej. los días, las zonas etc.

[1] O extraños, aberrantes o extremos.

Capítulo 4

Los datos

Una variable se puede definir como: *"una propiedad con respecto a la cual los individuos de una población difieren de algún modo verificable"*. Si en todos los individuos de la población la propiedad no difiere, no es de interés estadístico. Puede ser de interés la longitud, el peso, el contenido en ácidos grasos, el color etc. En cambio no son de interés, la presencia de vértebras en mamíferos o el número de patas de un ave.

Hay muchas formas de clasificar las variables y los datos.

Las variables pueden clasificarse en tres grandes grupos:

- **atributos**: no pueden medirse ni ordenarse. Se deben expresar cualitativamente, como son los colores.
- **variables clasificables en rangos**: no pueden medirse pero sí pueden clasificarse en orden dependiendo de su magnitud.
- **variables medibles**: son aquellas cuyos valores pueden expresarse de forma numéricamente ordenada.

  Pueden ser de dos clases:

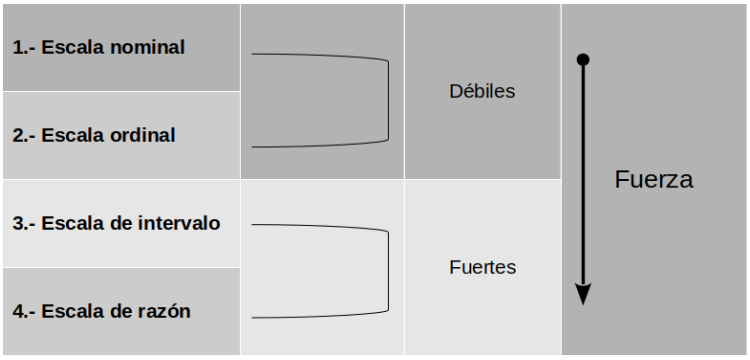
  - **variables discretas** o merísticas, que sólo tienen valores numéricos fijos, sin posibilidad de valores intermedios.

  - **variables continuas**, que pueden tomar un número infinito de valores entre dos puntos determinados.

4.1. Escalas de medida

Para cada tipo de variable se usará una escala de medida adecuada. En el caso de variables cualitativas o atributos se deberá emplear la escala de medida nominal, que simplemente proporciona características cualitativas de la variable. En las variables clasificables en rangos, la escala de medida a usar será la ordinal, y las variables medibles (continuas o discretas) se medirán en escala de intervalos o de razones (también llamada de cocientes).

Escalas de medida:



Cuadro 4.1: Escalas de medidas.

Cualidades para cada tipo de escala:

1. Medida de datos no ordenados, ej.: Macho, Hembra etc. se pueden asignar claves numéricas 1/2.
2. No se pueden asignar a números reales pero sí a clases que tengan un orden o rango, ej.: picoplancton, nanoplancton, microplancton o análisis granulométrico.
3. Podemos asignar un valor númerico concreto a cada intervalo pero no existe un cero absoluto. Ej.: tiempo.
4. Además de lo anterior, en la escala de razón o cocientes existe un cero absoluto, ej.: altura, peso.

Las escalas fuertes, utilizan prácticamente las mismas técnicas estadísticas.

Escala de medida	**Índice de Tendencia Central**	**Transformación**	**Test**
Nominal	Moda	Biunívocas	Igualdad
Ordinal	Mediana/Moda	Monótonas	Orden creciente
Intervalos	Media/Mediana/Moda	Lineales positivas	Igualdad de diferencias
Razones	Media/Mediana/Moda	Lineales positivas	Igualdad de razones

La moda es un índice de tendencia débil. La mediana es un índice de tendencia central muy bueno (mejor que la moda) y además es más "robusto" que la media.

4.2. Tipos de datos y su preparación

Una posible clasificación de los diferentes tipos de datos, podría ser:

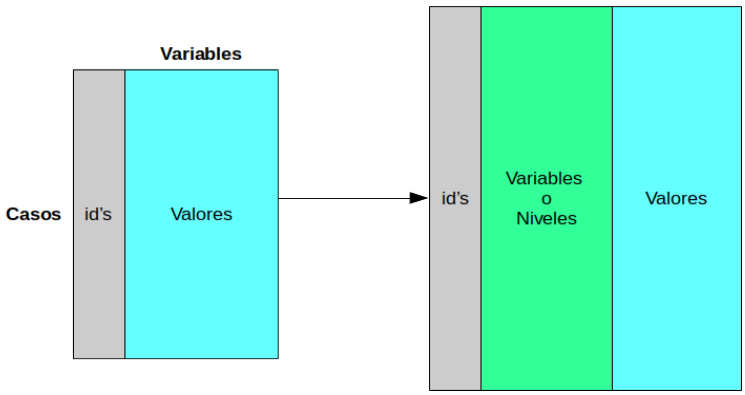
<table>
<tr><th colspan="4">Tipos de datos</th></tr>
<tr><td rowspan="2">Doble estado</td><td colspan="3">Presencia/Ausencia</td></tr>
<tr><td colspan="3">Excluyentes</td></tr>
<tr><td rowspan="4">Multiestados</td><td rowspan="2">Cualitativos</td><td colspan="2">sin secuencia lógica (nominal)</td></tr>
<tr><td colspan="2">con secuencia lógica (ordinal)</td></tr>
<tr><td rowspan="2">Cuantitativos</td><td>discontinuos</td><td rowspan="2">(intervalos/razones)</td></tr>
<tr><td>continuos</td></tr>
</table>

Cuadro 4.2: Tipos de datos.

Aunque existen más tipos de clasificaciones, en esencia sólo existen cuatro tipos de datos asociados las cuatro escalas de medida, es decir: **nominales**, **ordinales** , **intervalos**, **razones o cocientes**.

Preparación, manipulación y buenas prácticas sobre los datos:

A veces es necesario tener los datos en formato ancho y largo. El formato ancho, tiene una columna para cada variable; mientras que en el formato largo, cada fila es una combinación única de variable de identificación.



Cuadro 4.3: Formato ancho y formato largo en los marcos de datos.

Decálogo:

1. Tener identificadores para los casos.
2. No truncar, ni redondear nunca.
3. Prever suficientes columnas.
4. Uso del punto decimal. No usar la coma (delimitador en ASCII) y bien indentados.
5. Asignar valores a los datos que faltan. Decidir una codificación que sea clara para los "missing values"(NA o mv).
6. Mejor números que letras, pero no ceros (en algunas técnicas se consideran NA). Conviene codificar (de forma identificable) los datos no numéricos.
7. No bajar la escala de medida aunque se pueda hacer, ni por supuesto subirla. Los datos deben de estar con toda su significación (precisión) y en escala de razón si es posible.
8. Las variables pueden agruparse de modo significativo.
9. Dos o más variables se pueden combinar sin perder información.
10. Si se transforman, no perder la información original. Pensar bien en las implicaciones sobre la alteración y nueva significación de la información transformada.

4.3. Exactitud, precisión

Exactitud: es la proximidad de un valor medido o calculado al valor real.

Precisión: es la proximidad de dos medidas repetidas de la misma cantidad. A menos que el instrumento de medida este sesgado, la precisión conducirá a la exactitud.

La mayoría de las variables continuas son aproximadas, el valor exacto de la medida individual es desconocido. La última cifra implicará precisión, los límites entre los cuales creemos que se encuentra la verdadera medida. Por ej. una medida de 13.3, implica que la verdadera longitud se encuentra entre 13.25 y 13.35. Entre estos límites implícitos, no sabemos donde se encuentra la longitud real. Si quisiésemos decir que el valor es 13.30 deberíamos haber añadido esa cifra significativa, estableciéndose los límites implícitos entre 13.305 y 13.295. Si este no es el propósito no se debería añadir la última cifra significativa.

Cuando se quiere reducir cifras significativas se realiza el redondeo[1]

[1] Las reglas son muy "sencillas": el número a redondear no se cambia si va seguido de uno menor que 5 y sí se cambia (aumentando en uno) si va seguido de un número mayor de 5 ó 5 seguido de otros números distintos de cero. Si el número es 5 solo o seguido de ceros se cambia, aumentando en uno si el anterior es impar, pero no se cambia si es par.

Capítulo 5

Estadística descriptiva

Como se mencionó anteriormente, la estadística descriptiva realiza estimaciones de los estadísticos de la **muestra**. **No** hace inferencias sobre parámetros de la población, simplemente describe y resume la información proporcionada por los datos.

Existen principalmente tres tipos de estadísticos[1] descriptivos:

5.1. Medidas de centralización

También conocidas como medidas de localización:

- **Media aritmética**: actúa como centro geométrico o de gravedad del conjunto de puntos, y se define como:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{5.1}$$

  Existen otras medias tales como la geométrica, o la armónica, que se adaptan mejor a datos transformados (logarítmica o inversamente).

- **Mediana**: es un valor tal que, ordenados en magnitud los datos, el 50 % es menor que ella y el 50 % es mayor. Al ordenar los datos sin agrupar, la mediana es el valor central.

- **Moda**, es el valor más frecuente en los datos.

5.2. Medidas de dispersión

- **Desviación típica o estándar**: es un promedio de las desviaciones de las puntuaciones[2] con respecto a su media[3]. Se elevan al cuadrado las desviaciones para hacerlas positivas y que no se anulen al sumarlas. El cuadrado de la desviación típica se denomina varianza, s^2:

$$s_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \tag{5.2}$$

- **MEDA**: medida de dispersión que se asocia a la mediana (Med). Es la mediana de las desviaciones absolutas. Tiene la ventaja de no verse afectada por datos extremos. Es por lo tanto una medida robusta o resistente:

$$MEDA_{\bar{x}} = mediana|x_i - \bar{x}| \tag{5.3}$$

[1] O estadígrafos, que se definen como: las distintas medidas descriptivas que pueden extraerse de una muestra específica, a fin de ser usado como una estimación sobre un parámetro

[2] Puntuación directa

[3] Puntuación diferencial

- **Rango o Recorrido**: diferencia entre el valor máximo y mínimo.
- **Cuantil**: valor que separa en un tanto por ciento dado los datos: cuartiles(4) 25 %, quintiles(5) 20 %, deciles(10) 10 % y percentiles(100) 1 %. En general, cuantiles de orden p[4] etc.

5.3. Medidas de forma

- **Coeficiente de asimetría**: indica como de simétricos son los datos respecto a la media. Es adimensional:

$$CA_{\bar{x}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{ns^3} \tag{5.4}$$

  Otra medida de asimetría menos utilizada es respecto a la mediana:

$$CA_{mediana} = \frac{\bar{x} - mediana}{s} \tag{5.5}$$

- **Coeficiente de apuntamiento**: indica la forma de la distribución. Se define como el momento de orden 4 respecto a la media dividido por la desviación típica elevada a cuatro. También es denominado índice de curtosis.

$$CAp_{\bar{x}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{ns^4} \tag{5.6}$$

[4] Kendall, 1940

Capítulo 6

Teoría del muestreo

La estadística tiene tres mundos claramente diferenciados: el mundo de la **Población**, el mundo de la **Muestra** y el mundo del **Muestreo**[1]. Debemos saber en todo momento en qué mundo nos encontramos.

Una población es la totalidad de observaciones individuales sobre las cuales se hacen inferencias y que existen en cualquier parte del mundo o al menos dentro del área de muestreo claramente delimitada en el espacio y en el tiempo. Las poblaciones pueden ser finitas, con un número determinado de elementos; o infinitas, cuyo número de elementos no se puede determinar.

De la población extraemos **muestras**, que son conjuntos de observaciones individuales seleccionadas por un procedimiento específico. Las observaciones individuales miden el **carácter** o **variable** en los individuos de la población. En un mismo individuo, o unidad mínima de muestreo, pueden medirse multitud de variables o caracteres.

"Existe la certeza, que al inferir de la muestra a la población cometemos errores o dicho de otro modo, la única certeza que tenemos es la de que cometemos errores" E.Cadima.

Pero lo que hay que hacer es controlar el error, no hay que tener miedo al error, sólo hay que controlarlo. Lo que significa que hay que controlar la precisión de la inferencia, controlando todo el **proceso** para tener una determinada confianza o credibilidad en la inferencia al hacer la "extrapolación".

"El objetivo de la estadística es conocer lo desconocido, que es la población."

[1] No confundir con el proceso de selección de la muestra

Los **tres mundos** de la Estadística son:

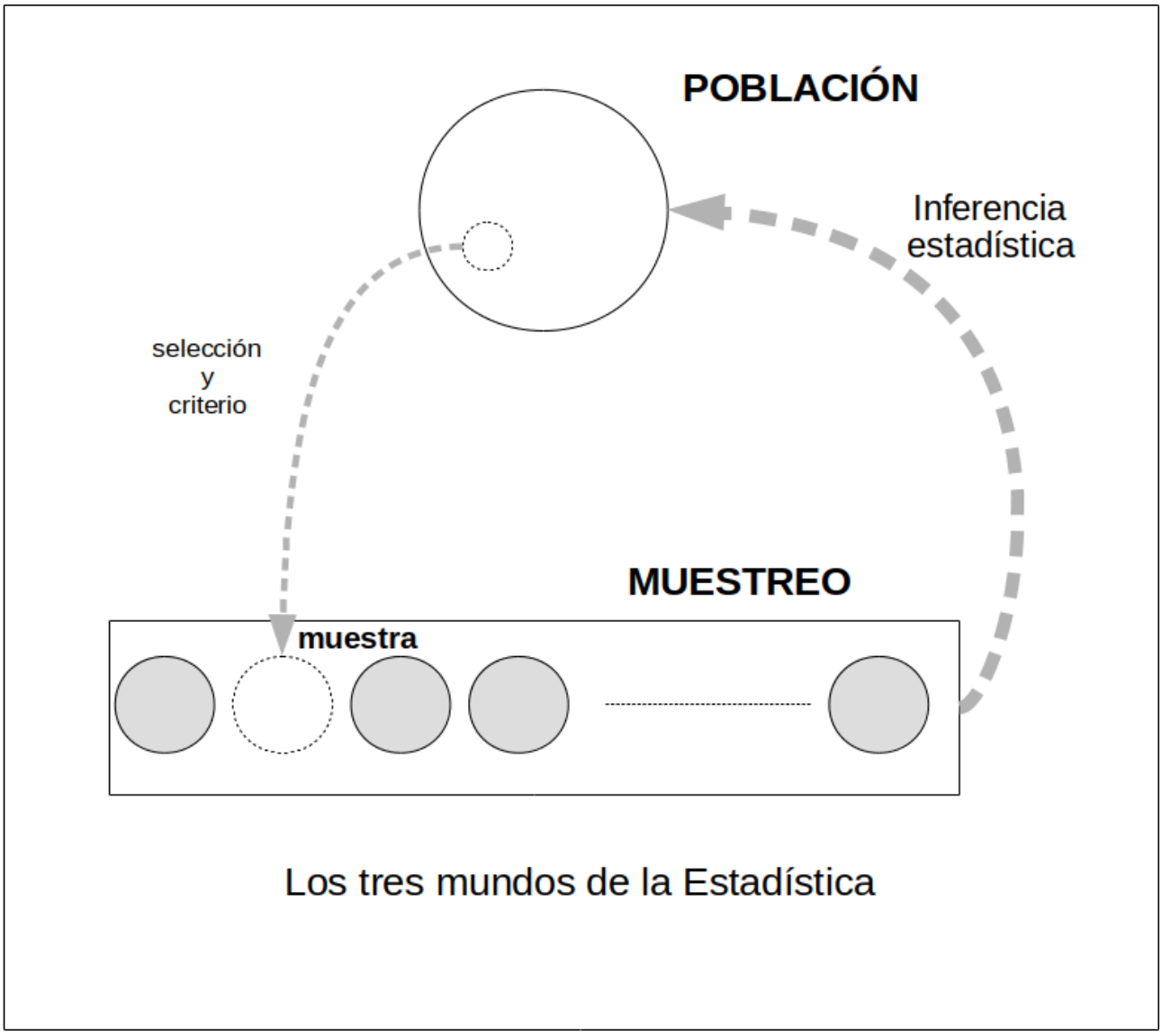


Figura 6.1: Los tres mundos de la Estadística.

6.1. Población, muestra y Muestreo

6.1.1. El mundo de la Población

Normalmente el mundo de la población nos es desconocido, mientras que la muestra siempre es conocida (**Estadística descriptiva**) y por eso necesitamos extraer muestras y tener una teoría que nos permita "conocer" a la población (**Estadística inferencial**).

Las poblaciones pueden ser:

1. Finitas:con número finito de elementos N, que es el tamaño de la población[2].
2. Infinitas: con número infinito de elementos, se denominan poblaciones infinitas.

[2] N, en mayúsculas

La población consta de:

1. Elementos.
2. Características de los elementos (X).

Si hago una lista de los elementos (X_i), lo que hago es una distribución de frecuencias. Si son muchos los elementos, hacemos: gráficos, tablas o histogramas[3] que son diferentes maneras visuales de representar a la población. Así es como empezamos a "resumir" a la población.

Otras maneras de "resumir" los elementos es por medio de los parámetros poblacionales, expresados con letras griegas, como por ej.: media de la población, como μ. Si no, usamos siempre letras mayúsculas. Otro parámetro típico es la varianza σ^2. También podemos escribir la varianza (para esquemas de muestreo) como S^2 o varianza modificada.

$$\mu = \frac{\sum X_i}{N} \qquad \mu\ (Media\ aritmética) \tag{6.1}$$

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N} \qquad \sigma^2\ (Varianza) \tag{6.2}$$

$$\sigma = \sqrt{\sigma^2} \qquad \sigma\ (Desviación\ típica) \tag{6.3}$$

$$S^2 = \frac{\sum (X_i - \mu)^2}{N-1} = \frac{S_{XX}}{N-1} \qquad S^2\ (Varianza\ modificada) \tag{6.4}$$

$$S = \sqrt{S^2} \tag{6.5}$$

$$CV = \frac{\sigma}{\mu} \qquad CV\ (Coeficiente\ de\ variación) \tag{6.6}$$

$$CV^2 = \frac{\sigma^2}{\mu^2} \qquad CV^2\ (Variación\ relativa) \tag{6.7}$$

6.1.2. El mundo de la muestra

Estamos en el mundo de la muestra, de tamaño n. Se puede hacer una distribución de los elementos, gráficos etc., o resumirla con estadísticos:

$$\bar{x} = \frac{\sum x_i}{n} \qquad \bar{x}\ (Media\ aritmética\ muestral) \tag{6.8}$$

$$s^2_{\bar{x}} = \frac{S_{xx}}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n-1} \qquad s^2\ (Varianza\ muestral) \tag{6.9}$$

$$s_{\bar{x}} = \sqrt{s^2_{\bar{x}}} \qquad s\ (Desviación\ típica\ muestral) \tag{6.10}$$

$$cv_{\bar{x}} = \frac{s_{\bar{x}}}{\bar{x}} \qquad cv_{\bar{x}}\ (Coeficiente\ de\ variación) \tag{6.11}$$

$$cv^2_{\bar{x}} = \frac{s^2_{\bar{x}}}{\bar{x}^2} \qquad cv^2_{\bar{x}}\ (Variación\ relativa.\ Cohcran) \tag{6.12}$$

[3] Para variables continúas

Estos dos últimos estadísticos 6.11 y 6.12 sí que nos permiten **comparar** la dispersión entre dos distribuciones.

Cuantil de orden p (cuartiles, quintiles, deciles, percentiles etc.). Si p es 25 % =¿el cuantil es de orden 25 %, será aquel valor que tenga el 25 % de las frecuencias acumuladas. Por ej. la Mediana es el cuantil de orden 50 %.

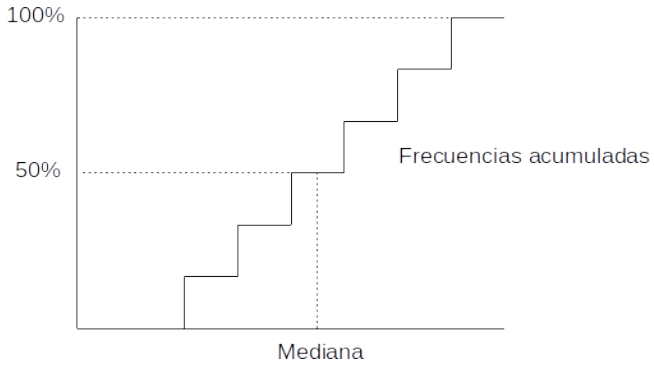


Figura 6.2: Los cuantiles.

Nota sobre la Suma de cuadrados (S_{xx}):

$$S_{xx} = \sum (x_i - \bar{x})^2 \qquad definición \qquad (6.13)$$

$$S_{xx} = \sum x_i (x_i - \bar{x})^2 \qquad también\ teórico \qquad (6.14)$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \qquad (6.15)$$

$$S_{xx} = \sum x_i^2 - n\bar{x}^2 \qquad (6.16)$$

$$S_{xx} = \sum x_i^2 - \bar{x}(\sum x_i) \qquad (6.17)$$

6.1.2.1. Criterio de selección de la muestra

Debemos de tener siempre un criterio y no extraer la muestra de cualquier manera (criterio probabilístico). Si sacamos la muestra sin criterio, se dice que se saca una muestra al azar[4].

Un criterio puede ser, el criterio aleatorio simple: donde cualquier elemento de la población tiene la misma probabilidad de ser seleccionado y formar parte de la muestra. Se dice entonces, que se extrae una muestra aleatoria simple (m.a.s. o a.s.).

Otros tipos de criterios de selección de la muestra pueden ser: estratificado, por conglomerados, multietápico o polietápico, sistemático, proporcional etc.

Cada elemento extraído de la población tiene una probabilidad conocida, si esta probabilidad es igual para todos los elementos, decimos que es un m.a.s. y si otorgamos diferente probabilidad por ej.: para las filas de clase, tenemos que la 1ª fila tiene el 10 %, la 2ª fila el 5 %, etc. sería una muestra aleatorio pero no será simple.

Para tener un criterio, la muestra tiene que ser siempre aleatoria, es decir: tener una probabilidad conocida. Si ésta es igual para todos los elementos es una m.a.s. o también, todas las muestras de tamaño n tienen la misma probabilidad de ser seleccionadas.

[4] ¿Qué diferencia hay entre azar y aleatorio?

Existen dos maneras de extraer las muestras:

- Muestras con reposición
- Muestras sin reposición

Por ejemplo, tenemos una Población P con tamaño $N = 5$, cuyas puntuaciones son: $X_i = \{5, 8, 2, 7, 3\}$ y extraemos una muestra de tamaño $n = 3$ como puede ser el subconjunto $(5, 8, 2)$. En total, tendremos 10 muestras posibles[5] $=>$ $(5, 8, 2)$.....$(2, 7, 3)$

El criterio elegido es un a.s., puede ser expresado como: que cualquiera de las 10 muestras posibles tiene la misma probabilidad de ser sacada. Es decir, que todas las muestras posibles tienen la misma probabilidad; pero para saber esto, tengo que conocer cómo es la población, lo que habitualmente no suele suceder.

El proceso anterior es un ejemplo de a.s. **sin** reposición. Otro sistema diferente es **con** reposición, donde cada elemento puede aparecer más de una vez en la muestra, en cambio sin reposición, cada elemento sólo aparece una vez.

Muestra a.s. posibles	Con reposición	$C'm,n=m^n=N^n$
	Sin reposición	$Cm,n=\frac{m!}{(m-n)!n!}=\binom{N}{n}$

Figura 6.3: Muestras posibles con y sin reposición.

Teóricamente es más fácil estudiar la muestra con reposición que sin reposición, pero esto tiene importancia solamente cuando la diferencia entre población y muestra es pequeña. Si estas diferencias son grandes, no existe prácticamente diferencia entre con o sin reposición. Al cociente n/N se le denomina fracción de muestreo y a $(1 - n/N)$ se le llama fracción de corrección para poblaciones finitas.

[5] ¿Por qué 10 muestras?

6.1.3. El mundo del Muestreo

El **Muestreo** es el conjunto de todas las muestras posibles.

Por ejemplo, de las 10 muestras anteriores, el conjunto de esas 10 muestras posibles es el **Muestreo**. Vamos a introducirnos en la teoría del muestreo que nos enseñará muchas cosas interesantes.

Decíamos que los **tres mundos** de la Estadística eran:

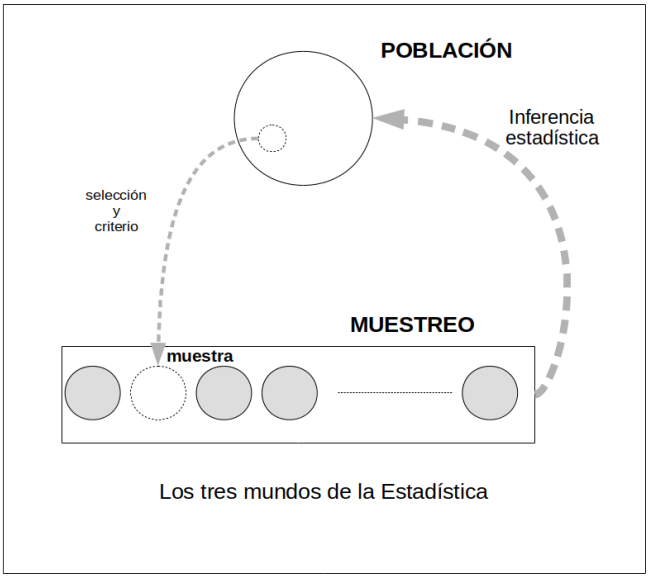


Figura 6.4: Los tres mundos de la Estadística.

Que a su vez son los tres mundos de/para la inferencia estadística.

Lo que conozcamos de la teoría del muestreo(TM), del mundo del muestreo, será lo que nos servirá para hacer **inferencia estadística**. La inferencia se hace fundamentalmente en el mundo del muestreo, aunque se haga ("aparentemente") a partir de la muestra. Se hace "aparentemente desde" la muestra porque conozco el muestreo (el conjunto de las muestras posibles).

En la mano la muestra, pero la mente en el mundo del MUESTREO

Llamamos estadísticos a la: $\bar{x}, s^2, s, mediana$ etc. pero en general los llamaremos $\mathbf{u}$. Extraigo de una muestra el estadístico u, pero podría haber sacado otras muestras posibles también con su estadístico u_i, ya que cada muestra del Muestreo puede tomar un valor del estadístico u_i.

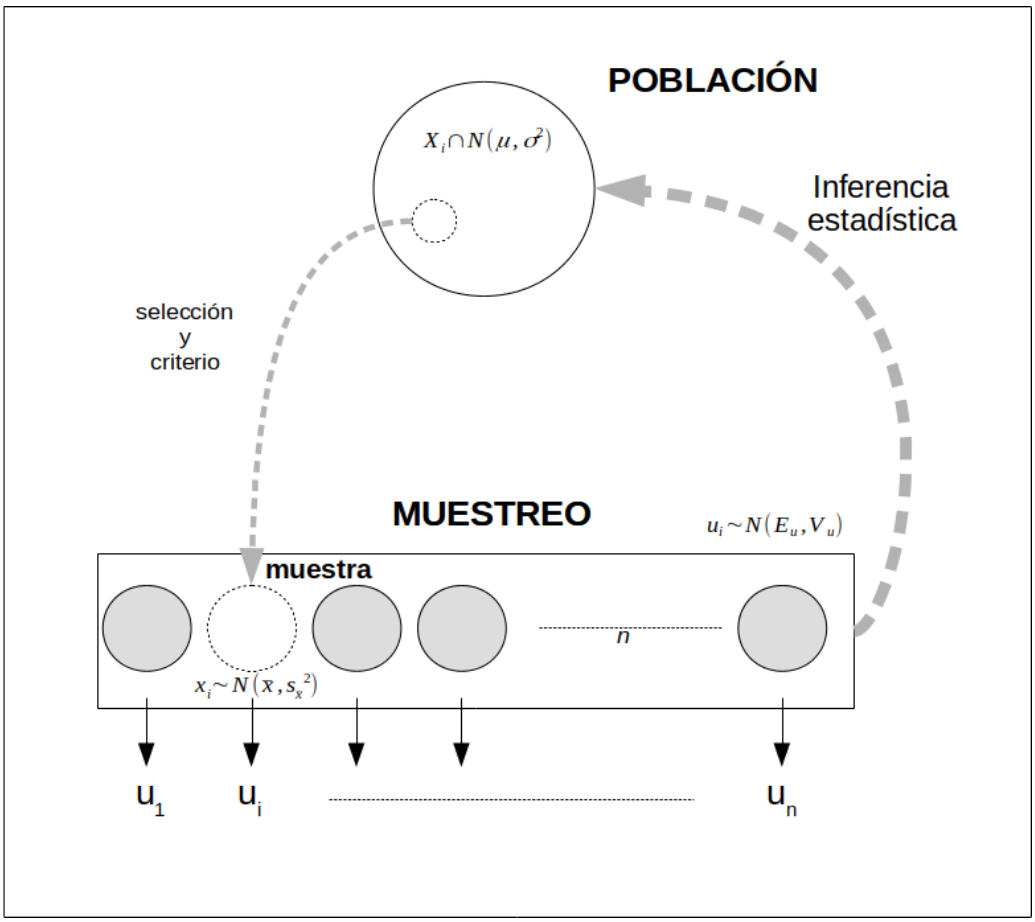


Figura 6.5: Los tres mundos de la Estadística.

Pero sólo conozco un **u** (el que extraigo), pero si tuviera todos los estadísticos u_i que puedo obtener de cada una de las muestras posibles, podría hacer una lista, una gráfica o un histograma, esto es: tener y representar la distribución del estadístico u en el muestreo. Es decir, tener una la lista de todos los valores de u de todas las muestras posibles. Esta distribución de u es fundamental, para la inferencia estadística. En esencia esto es de lo que nos habla la **teoría del muestreo**, que a su vez habla el lenguaje de la probabilidad, la base fundamental de la estadística.

- Las características en la población => **parámetros**
- Las características en la muestra => **estadísticos**
- Las características en el muestreo => **valores esperados** o parámetros (en el muestreo)

6.2. Distribución de los estadísticos

Para estudiar la distribución del estadístico u en el **MUESTREO**, recurro a la teoría de probabilidad que nos da la distribución de u en el Muestreo, es decir la Teoría del Muestreo (TM) nos responde sobre cómo es u:

- La **Esperanza**

  La media de la distribución del estadístico u en el Muestreo μ_u o el más usado $E[u]$, que es **el valor esperado o esperanza del estadístico u** en el Muestreo (la media de la distribución del estadístico **u**).

- La **Varianza**

  Para la varianza: $\sigma_u^2 = V[u]$, **varianza de la distribución del estadístico u** en el Muestreo.

- El **Error**

  La "desviación típica" en el Muestreo: $\sigma_u = \sqrt{\sigma_u^2}$ se llama **error del estadístico u** en el Muestreo que normalmente se representa como: $\mathcal{E}_u = \sqrt{V[u]}$, a veces se denomina simplemente error de u, o sea error es el nombre que damos a la desviación típica de la distribución del estadístico **u** en el Muestreo.

- El **Comportamiento**

  En el Muestreo podemos hacer afirmaciones como cual es: $Prob\{u \leq ?\} = 0{,}99$. Es decir, estamos hablando de que sabemos **cómo se distribuye u**.

Addendum:

- Si de una población de tamaño N saco una muestra de tamaño N:
  Población = muestra = Muestreo
- Si de una población de tamaño N, tomo una muestra de tamaño 1:
  Población = Muestreo

Conclusión:

¡Hay que separar claramente los tres mundos de la Estadística!

¡Nunca debemos confundirlos!

Capítulo 7

Probabilidades

¿Qué es la probabilidad?[1]

La probabilidad, con fines prácticos, la podríamos definir[2] como: "una frecuencia a priori; y la frecuencia, como una probabilidad a posteriori"[3]

La función de probabilidad(P), no es más que la probabilidad de que una variable aleatoria(X) tome un determinado valor: $P(x_i) = p_i$.

Con fines didácticos, nos fijaremos simplemente en el caso de variables aleatorias continuas.

7.1. Función de distribución de probabilidad $F(x)$

La función de distribución nos da la probabilidad de que la variable tome valores iguales o menores que un valor dado de la variable. Que no es más que el resultado de acumular sus probabilidades:[4]:

$$\boldsymbol{P(X \leq x) = F(x)} \tag{7.1}$$

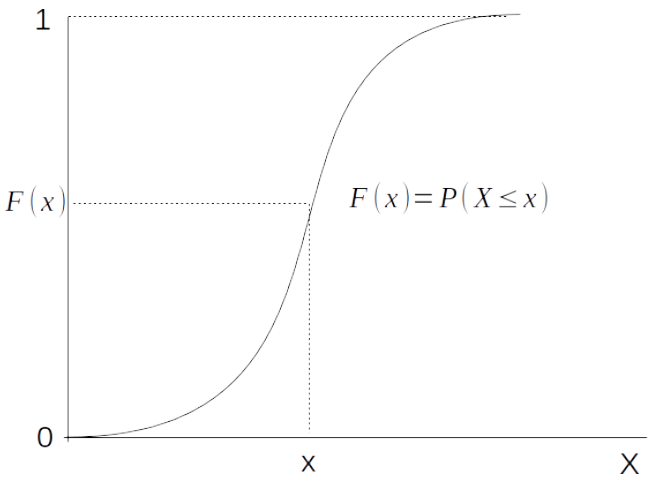


Figura 7.1: Función de distribución.

La función de distribución de probabilidad siempre está entre 0 y 1. Si conozco la función de probabilidad puedo conocer cualquier valor de la variable o sus intervalos (a, b):

$$\boldsymbol{P(a \leq X \leq b) = F(b) - F(a)}^{5} \tag{7.2}$$

[1] ¿Qué diferencias hay entre: probabilidad, frecuencia, posibilidad, (in)certidumbre, verdad, verosimilitud...?
[2] Esta es una definición muy frecuentista que a un Bayesiano le parecerá como que...
[3] ¿Y viceversa?
[4] Para las distribuciones continuas, no existe el valor puntual $\Rightarrow P(X = x) = 0$
[5] $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$

7.2. Función de densidad de probabilidad (pdf) $f(x)$

La función de densidad de probabilidad (pdf) permite obtener **áreas** de probabilidad. La probabilidad de $\boldsymbol{X \leq x}$ se representa como el área que hay entre el origen y **x**.

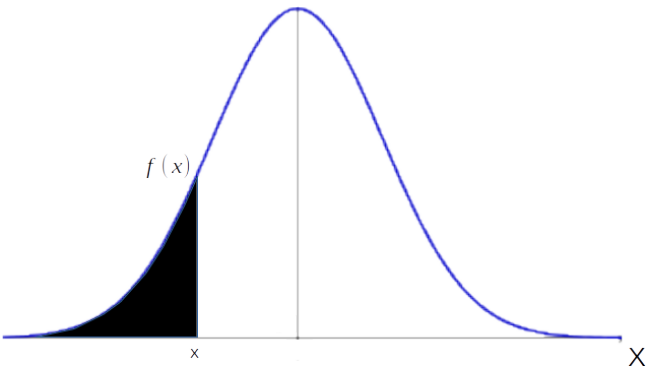


Figura 7.2: Función de densidad.

Para el caso de una variable aleatoria continua X, y si existe una función de distribución $F(X)$, existe a su vez una función de probabilidad $f(x)$ cuya relación será:

$$F(b) - F(a) = P(a \leq X \leq b) = \int_a^b f(x)dx \tag{7.3}$$

En general:

$$F(x) = \int_{-\infty}^{x} f(u)du$$

Es decir que la función de densidad de la probabilidad $f(x)$, no es más que la derivada, cuando sea derivable, de la función de distribución acumulada $F(x) \Longrightarrow f(x) = \dfrac{d}{dx}F(x)$ o dicho de otro modo, la función de distribución no es más que la integral de la función de densidad.[6]

Como propiedades, indicar que $f(x) \geq 0\ \forall x$; y que el área total bajo la función es 1: $\displaystyle\int_{-\infty}^{+\infty} f(x)dx = 1$. En este caso hemos puesto el rango $(-\infty, +\infty)$ como por ejemplo sucede en la función de la distribución normal.

[6] Imaginaos simplemente a una $f(x)$ como una "frecuencia relativa" y a $F(x)$ como una "frecuencia acumulada"

Capítulo 8

Funciones de distribución

8.1. Distribución normal

La función de densidad de probabilidad de la distribución normal $f(x)$ se define por dos parámetros, la media(μ) y la varianza(σ^2).

$$X \cap N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8.1}$$

- Normalización de variables:
  Existen tantas distribuciones normales como medias y varianzas haya. Para calcular probabilidades, y evitar el problema anteriormente citado, se recurre a una distribución especial que se denomina **unitaria**, reducida o **tipificada**, donde la media es cero y la varianza o desviación típica es uno. A esta función se le llama $f(z)$.

  Todas las distribuciones normales se pueden reducir a esta unitaria, y para ello se convierte x a z mediante el proceso denominado tipificación:

  $$z = \frac{x-\mu}{\sigma}$$

  De esta forma $Z \cap N(0,1)$ sólo hace falta tener una tabla de asignación para el cálculo de probabilidades de todas las posibles distribuciones normales.

- Propiedades y características:
  Es simétrica respecto a la media $\Rightarrow$ la media = la mediana = la moda = μ.

  La distribución normal tiene una propiedad muy importante:

  $$x_i \cap N(\mu_i, \sigma_i^2) \Rightarrow \sum x_i \cap N(\sum \mu_i, \sum \sigma_i^2) \qquad i = 1, \ldots, k \quad \textit{independientes}$$

  Si unas determinadas variables x_i se distribuyen normalmente, la suma de esas variables se distribuirá como una normal de media, la suma de las medias y de varianza, la suma de las varianzas.

8.2. Distribución t de Student

Las desviaciones de las medias muestrales respecto a la media paramétrica se distribuyen como una normal, si dividimos estas desviaciones por la desviación típica paramétrica se distribuirán como una normal de media 0 y desviación típica 1 (tipificación).

$$\bar{X} = \frac{(X_1 + \ldots + X_n)}{n} \qquad ; X_1, \ldots, X_n \quad \textit{variables aleatorias independientes} \tag{8.2}$$

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{(n)}} \qquad Z \cap N(0,1)$$

Pero como σ suele ser desconocida y utilizamos como estimador de la desviación típica paramétrica(σ) la desviación típica muestral[1] (S_n), resulta que este cociente ya **no** se distribuye normalmente, sino que toma un forma más aplanada y abierta, ya que la varianza paramétrica "fluctúa" más. Esta es precisamente la denominada distribución t de Student (seudónimo de W.S. Gossett).[2]

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{(n)}}$$

- Propiedades y características:

  Es simétrica alrededor de cero y toma valores desde $(-\infty, +\infty)$ y su forma depende de los grados de libertad ν.

  Al igual que ocurría con la normal, existen muchas distribuciones t, pero en este caso **no** se pueden reducir a una unitaria. Así que existen tablas de la función de distribución de la probabilidad para diferentes grados de libertad, aproximándose a una normal de media 0 y desviación típica 1 cuando los grados de libertad son muy altos ($n \geq 30$).

8.3. Distribución χ^2

Es una distribución de densidad debida a K. Pearson cuyos valores van de 0 a $+\infty$. La distribución es asimétrica y se aproxima asintóticamente al eje horizontal en la cola de la derecha.

No existe una sola distribución χ^2 sino que hay una para cada número de grados de libertad (ν). Así que, χ^2 es función de los grados de libertad ν. La curva empieza siendo en forma de L para $\nu = 1$ pero a medida que los grados de libertad aumentan, se va haciendo más simétrica aproximándose a una distribución normal.

Se define como[3]:

$$Sea\ X = Z_1^2 + ... + Z_\nu^2 \quad ; Z_i\ variables\ aleatorias\ independientes\ tipificadas \Rightarrow Z \cap N(0,1)$$
$$Entonces \quad X \cap \chi_\nu^2$$

- Propiedades y características:

  .-La distribución χ^2 no se puede reducir a una unitaria.

  .-El valor esperado de esta distribución (la media en el muestreo) es los grados de libertad.

$$E[\chi_\nu^2] = \nu \tag{8.3}$$

  .- Si $z \cap N(0,1) \Rightarrow z^2 \cap \chi_{(1)}^2$

  .- $\left.\begin{array}{l} Si\ x \ \cap \ \chi_{(\nu_1)}^2 \\ Si\ y \ \cap \ \chi_{(\nu_2)}^2 \end{array}\right\} \quad x, y\ independientes \Rightarrow (x+y) \ \cap \ \chi_{(\nu_1+\nu_2)}^2)$

  Si tengo muchas variables que se distribuyen como una χ^2, la suma de estas variables se distribuye como un χ^2 con la suma de los grados de libertad.

$$Si \quad \boldsymbol{x_i} \ \cap \ \boldsymbol{\chi^2_{(\nu_i)}} \Rightarrow \sum \boldsymbol{x_i} \ \cap \ \boldsymbol{\chi^2_{(\sum \nu_i)}}$$

Esta última propiedad permite, entre otros tratamientos, realizar el análisis de varianza.

---

[1] Ojo con la notación usada aquí y su denominación. Este concepto se entenderá mejor cuando se vea el capítulo sobre Inferencia

[2] $f(t) = \dfrac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}(1 + t^2/\nu)^{-(\nu+1)/2} \quad ; \Gamma\ es\ la\ función\ gamma$

[3] $f(x;\nu) = \begin{cases} \dfrac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases}$

Capítulo 9

Inferencia estadística

9.1. Introducción

Partimos de la base de que el Método Estadístico se basa en el razonamiento inductivo, o sea que hace proposiciones que van de lo "particular a lo general"[1]. Aplicando el método estadístico, la inferencia estadística nos permite sacar conclusiones sobre la población a partir de una muestra representativa.

El proceso nos permite en general, bajo un determinado marco probabilístico, realizar hipótesis sobre parámetros poblaciones desconocidos y desarrollar formas para comprobar la verosimilitud de tales afirmaciones a través de los estimadores.

9.1.1. Esencia de la Teoría del Muestreo(TM)

En esencia y sencillamente, la TM lo que nos tiene qué decir es: cual es el comportamiento del estadístico **u** en el Muestreo.

Decir cual es el comportamiento, significa saber:

1. Cual es **el valor esperado de u** $\Rightarrow E[u]$
2. Cual es **la varianza de u** $\Rightarrow V[u]$ *o* $Var[u]$
3. Cual es **el error de u** $\Rightarrow Error[u]$
4. Cual es **la distribución de u** $\Rightarrow u \cap ?$

---

1. El valor esperado o esperanza $E[u]$ es la media del estadístico **u** en el muestreo. Media de todos los **u** posibles, de todo el conjunto de medias posibles.
2. $V[u] = s_u^2$
3. La desviación típica del estadístico **u** en el muestreo, es el error estadístico[2]. $\sqrt{V[u]} = Error[u] = s_u$
4. $u \cap$[3] ? ¿Qué tipo de distribución tiene **u**?.

Estas son las cuatro características[4] principales a las que tiene que responder la Teoría del Muestreo (TM). Pero a veces la teoría no puede responder(demostrar) de forma analítica a estas cuatro preguntas y aquí es cuando entran de lleno las diferentes técnicas de remuestreo por simulación numérica[5]. Lo que hacen simplemente es simular (generar) numéricamente por ordenador un conjunto de muestras posibles en el mundo del muestreo.

[1] Realmente esto es una simpleza y el proceso cognitivo es mucho más complejo. Para profundizar en ello, sugiero leer lo que la filosofía de la ciencia nos dice sobre el entorno del método científico: método hipótetico-deductivo, razonamiento deductivo, inductivo, abductivo, analógico etc.

[2] *O* $error[u]$ *o* "$\mathcal{E}[u]$"

[3] *O* $\dot{\cap}$ *o el mas habitual* $\sim$

[4] Que realmente es sólo una(?) desde el punto de vista numérico

[5] Bootstrap, Jackknife, Montecarlo etc.

9.1.2. Dependencias de la TM

La TM (E, V, Error, ∩) depende de:

1. La población, i.e. de los parámetros de la población.
2. Criterio de selección de la muestra.
3. Estadístico elegido.
4. Tamaño de la muestra.

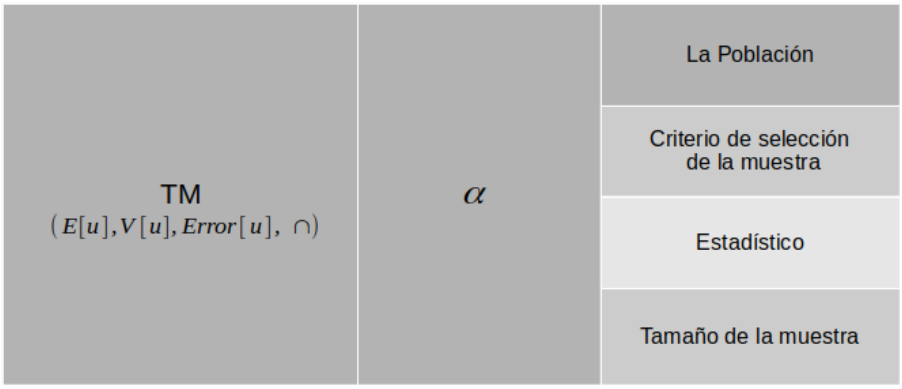


Figura 9.1: Dependencia de la TM.

9.2. Tipos de inferencias

1. **Estimación** de parámetros de la población
2. **Prueba de hipótesis**. Hipótesis que hacemos sobre los parámetros de la población

9.2.1. Estimación de parámetros

9.2.1.1. Cualidades de los estimadores

Para estimar un parámetro, tengo que seleccionar un estadístico. Pero, ¿qué estadístico tengo que seleccionar?. El estadístico seleccionado se llama estimador, por ejemplo: si quiero estimar μ (media de la población), selecciono $\bar{x}$ (media aritmética de la muestra), a $\bar{x}$ se le llama estimador de μ, pero existen otros posibles estimadores de μ, como por ejemplo la mediana.

Pero todo esto no nos dice nada. Lo que se quiere saber es si la estimación es buena o mala. Para solucionar este problema, recurrimos, como siempre, a la TM que nos da resultados concretos sobre los estimadores. Por ejemplo:

Parámetro	Estimador	TM	Cualidad
μ	$\bar{x}$	$E[\bar{x}] = \mu$	**no sesgado**

La TM nos dice que el valor esperado de $\bar{x}$ (media del conjunto de todos las $\bar{x}$ posibles) es μ (la media de la población), y nos dice también que es un estimador **no sesgado** o **centrado**, y que además no depende del criterio de selección de la muestra[6]. Que sea no sesgado o no viciado, es una buena característica para un estimador, quiere decir que con un sólo valor de la muestra, podemos utilizarlo como estimador de la población[7]. Si hiciéramos esto miles de veces, por ejemplo: por simulación; tendríamos la misma probabilidad de que el valor se encontrase tanto a la derecha como a la izquierda de la media poblacional.

[6] Se asume que la media muestral utilizada como estimador se ha calculado de forma coherente con el criterio de selección de la muestra

[7] De aquí viene la confusión de pensar que la media muestral es un estimador directo de la media poblacional, y que no es estimado a través del muestreo (TM)

Con respecto a los estadísticos de dispersión la TM nos dice que:

Parámetro	Estimador	TM	Cualidad
σ^2	s^2	$E[s^2] = \sigma^2$	**no sesgado**

La TM dice que sólo es no sesgado si el criterio es un m.a.s[8] sin reposición

Parámetro	Estimador	TM	Cualidad
σ	s	$E[s] \neq \sigma$	**es sesgado**

La TM dice que el valor esperado de s en el muestreo **no es** la desviación típica poblacional. El estimador es sesgado o no centrado; y el sesgo es grave para un tamaño de muestra menor que 30, si $n > 30$ el sesgo es muy pequeño. Esta cualidad es llamada **consistencia** de un estimador.

La TM dice que: $E[s] \neq \sigma$, pero si : $n \to \infty \Rightarrow E[s] = \sigma$; a medida que el tamaño muestral se incrementa, s es mejor estimador de σ. Se dice entonces que s es un estimador consistente, esta cualidad es más importante que la de sesgado.

Decíamos que si queremos estimar μ podíamos utilizar tanto la media como la mediana($\tilde{x}$). Pero, ¿con cuál de las dos nos quedaríamos?.

Parámetro	Estimador	TM	Cualidad
μ	$\bar{x}$	$E[\bar{x}] = \mu$	no sesgado
μ	$\tilde{x}$	$E[\tilde{x}] = \mu$	no sesgado

Pero si miramos cómo son las varianzas en el muestreo de estos dos estadísticos, vemos que: $V[\bar{x}] < V[\tilde{x}]$. Esto quiere decir que la media aritmética como estimador de μ es más **eficiente** o **preciso** que la mediana. Que sea más eficiente (formalmente en estadística) significa que la varianza en el muestreo es menor.

Con estos ejemplos se han visto tres[9] de las principales cualidades/propiedades de los estimadores puntuales de parámetros de la población, que son:

1. **Sesgo**[10]
2. **Consistencia**
3. **Eficiencia**[11]

[8] Muestreo aleatorio simple

[9] Además en este contexto son importantes los conceptos de:

- Error cuadrático medio (ECM) $\Rightarrow ECM(\vartheta) = E[(\hat{\vartheta} - \vartheta)^2]$; $\hat{\vartheta}$ *es el estimador* y;
- el concepto de Robustez de un estimador a las alteraciones del modelo

[10]

- Si es centrado decimos que: $E[\hat{\vartheta}] = \vartheta$.
- Si no es centrado $\Rightarrow sesgo(\hat{\vartheta}) = E(\hat{\vartheta}) - \vartheta$

[11] O **Precisión**

9.2.1.2. Estimaciones puntuales de los parámetros

Aquí veremos sólo las estimaciones puntuales y por intervalos, pero quedan dos métodos de estimación muy importantes a considerar: la estimación de máxima verosimilitud y la estimación bayesiana.

Distribución de la media en el muestreo:

1.

$$E[\bar{x}] = \mu \tag{9.1}$$

μ = media de la población
$\bar{x}$ = media de la muestra
E = media del muestreo

Se ve cómo están relacionados los tres mundos de la estadística:

"La media en el muestreo de la media de la muestra es la media de la población"

2.

$$V[\bar{x}] = \frac{\sigma^2}{n} \tag{9.2}$$

Esto es cierto, cuando población es infinita o cuándo el muestreo es con reposición.

Si N es finita y el criterio del muestreo es un aleatorio simple(a.s.) sin reposición, entonces[12]:

$$V[\bar{x}] = \left(1 - \frac{n}{N}\right) \frac{S^2}{n} \tag{9.3}$$

$(N-1)\ S^2 = N\sigma^2$; S^2 es la varianza modificada. Sabiendo una varianza se puede deducir la otra. Normalmente como la corrección para poblaciones finitas $(1-\frac{n}{N})$ es pequeña, salvo en casos excepcionales, se suele usar el conocido:

$$V[\bar{x}] = \frac{s^2}{n} \tag{9.4}$$

donde s^2 es la varianza de la muestra.

Como la población suele ser muy grande: $\left(1 - \frac{n}{N}\right) * \frac{S^2}{n} \cong \frac{\sigma^2}{n}$

La varianza en el muestreo de $\bar{x}$ es igual a la varianza de la población dividido por el tamaño de la muestra. $V[\bar{x}]$ es más pequeño a medida que el tamaño de la muestra aumenta, es decir: la distribución estará más concentrada entorno a la verdadera media de la población, luego la estimación es más precisa[13].

3.

$$Error[\bar{x}] = \sqrt{V[\bar{x}]} = \frac{\sigma}{\sqrt{n}} \tag{9.5}$$

La desviación típica en el muestreo ($Error[\bar{x}]$), cuánto mayor sea el tamaño de la muestra, menor será su error[14].

[12] En algunos libros aparece: $V[\bar{x}] = \frac{N-n}{N-1} * \frac{\sigma^2}{n}$ para N finito y a.s. sin reposición

[13] Y más eficiente. ¡Ojo! hasta un determinado punto.

[14] Repito, hasta un determinado punto. Es una función inversa y asintótica. Por ello es útil, entre otras cosas, para el cálculo de un tamaño de muestra adecuado u óptimo en función del error que deseemos.

Si desconocemos σ^2, podremos utilizar la varianza muestral s^2 para calcular el ($\underline{error}[\bar{x}]$). Cuando hacemos esto se denomina $var[\bar{x}]$ en vez de $V[\bar{x}]$.

Si

$$var[\bar{x}] = \sqrt{\frac{s^2}{n}} \quad \Rightarrow \quad error[\bar{x}] = \frac{s}{\sqrt{n}} \tag{9.6}$$

Cuando usamos esta aproximación ($\frac{s^2}{n}$), la $\bar{x}$ se distribuye como una distribución t de Student. En la práctica es lo que se utiliza, pero ... si $n > 30$ la t-Student se comporta como una distribución Normal[15].

4.

$$\bar{x} \cap N(E, V) \tag{9.7}$$

la media de la muestra tiene una distribución normal con media, el valor esperado de ($E[\bar{x}]$) y varianza $V[\bar{x}]$. Entonces $\bar{x} \dot{\cap} N(E, V) \Rightarrow N(\mu, \frac{\sigma^2}{n})$ $\quad ; donde \dot{\cap}\ o\ \sim\ \rightarrow$ "*aproximadamente*"

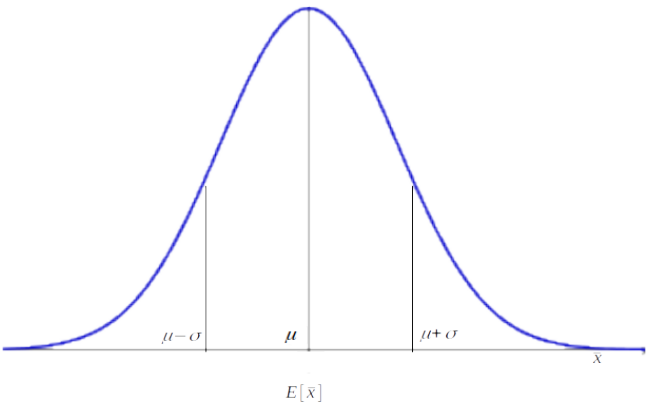


Figura 9.2: TM sobre la media.

Es decir que la TM nos dice también, que la media muestral en el mundo del muestreo se distribuye como una normal de media la esperanza y desviación típica, el error típico. Es fundamental conocer la distribución de los estadísticos en el muestreo.

Para cualquier población, la aproximación a la Normal es mejor cuando n es grande ($n > 30$) y cuando la población es simétrica. Deja de ser aproximada para ser exacta, cuando la Población es Normal.

[15] Como info suplementaria se recomienda ver el Teorema del límite central o central del límite

Distribución de la s^2 en el muestreo

1.

$$E[s^2] = \sigma^2 \tag{9.8}$$

2. Por ahora vamos a dejar de ver como son: $V[s^2]$ *y el* $Error[s^2]$, siendo de mayor más interés ver cómo es la distribución de s^2:

$$s^2 = \frac{S_{xx}}{(n-1)} \quad ; \quad s^2 \textit{ varianza de la muestra}$$

$$\left(\frac{S_{xx}}{\sigma^2}\right) \cap \chi^2_{(n-1)} \Rightarrow E\left[\frac{S_{xx}}{\sigma^2}\right] = n-1 \rightarrow E\left[\frac{S_{xx}}{(n-1)}\right] = \sigma^2 \tag{9.9}$$

Por eso decimos que[16]: $E[s^2] = \sigma^2$
Nota: igual veis un desarrollo análogo, ya que realmente $E[s^2] = \sigma^2$ es una "aproximación"[17]:

$$E[\frac{1}{n}\sum(x_i - \bar{x})^2] = \frac{(n-1)}{n}\sigma^2$$

$$\frac{(n-1)s^2}{\sigma^2} = \frac{1}{\sigma^2}\sum(x_i - \bar{x})^2 \sim \chi^2_{(n-1)} \tag{9.10}$$

En resumen, si la variable X tiene una distribución normal, la distribución de $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

Por lo tanto, entre otras propiedades, la varianza muestral no es simétrica, tiene asimetría positiva.

9.2.1.3. Intervalos de confianza de la media y la varianza

Al estimar los parámetros obteníamos un valor puntual y como nos dice la teoría de muestreo: la distribución de los mismos. Esta distribución nos permite establecer la precisión de nuestra estimación gracias al cálculo de los intervalos de confianza[18] (C) de nuestros estimadores.

La teoría del muestreo nos da los datos necesarios para calcular los intervalos de confianza de los estimadores, diciéndonos que la media se distribuye como una normal de media μ y desviación típica, el error de la media. Si tipificamos, se distribuirá como una normal de media cero y desviación típica 1 ($Z \sim N(0,1)$).

Como vimos en la ecuación (7.3), definíamos la función de densidad de probabilidad (pdf) para un intervalo dado como:

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{9.11}$$

Que nos permite calcular una densidad entre dos valores dados (a, b) y viceversa, a partir de una determinada área establecer cuales son los valores que acotan tal densidad.

[16] Para entender $E\left[\frac{S_{xx}}{\sigma^2}\right] = n-1$ ver ecuación 8.3

[17] Realmente tiene trampa esta definición, ¿sabéis cual es? ¿$E[s^2]$ es sesgado o insesgado?

[18] O credibilidad. ¿Qué diferencias hay?

Para una distribución normal tipificada, las densidades correspondientes a 1, 2 y 3 desviaciones(σ), comprenden aproximadamente el 68.3 %, 95.4 % y 99.7 % respectivamente de la densidad. A estas áreas o densidades les podremos denominar **Confianzas** y al resto de la densidad[19] le llamaremos α o **nivel de significación**: $\alpha = 1 - C$, que tiene una enorme importancia en la estadística y que se verá en mayor profundidad en la sección siguiente sobre el Contraste o Prueba de hipótesis.

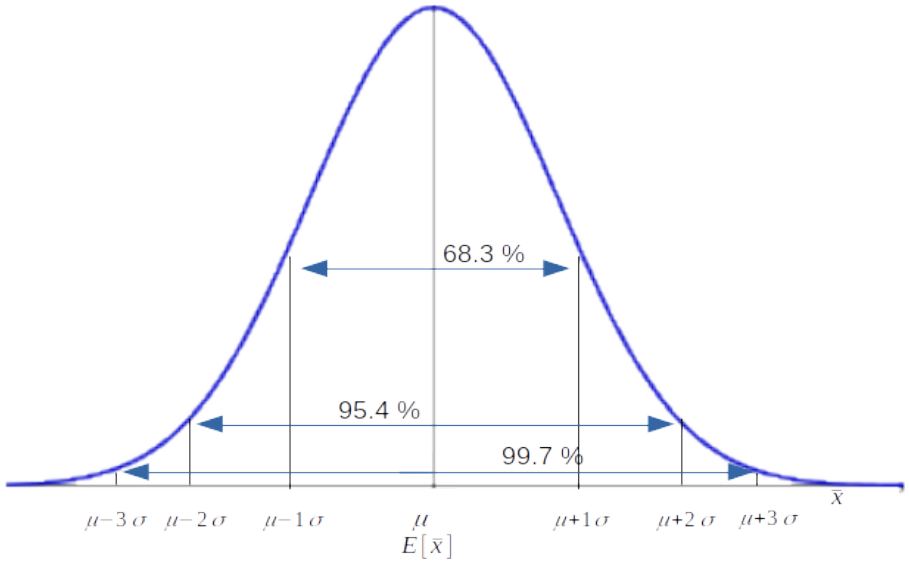


Figura 9.3: Densidades de la distribución normal.

- **Para la media**

  En el caso de la desviación normal tipificada (Z), los intervalos de confianza se construirán habitualmente[20] de la siguiente manera:

$$P(-z \leq Z \leq +z) = 1 - \alpha = C \tag{9.12}$$
$$P\left(-z \leq \frac{\bar{x} - \mu}{s/\sqrt{n}} \leq +z\right) = 1 - \alpha = C$$
$$P\left(-z\frac{s}{\sqrt{n}} \leq \bar{x} - \mu \leq +z\frac{s}{\sqrt{n}}\right) = 1 - \alpha = C$$
$$P\left(-z\frac{s}{\sqrt{n}} \leq \mu - \bar{x} \leq +z\frac{s}{\sqrt{n}}\right) = 1 - \alpha = C$$
$$P\left(\bar{x} - z\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z\frac{s}{\sqrt{n}}\right) = 1 - \alpha = C$$

$$P(L_1 \leq \mu \leq L_2) = 1 - \alpha = C \tag{9.13}$$

$$L_1 = \bar{x} - z\frac{s}{\sqrt{n}} \qquad y \qquad L_2 = \bar{x} + z\frac{s}{\sqrt{n}}$$

L_1 y L_2, son los límites del intervalo de confianza.

[19] $\int_{-\infty}^{+\infty} f(x)dx = 1$

[20] ¿Por qué habitualmente?

Si el tamaño de las muestras es pequeño y no conocemos la desviación típica paramétrica, entonces debemos utilizar la distribución **t**, para usar la desviación típica del muestreo $s_{\bar{x}}$[21]:

$$Var[\bar{x}] = \frac{s^2}{n} \quad ; \quad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Donde s^2 y s son la varianza y desviación típica de la muestra. Siendo los límites del intervalo de confianza:

$$L_1 = \bar{x} - ts_{\bar{x}} \quad ; \quad L_2 = \bar{x} + ts_{\bar{x}} \tag{9.14}$$

Debido a que:

$$\begin{aligned} C &= P(-t \leq t_{n-1} \leq +t) \\ C &= P(\bar{x} - ts_{\bar{x}} \leq \mu \leq \bar{x} + ts_{\bar{x}}) \end{aligned} \tag{9.15}$$

Si se supone que los límites son simétricos este intervalo da la mínima amplitud, si se hacen no simétricos la amplitud del intervalo de confianza es mayor para esa confianza. Así que la teoría de muestreo aconseja que sean simétricos.

- **Para la varianza** La teoría de muestreo en el caso de la varianza nos informa en el muestreo sobre la esperanza, la varianza y el error, así como su distribución. Por lo tanto, se puede calcular perfectamente el intervalo de confianza de la varianza.

  Según la teoría del muestreo:

$$\frac{S_{xx}}{\sigma^2} \cap \chi^2_{n-1}; \qquad Error\left(\frac{S_{xx}}{\sigma^2}\right) = \sqrt{2(n-1)}$$

$$C = P\left(L_1 \leq \frac{S_{xx}}{\sigma^2} \leq L_2\right); \qquad Como\ \frac{S_{xx}}{\sigma^2} \cap \chi^2_{n-1} \Rightarrow E\left(\frac{S_{xx}}{\sigma^2}\right) = n-1$$

$$L_1 = (n-1) - \chi^2_{n-1(\alpha/2)} Error\left(\frac{S_{xx}}{\sigma^2}\right)$$

$$L_2 = (n-1) + \chi^2_{n-1(\alpha/2)} Error\left(\frac{S_{xx}}{\sigma^2}\right)$$

Entonces:

$$\frac{S_{xx}}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \cap \chi^2_{n-1}$$

$$P(-\chi^2_{n-1} \leq \chi^2 \leq +\chi^2_{n-1})$$

$$P\left(\chi^2_{n-1(1-\alpha/2)} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{n-1(\alpha/2)}\right) = C$$

$$P\left(\frac{(n-1)s^2}{\chi^2_{n-1(\alpha/2)}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{n-1(1-\alpha/2)}}\right) = C$$

$$Si \ \ (n-1)s^2 = \sum y^2 \quad \Rightarrow \quad \boldsymbol{P\left(\frac{\sum y^2}{\chi^2_{n-1(\alpha/2)}} \leq \sigma^2 \leq \frac{\sum y^2}{\chi^2_{n-1(1-\alpha/2)}}\right) = C} \tag{9.16}$$

[21] Mejor notación sería: $error[\bar{x}]$ pero se suele usar también esa notación.

9.2.2. Contraste o Prueba de hipótesis

Prácticamente cualquier suposición puede transformarse en una hipótesis que se puede probar. La hipótesis es lo que queremos probar y en cambio una suposición, al igual que una asunción, no necesita ser probada.

Las hipótesis se hacen siempre sobre el mundo de la Población.

Esto no hay que olvidarlo nunca. La hipótesis se denomina tradicionalmente como H_0, por ej.: podría ser que la media de mi población es igual a 4 $\Rightarrow$ $H_0 : \mu = 4$; o, que la varianza de mi población es igual a 5 $\Rightarrow$ $H_0 : \sigma^2 = 5$; o que tengo una población (1) que tiene como media μ_1 y otra población(2) que tiene como media μ_2 y por lo tanto podríamos plantear una hipótesis del tipo $\Rightarrow$ $H_0 : \mu_1 = \mu_2$

Siempre sobre la Población, esto es lo que tenemos que probar. Para aceptar (no rechazar)[22] o no(rechazar) una hipótesis tenemos que seguir siempre el mismo proceso, que este consta de 5 partes:

- Proceso de realización

  1. Formular la hipótesis $\Rightarrow$ H_0 / H_1.

  2. Definir el nivel de confianza[23].
     Definir un criterio de selección y el tamaño de la muestra.
     Definir el estadístico que se utilizará.

  3. Definir el criterio de decisión sobre rechazar la H_0 o no hacerlo. Esto se hace en base a la TM.

  4. Seleccionar la muestra y calcular el valor del estadístico de contraste.

  5. Toma de decisión: rechazar o no rechazar H_0.

		HIPÓTESIS	
		VERDADERA	FALSA
DECISIÓN	RECHAZAMOS	Incorrecta	Correcta
	NO RECHAZAMOS	Correcta	Incorrecta

Figura 9.4: Cuadro de decisión.

Otra decisión podría ser: ni rechazar, ni no-rechazar[24]. En estadística, nuestras decisiones son tomadas con probabilidad, nunca tenemos la certeza (casi nunca). Lo que tenemos es: la probabilidad de estar en lo correcto o en lo incorrecto (**no**, verdadero o falso). Por eso en estadística, hacemos otro cuadro hablando de probabilidades.

[22] ¿Qué diferencia hay entre aceptar y no rechazar?
[23] O nivel de significación(α) e idealmente β pero ... **Siempre a priori**. Esto es crucial!
[24] ¿Mande?! ;)

- Probabilidades asociadas

		HIPÓTESIS	
		VERDADERA	FALSA
DECISIÓN	RECHAZAMOS	α Error tipo I	$1-\beta$
	NO RECHAZAMOS	$1-\alpha$	β Error tipo II

Figura 9.5: Cuadro de probabilidades sobre la decisión.

Donde $\boldsymbol{\alpha}$ es: **la probabilidad de rechazar una hipótesis verdadera** y $\boldsymbol{\beta}$ es: **la probabilidad de no rechazar una hipótesis falsa**. También son conocidos como error de tipo I (α) y tipo II[25] (β) respectivamente.

$(1-\alpha)$ es la confianza(C)[26] que se puede interpretar como: **tener una confianza del 95 % de no rechazar la hipótesis cuando es verdadera**. En prueba de hipótesis, en vez de la confianza, lo habitual es hablar más del error; y a $\boldsymbol{\alpha}$ se le llama nivel de significación. Por lo tanto, tener un $\boldsymbol{\alpha}$ de 5 % implica tener una confianza del 95 %, es decir, significa tener una probabilidad del 5 % de rechazar la hipótesis cuando esta se verdadera.

$$C = 1-\alpha = P(no\ rechazar\ /\ H_0\ verdadera)$$
$$\alpha = P(rechazar\ /\ H_0\ verdadera)$$

Ejemplo[27]:

1. $H_0 : \mu = 4$, con σ^2 conocida $\Rightarrow \sigma^2 = 12{,}1$
2. nivel de confianza y significación, 95 % y 5 % respectivamente.
   criterio de selección de la muestra: m.a.s sin reposición, de tamaño $n = 10$.
   estadístico: $\bar{x}$
3. $$C = P(\bar{x} - z * s_{\bar{x}} \leq \mu \leq \bar{x} + z * s_{\bar{x}})$$
   $$C = P(\mu - z * s_{\bar{x}} \leq \bar{x} \leq \mu + z * s_{\bar{x}})$$

   Esta última expresión es la que se usa ya que es la probabilidad de no rechazar H_0 cuando es verdadera y cuando H_0 es verdadera $\mu = 4$.

   Ahora se calculan los valores que se necesitan, por ejemplo el error de la media muestral.

   $$s_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{12{,}1}{10} = 1{,}21 \quad \Rightarrow \quad s_{\bar{x}} = 1{,}1$$

   Como sabemos que:

   $$C = P(-z \leq Z \leq +z) \quad \Rightarrow \quad 0{,}95 = P(-1{,}96 \leq Z \leq +1{,}96) \qquad ; z = 1{,}96$$
   $$0{,}95 = P(4 - 1{,}96 * 1{,}1 \leq \bar{x} \leq 4 + 1{,}96 * 1{,}1)$$

[25] que realmente es el más grave.
[26] $(\boldsymbol{1-\beta})$ es la Potencia del contraste.
[27] De nuevo, ojo con la notación!

Quedando finalmente:

$$0{,}95 = P(1{,}84 \leq \bar{x} \leq 6{,}16) \quad / \quad H_0 : \mu = 4 \qquad \Rightarrow H_0 : es\ verdadera$$

Si a partir de la la muestra se obtiene una media comprendida entre 1,84 y 6,16, se tiene un 95 % de confianza de no rechazar H_0, y esta es la regla de decisión. Esta decisión tiene una probabilidad del 95 % de no rechazarse y un 5 % de rechazarse.

A la región de rechazo se le denomina también **región crítica** y a la de no rechazo, **región de aceptación**.

4. Selección de la muestra y cálculo de la media.

5. Si la muestra cae en la región de no rechazo entonces H_0 no es rechazada, con un nivel de significación α (H_0 no es rechazada con un nivel de significación del 5 %). Si cae en la región de rechazo, ahí sí que se puede decir que es significativamente diferente.

Capítulo 10

Regresión lineal simple. Efectos fijos (I)

La regresión lineal simple, múltiple, el análisis de varianza (ANOVA), el de covarianza (ANCOVA) etc. surgen del Modelo Lineal General (GLM) y por reducción de los modelos aditivos generales (GAM).

Modelo Lineal General:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

La regresión se podría definir como: el estudio de la relación que existe entre las variables explicativas y las variables respuestas. En el caso de la regresión lineal, esta es en base a un modelo lineal[1] y en el caso más simple sería a través de una sola variable explicativa y una variable respuesta, mal llamadas independiente y dependiente respectivamente.

La regresión (relación) y el ANOVA (diferenciación) son caras de la misma moneda. Son las técnicas de análisis mayoritariamente usadas en estadística, sobre todo la regresión y la correlación.

Para su estudio [2] [3] [4] [5]

10.1. En la Población

En la regresión lineal, las diferentes posibilidades de y son y_i, para un x_i hay diferentes y_i

Las medias de los diferentes y_i para cada x están sobre una recta. $Y = media\ de\ las\ y\ para\ cada\ x$ definida por la ecuación:

$$Y = A + BX$$

para cada valor de x la media de todos los y_i están sobre una recta, con dos suposiciones principales:[6]

1. Para cada x_i las y_i están distribuidas normalmente

$$Y \cap N(\mu, \sigma^2)$$

$$\mu = media\ de\ las\ y\ para\ cada\ x \Rightarrow \mu = Y$$

$$\boxed{Y \cap N(A + BX,\ \sigma^2)} \tag{10.1}$$

2. Las varianzas de todos los y_i son constantes

$$\boxed{\sigma^2 = cte} \tag{10.2}$$

[1] Lineal se refiere a la ecuación, o sea que los parámetros tienen que ser lineales. x puede ser x^2, $\ln x$ o $\sqrt{x}$.
[2] Seber, G. A. F. (1977), Linear Regression Analysis, Wiley, New York
[3] Seber, G. A. F. & Lee, A. J. (1998),Linear Regression Analysis, Wiley
[4] Draper, N. R. & Smith, H. (1998),Applied Regression Analysis, third edn,Wiley
[5] Faraway, J. J. (2005),Linear Models with R, Chapman & Hall/CRC
[6] Son asunciones. Por lo tanto, no se tiene que probar nada y simplemente se aceptan.

Como comentario, decir que es una propuesta robusta[7], i.e. que aunque las suposiciones del modelo no se cumplan estrictamente, la conclusión a la que se llega es la misma.

Pero esto mismo es importante verlo bajo el "prisma"de los residuos. Otra aproximación al modelo podría ser: que para cada x_i, valor fijo de X, se cumple la ecuación

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

donde β_0 y β_1 son constantes desconocidas.

Las hipótesis básicas del modelo son:

1. **Independencia en los residuos**. Sin correlación (auto-correlación) de los residuos $Cor(e_i, e_j) = 0$. Cualquier par de errores e_i y e_j son independientes.
2. **Esperanza nula** de los residuos es: $E(e_i) = 0$
3. **Varianza constante** de los residuos (homocedasticidad): $Var(e_i) = \sigma^2$
4. **Normalidad** de los residuos: $e_i \sim N(0, \sigma^2)$

De lo que se deduce:

- Cada valor y_i de la variable aleatoria Y tiene una distribución

$$\boxed{(Y \mid X = x_i) \approx N(\beta_0 + \beta_1 x_i,\ \sigma^2)}$$

- Las observaciones y_i de la variable Y son independientes.

Si las hipótesis del modelo son ciertas, gráficamente se tiene que:

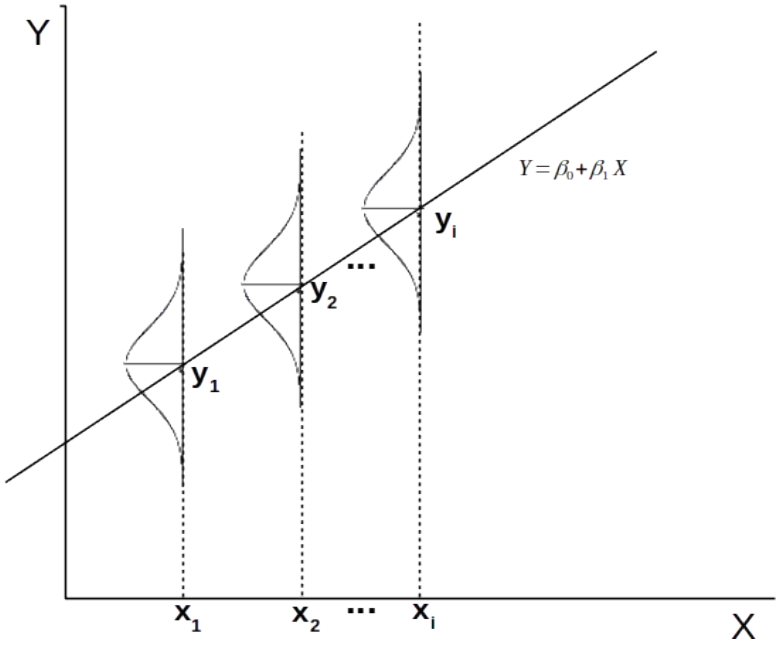


Figura 10.1: Modelo gráfico de la recta de regresión.

[7] Robusta para la construcción de la recta, pero sí es más sensible para los contrastes de hipótesis asociados.

10.2. En la muestra

Criterio de selección

Las X **no** tienen criterio probabilístico, se decide arbitrariamente (por el investigador) cuánto vale. Por eso, a veces se le llama "media variable o semi-variable". Por lo tanto, X **NO** es una variable aleatoria, no tiene una probabilidad asociada. Su selección es arbitraria, tiene cierta libertad, pero una vez seleccionada las x_i, las y_i **SÍ** que se seleccionan de forma aleatoria.

El caso más sencillo, para cada x, es que la muestra tenga un tamaño 1. Lo que quiere decir: que para cada x sólo se escoge una y de todas las posibles y_s.

Generalmente, el criterio de selección suele ser que: para cada valor arbitrario de x_i elijo un y_i aleatorio simple (a.s.), con lo cual la muestra total tiene n pares de valores.

A la X se le llamaba variable independiente, pero es un nombre antiguo y lo correcto, en este contexto, es llamarle: variable auxiliar, variable regresora o en general, variable explicativa.

10.3. En el Muestreo

En el mundo del Muestreo (no confundir con el hecho de muestrear) es fundamental saber qué inferencia se pretende realizar; i.e. responder a: ¿Cuál es el objetivo que se quiere conocer?

En la Población para cada valor de X, la Y (media de y para cada x) está sobre una recta: es una recta. La distribución de Y para cada X es una normal con $\mu = A + BX$ y con la misma varianza σ^2. También lo podemos expresar como $y \cap (Y,\ \sigma^2)$. Es decir, $\boxed{Y = A + BX}$

Para resolver la ecuación anterior, ¿Cuántos parámetros se tienen que calcular? Se tiene A y B y $\cdots$ σ^2 (común para todos los x) e....$Y_{(media)}$ que también lo desconozco, pero si se conoce A y B ya se puede calcular.

Con lo cual el objetivo que se pretende es: que a partir de la muestra, **estimar** A, B y σ^2 para posteriormente calcular la Y. Es a partir de este punto cuando ya se pueden hacer pruebas de hipótesis.

Estimación de $A, B,\ \sigma^2$ e Y

Observación		Pretensión	Estimación
x	y	Y	$\hat{Y}$
x_1	y_1	$A + B\,x_1$	$a + b\,x_1$
x_2	y_2	$A + B\,x_2$	$a + b\,x_2$
⋮	⋮	⋮	⋮
x_i	y_i	$A + B\,x_i$	$a + b\,x_i$
⋮	⋮	⋮	⋮
x_n	y_n	$A + B\,x_n$	$a + b\,x_n$

Figura 10.2: Tabla de regresión.

Por lo tanto, se tendrán que estimar estos tres parámetros: a, b *e* $\hat{Y}$ que son los estimadores puntuales de A, B *e* Y respectivamente. Obviamente, los más importantes son los dos primeros (a y b).

10.3.1. Métodos de estimación de los parámetros

Método de mínimos cuadrados

Este método de estimación es **uno** de los más habitualmente utilizados. Concebido para el cálculo de la estimación puntual de a y b, basado en reducir (minimizar) la distancia al cuadrado entre el y de la muestra menos el estimador ($\hat{Y}$) de y:

$$\boxed{\sum_{i=1}^{n}(y_i - \hat{Y}_i)^2 = S^2}$$

esa diferencia de valores se llama *residuo* y por tanto S^2 es la suma de los cuadrados de los residuos. Que también se puede expresar como:

$$\boxed{S^2 = \sum[y_i - (a + bx_i)]^2}$$

Para calcular a y b, simplemente se calculan las derivadas parciales de la suma de cuadrados de los residuos con respecto a cada estimador para que estas derivadas sean igual a cero, ya que lo que se busca es minimizar esa función para cada parámetro.

Este método sencillo de ajuste se utiliza en muchas ocasiones en estadística, ya que los estimadores que genera son:

- No sesgados
- Eficientes(Varianza pequeña en el Muestreo)
- Consistentes. Si sesgados, aumentando n tiende a ser insesgado

En resumen, para la estimación de los parámetros lo que se busca es minimizar la suma de cuadrados de los residuos $(y_i - \hat{Y}_i)^2$ en el ajuste; que se denomina también $S = S^2_{residual}$. Otra forma de verlo es: que lo que se busca es reducir la diferencia entre lo que se observa (y_i) y lo que se estima ($\hat{Y}$) o espera obtener a partir del ajuste.

Cálculo de los estimadores por mínimos cuadrados

Para el cálculo de los parámetros de la regresión lo primero que se necesita es recordar cómo se calcula la suma de cuadrados para x e y y la suma de cuadrados del producto $x \cdot y$ (S_{xy}):

$$\boxed{\begin{aligned} S_{xx} &= \sum x_i^2 - \bar{x}(\sum x_i) \\ S_{yy} &= \sum y_i^2 - \bar{y}(\sum y_i) \\ S_{xy} &= \sum x_i \cdot y_i - \bar{x}(\sum y_i) \\ &\quad o \\ S_{xy} &= \sum x_i \cdot y_i - \bar{y}(\sum x_i) \end{aligned}}$$

Una vez obtenidos las suma de cuadrados de las variables y obviamente sus medias aritméticas, ya se pueden calcular todas las estimaciones puntuales necesarias sabiendo que:

1. Estimador de **b**

$$\boxed{b = \frac{S_{xy}}{S_{xx}}} \tag{10.3}$$

2. Estimador de **a**

$$\boxed{a = \bar{y} - b\ \bar{x}} \tag{10.4}$$

3. Estimador de $\hat{\mathbf{Y}}$

$$\boxed{\hat{Y} = a + b\ X} \tag{10.5}$$

4. Estimador para $\sigma^{\mathbf{2}} \rightarrow s^2$

$$\boxed{s^2 = \frac{S^2_{residual}}{n-2}} \tag{10.6}$$

Donde la $S^2_{residual}$ ([8]) es:

$$S^2_{residual} = S_{yy} - \frac{S^2_{xy}}{S_{xx}}$$

$$o$$

$$S^2_{residual} = S_{yy} - b^2 S_{xx}$$

[8] Cuidado con la notación: ya que puede aparecer tanto $S_{xx/yy/xy}$ *como* $S^2_{x/y/xy}$

10.3.2. Teoría de Muestreo (TM) sobre los parámetros

Los estimadores de nuestros parámetros son:

Parámetros	A	B	σ^2	Y
Estimadores	a	b	s^2	$\hat{Y}$

Y la TM, ¿Qué es lo que dice sobre cada uno de estos estimadores?

1. **Estimador de a**

   - Esperanza

   $$E[a] = A \qquad \rightarrow no\ sesgado$$

   - Varianza([9])

   $$V[a] = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] \qquad \rightarrow \sigma^2\ conocida$$

   $$var[a] = s^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] \qquad \rightarrow \sigma^2\ desconocida$$

   - Error

   $$Error[a] = \sqrt{V[a]} \qquad \rightarrow \sigma^2\ conocida$$
   $$error[a] = \sqrt{var[a]} \qquad \rightarrow \sigma^2\ desconocida$$

   - Comportamiento (Distribución)[10]

   $$a \sim N(E,V) \rightarrow Si\ \sigma^2\ desconocida:\ a \sim t_\nu$$

2. **Estimador de b**

   - Esperanza

   $$E[b] = B \qquad \rightarrow no\ sesgado$$

   - Varianza

   $$V[b] = \frac{\sigma^2}{S_{xx}} \qquad \rightarrow \sigma^2\ conocida$$

   $$var[b] = \frac{s^2}{S_{xx}} \qquad \rightarrow \sigma^2\ desconocida$$

   - Error

   $$Error[b] = \sqrt{V[b]} \qquad \rightarrow \sigma^2\ conocida$$
   $$error[b] = \sqrt{var[b]} \qquad \rightarrow \sigma^2\ desconocida$$

   - Comportamiento

   $$b \sim N(E,V) \rightarrow \qquad Si\ \sigma^2\ desconocida:\ b \sim t_\nu$$

[9] Notación: Si σ^2 es desconocida (lo habitual), la Var[u]/Error[u] se suelen escribir con minúsculas $\Rightarrow$ var[u]/error[u]
[10] Notación: En este contexto, $\sim$ o a veces también $\dot{\cap}$, significa que: "se distribuye aproximadamente como".

3. **Estimador de $\hat{Y}$**

   - Esperanza

$$E[\hat{Y}] = Y = A + B\ X \qquad \rightarrow no\ sesgado$$

   - Varianza

$$V[\hat{Y}] = \sigma^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right] \rightarrow \sigma^2\ conocida$$

$$var[\hat{Y}] = s^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right] \rightarrow \sigma^2\ desconocida$$

   - Error

$$Error[\hat{Y}] = \sqrt{V[\hat{Y}]} \qquad \rightarrow \sigma^2\ conocida$$

$$error[\hat{Y}] = \sqrt{var[\hat{Y}]} \qquad \rightarrow \sigma^2\ desconocida$$

   - Comportamiento

$$\hat{Y} \sim N(E, V)$$

4. **Estimador**[11] **de s^2**

   - Esperanza

$$E[s^2] = \sigma^2$$

   - Varianza/Error
     (ahora no es relevante)
   - Comportamiento

$$\frac{S^2_{residual}}{\sigma^2} \sim \chi^2_{(n-2)}$$

[11] Recordar: la ecuación para la varianza residual en la ecuación (10.6)

10.4. La importancia de lo Residual

La suma de cuadrados de la "variación"total, no es más que la suma de las diferencias (distancias o dispersiones) cuadráticas de cada valor de la variable, en este caso de la variable respuesta (y), y su media aritmética ($\bar{y}$). Se puede expresar, a partir de las diferencias, como:

$$(y - \bar{y}) : S_{yy} = S^2_{Total}$$

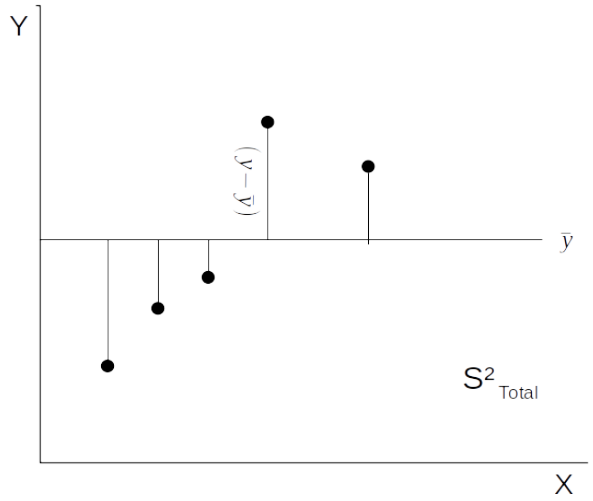


Figura 10.3: Suma de cuadrados Total.

Para cada valor de x_i hay un punto y sobre la recta horizontal de y, que no es más que la media aritmética de y: $\bar{y}$. Por lo tanto, cada segmento representa la diferencia de $(y_i - \bar{y})$. La S^2_{Total} da en general una medida básica y directa de la dispersión de los puntos. En el caso de la regresión, de la dispersión de los puntos de la variable respuesta. Precisamente, esto es lo que se quiere conocer y lo que da sentido al modelo de regresión. Saber cómo es la variación de la variable respuesta en relación (a la "influencia") de otra variable que por eso se llama variable explicativa. De ahí que se denomine **Suma de cuadrados total**.

Considerando e integrando ahora la recta de regresión, se observan otros tipos de diferencias, como son las diferencias de cada valor de y con respecto a la estimación de $\hat{y}$ derivada del modelo de regresión: $(y - \hat{y})$, a esta diferencia se le llama *residuo*. Por lo tanto, la **Suma de cuadrados residual**, es una medida de la variación residual, también llamada *variación no explicada por la recta de regresión*:

$$(y - \hat{y}) : S_{yy} - \frac{S^2_{xy}}{S_{xx}} = S^2_{Residual}$$

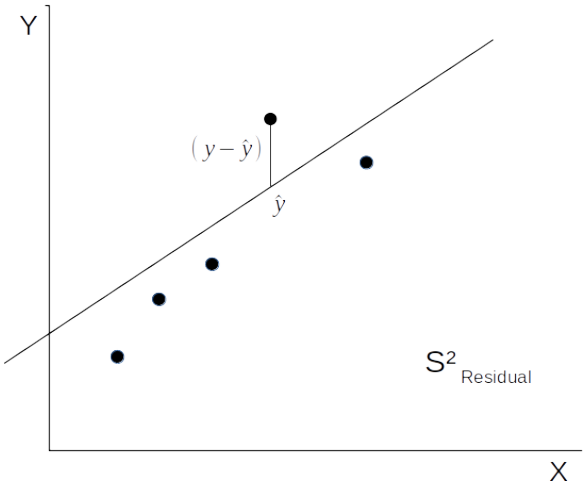


Figura 10.4: Suma de cuadrados Residual.

Dicho de otra manera, la variación residual es la variación no explicada por la recta (es residual), es lo que resta del valor observado de la recta después de que se haya ajustado un modelo de regresión. Resultando ser una herramienta muy poderosa a tener muy en cuenta porque da, entre otras cosas, una medida del ajuste.

La diferencia entre la variación total y la residual será la variación debida a la recta. Aquella parte de la variación que es *explicada* por la recta de regresión. Por lo tanto, la variación de la recta se mide a través de la **Suma de cuadrados de la recta**, que será:

$$(\hat{y} - \bar{y}) : \frac{S_{xy}^2}{S_{xx}} = S_{Recta}^2$$

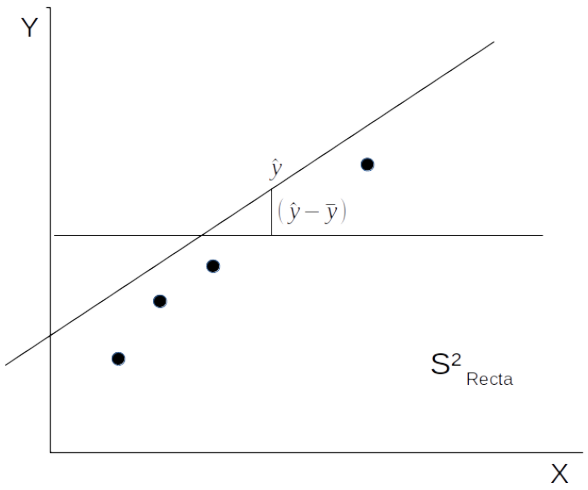


Figura 10.5: Suma de cuadrados de la Recta.

En resumen y bajo la perspectiva de las desviaciones, tenemos que:

$$(y - \bar{y}) \equiv (y - \hat{y}) + (\hat{y} - \bar{y})$$

Las desviaciones totales para cada puntuación, son idénticas a la suma de las desviaciones hasta la recta y las desviaciones de la recta hasta la media.

Si se eleva al cuadrado y se suma cada desviación, se obtiene que la variación total es igual a la suma de la variación residual más la variación explicada por la recta.

$$S^2_{Total} = S^2_{Residual} + S^2_{Recta}$$

Esta es la descomposición fundamental de la variación para el modelo de regresión. También se expresa, como: la variación total es la suma de la variación no explicada (residual) y la variación explicada por la recta. Evidentemente, cuánto mayor es la S^2_{Recta} menor será la $S^2_{Residual}$ lo que implica que será mejor el ajuste a la recta.

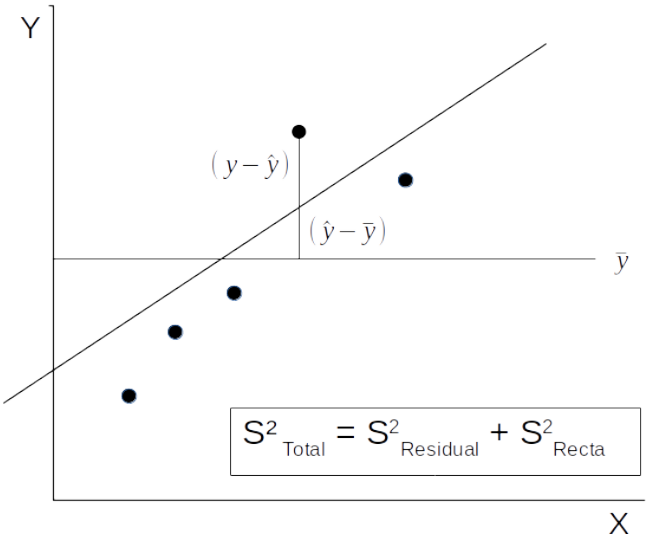


Figura 10.6: Suma de cuadrados Total de la recta de regresión.

Si la relación anterior se divide por la S^2_{Total} queda:

$$\boxed{1 = \frac{S^2_{Recta}}{S^2_{Total}} + \frac{S^2_{Residual}}{S^2_{Total}}}$$

Con lo cual se obtiene en porcentaje (o tanto por uno), sobre cual es la importancia relativa que explica o no, la variación total de la recta de regresión. A la relación $\frac{S^2_{Recta}}{S^2_{Total}}$ se le llama **coeficiente de determinación**[12].

El coeficiente de determinación (*c.d.*) es una medida de la bondad del ajuste, ya que indica el porcentaje de variación total que es explicado por la recta de regresión. Muchas veces se expresa como r^2 y es una notación muy equívoca. Esta denominación se suele, o se puede, relacionar erróneamente con el *coeficiente de correlación.* La correlación considera que la variable explicativa es aleatoria (modelo de tipo II, o de efectos variables), por lo tanto asume aleatoriedad en ambas variables, lo que **no** ocurre con el modelo de regresión simple(modelo de tipo I, o de efectos fijos),[13] donde la variable explicativa **no** es aleatoria.

[12] Realmente es una mala denominación. Sería más apropiado llamarlo: *coeficiente de explicación.*
[13] Este es un error conceptual muy extendido.

Con respecto a sus valores se puede interpretar que:

- Si $cd|r^2 = 0$ La recta no explica nada
- Si $cd|r^2 = 1$ La recta explica toda la variación
- $0 \leq cd|r^2 \leq 1$ La recta explica parte de la variación

Como resumen, se puede decir que el *coeficiente de determinación*, sirve para:

- Medida de la bondad de ajuste.
  Cuan cercanos son los puntos a la recta.
- Medida del grado de linealidad.
  "Acercamiento" residual.
- Medida directa de la mejora en el modelo.
  Mejora al introducir información al modelo: más datos, mejores datos, una u otra variable, mejores variables etc.

Nota: De alguna manera también se puede asumir que el *coeficiente de determinación* es independiente de las asunciones del modelo. [14] Esto se puede interpretar como que: si el *coeficiente de determinación* se incrementa, implica un mejor ajuste, siempre con respecto a si se cumple o no ese modelo concreto. Para eso, no se necesitan estrictamente las asunciones del modelo pero... lo invalidaría para dar niveles de significación (inferencia).

[14] Cuidado con la mala interpretación de esto.

10.4.1. Análisis visual de los residuos

El residuo en el modelo de regresión lineal simple se había definido como:

$$Residuo \;=\; Error \;= \hat{y}_i - y_i = \mathcal{E}_{ij} = \beta_0 + \beta_1 x_1 - y_i$$

Lo primero que hay que hacer es ver si se cumple:

1. $E[\mathcal{E}_i] = 0$ — Los residuos deben de estar aleatoriamente distribuidos entorno a su Esperanza nula
2. $cov(\mathcal{E}_i, \mathcal{E}_j) = 0$ — $\forall\ i \neq j \Rightarrow \nexists\ rachas$ — No existe auto-covariación/correlación
3. $Var[\mathcal{E}_i] = \sigma^2 = cte$ — Homocedasticidad
4. $\mathcal{E}_i \cap N(0, \sigma^2)$ — Distribución normal de los residuos

Patrones de representación visual sobre los residuos

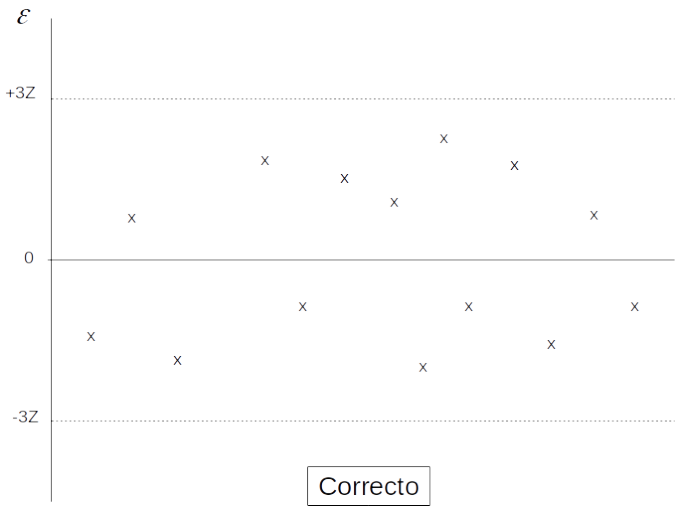


Figura 10.7: Representación correcta/esperable de los Residuos en un modelo de regresión.

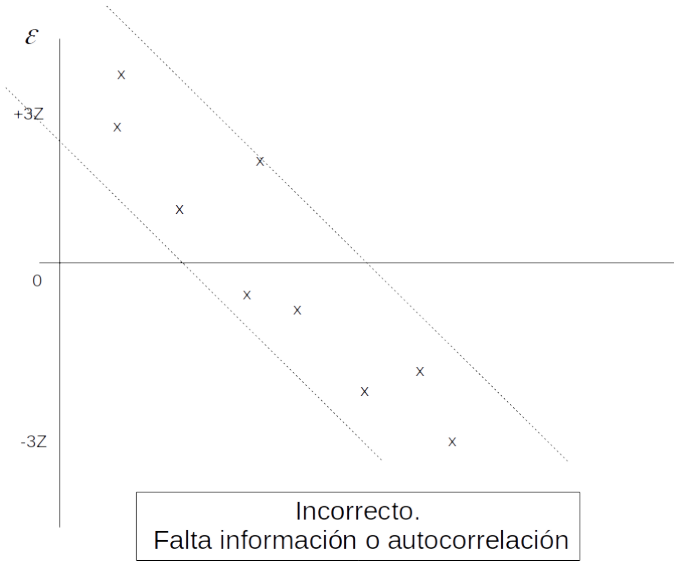


Figura 10.8: Representación incorrecta de los Residuos en un modelo de regresión.

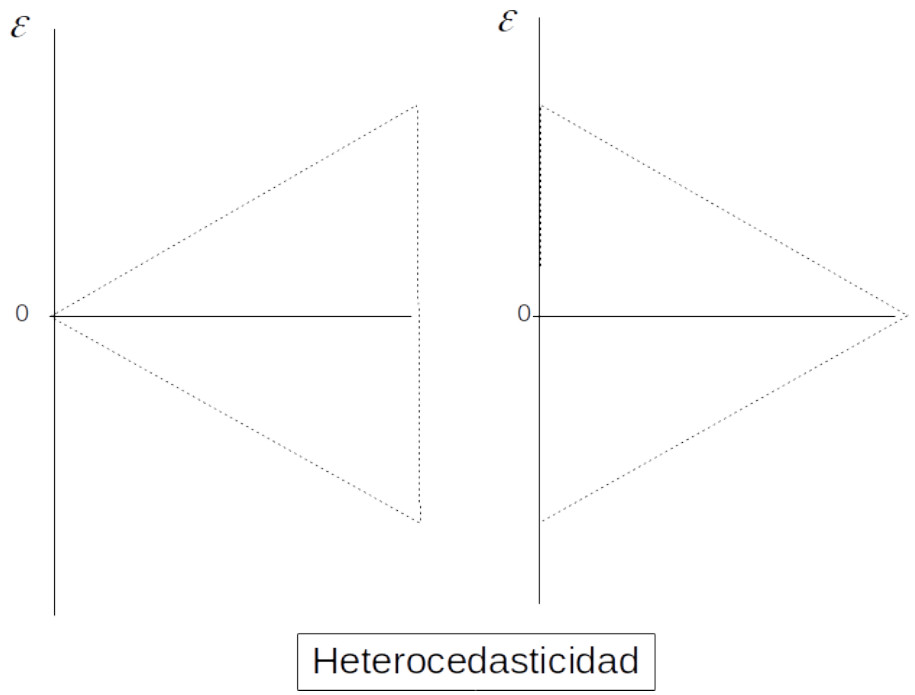


Figura 10.9: Representación con heterocedasticidad de los Residuos en un modelo de regresión.

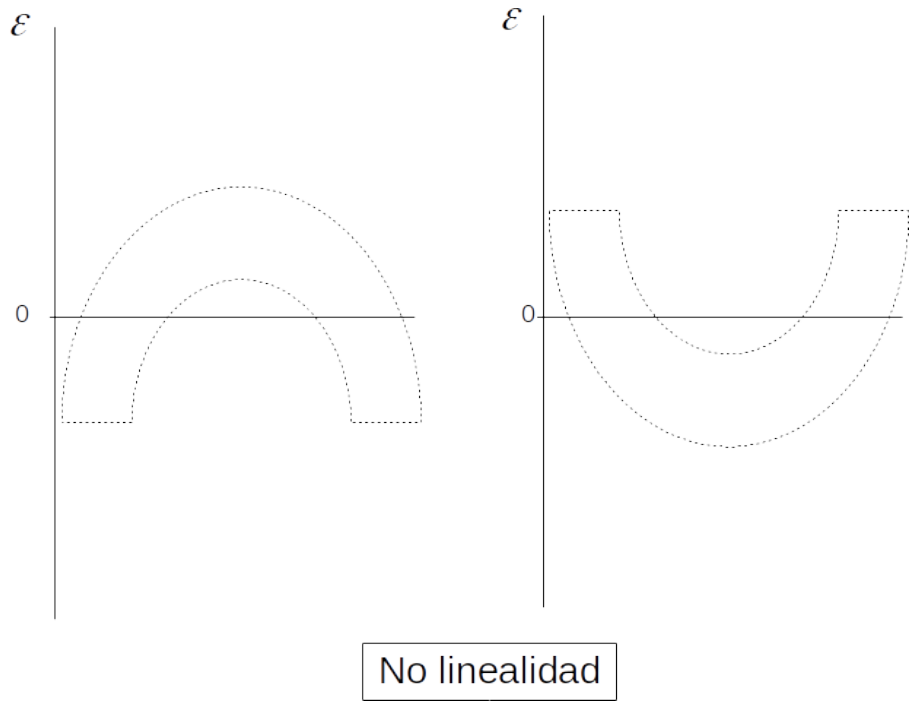


Figura 10.10: Representación con no-linealidad de los Residuos en un modelo de regresión.

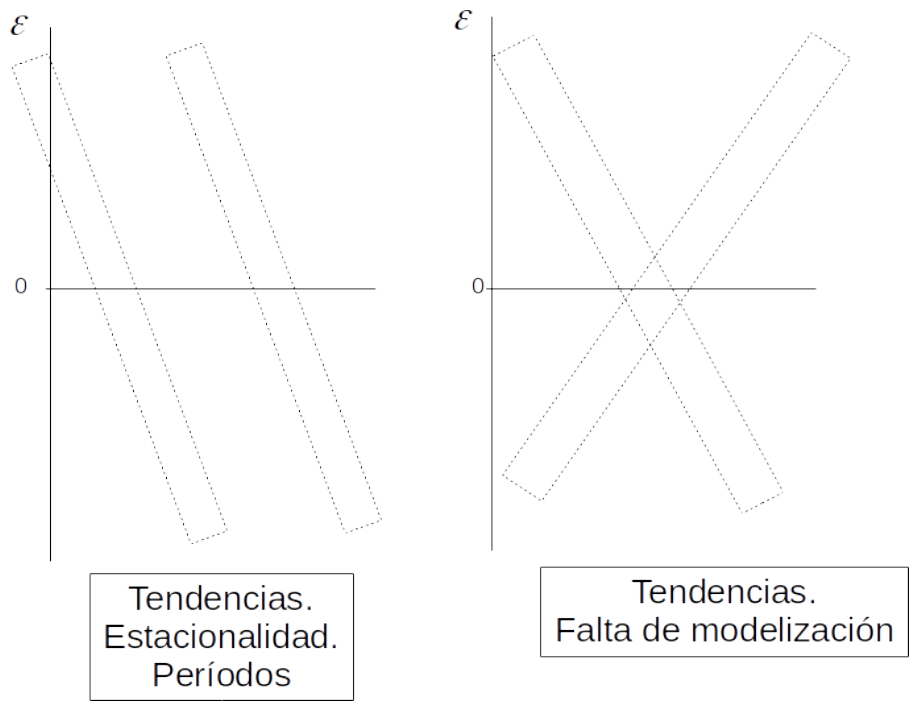


Figura 10.11: Representación con tendencias de los Residuos en un modelo de regresión.

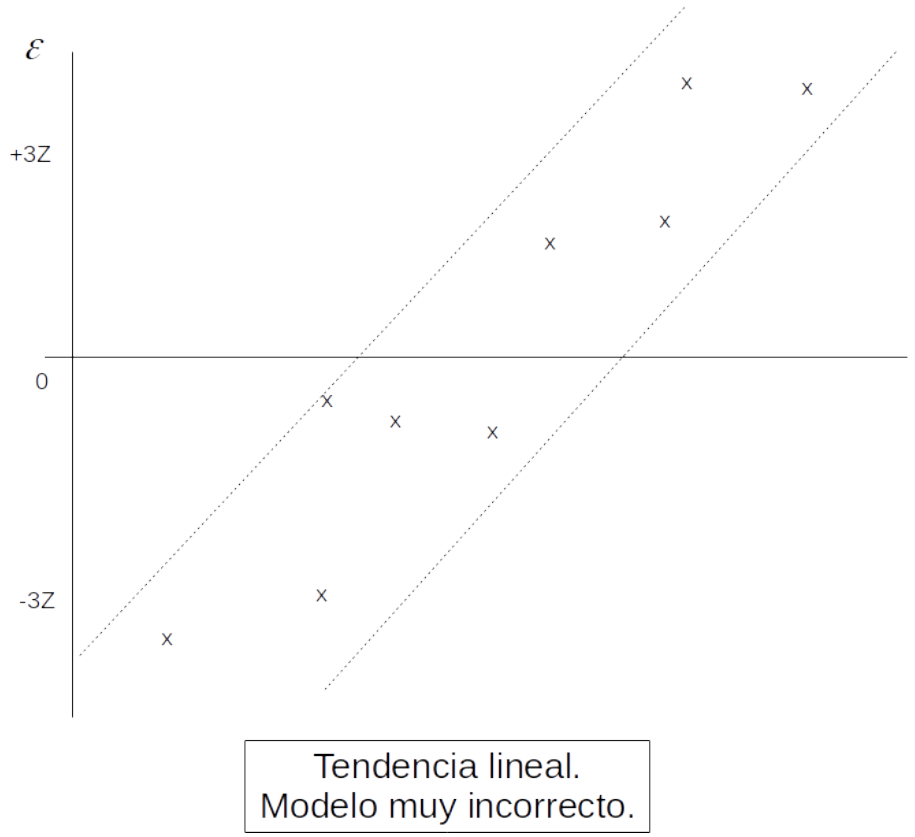


Figura 10.12: Representación con tendencias lineales de los Residuos en un modelo de regresión.

Anexo sobre cálculo de los estimadores

Para el cálculo de los estimadores y sobre todo para el de dos variables, el método de estimación se puede hacer, como ya se indicó anteriormente, a través del cálculo de diferenciales parciales de la función objetivo (la suma de cuadrados de los residuos) con respecto a cada parámetro.

Como la regresión simple es una simplificación del modelo de regresión múltiple (con varias variables explicativas) y este a su vez del Modelo Lineal General (GLM, en inglés), los cálculos son matriciales. Pero el análisis matemático, también ofrece otro tipo de aproximaciones más "sencillas"para resolver este problema.

Método analítico para más de dos variables

Se genera un sistema de dos ecuaciones partiendo del modelo básico de regresión. A la primera ecuación: se le aplica un sumatorio; y a la segunda: se le multiplica por x y se aplica el sumatorio a la ecuación, quedando[15]:

$$\begin{cases} 1: \sum y & = n \cdot a + b \cdot \sum x \\ 2: \sum xy & = a \cdot \sum x + b \cdot \sum x^2 \end{cases} \tag{10.7}$$

1. Si a la primera ecuación la dividimos por n, queda:

$$\frac{\sum y}{n} = \frac{n \cdot a}{n} + \frac{b \cdot \sum x}{n} \Longrightarrow \bar{y} = a + b\ \bar{x} \Longrightarrow \boxed{a = \bar{y} - b\ \bar{x}}$$

2. Si a la segunda ecuación, ahora le restamos el término $-\bar{y}\sum x$, queda:

$$\begin{aligned} \sum xy - \bar{y}\sum x &= a\sum x - \bar{y}\sum x + b\sum x^2 \quad ; \qquad Como\ a = \bar{y} - b\bar{x} \\ Sxy &= (-b\ \bar{x})\sum x + b\sum x^2 \\ &= b\ (\sum x^2 - \bar{x}\sum x) \qquad ; \qquad Como\ (\sum x^2 - \bar{x}\sum x) = Sxx \\ &= b\ Sxx \Longrightarrow \boxed{b = \frac{Sxy}{Sxx}} \end{aligned}$$

Solución alternativa

El sistema de ecuaciones en 10.7, también se puede resolver aplicando la *Regla de Kramer*, quedando:

$$a = \frac{\begin{vmatrix} \sum y & \sum x \\ \sum xy & \sum x^2 \end{vmatrix}}{\begin{vmatrix} n & \sum x \\ \sum x & \sum x^2 \end{vmatrix}}$$

$$b = \frac{\begin{vmatrix} n & \sum y \\ \sum x & \sum xy \end{vmatrix}}{\begin{vmatrix} n & \sum x \\ \sum x & \sum x^2 \end{vmatrix}}$$

[15] Muchas veces, por simplificación, se obvian los subíndices i en las ecuaciones.

Capítulo 11

Correlación. Efectos variables (II)

La diferencia fundamental entre la regresión (modelo I) de efectos fijos y la correlación (modelo II) de efectos variables es que en este último los pares de valores son aleatorios, i.e. varían y covarían las dos variables a relacionar.

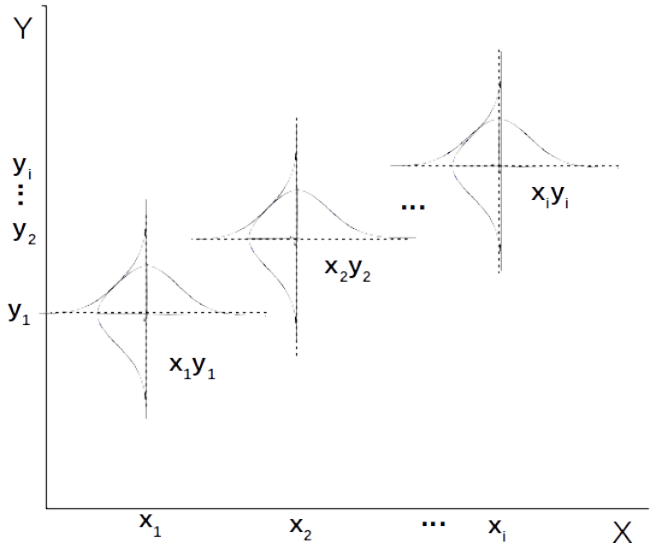


Figura 11.1: Correlación. Modelo II de efectos variables.

Se podría decir que la correlación mide la fuerza de la relación (covariación[1]) entre un conjunto de variables cuantitativas continuas[2] a través de un estadístico al efecto, que en el caso más sencillo, se denomina: coeficiente de correlación simple (r).[3]

La correlación no deja de ser una medida normalizada de la covariación entre dos variables, que está comprendida entre $-1 \leq r \leq +1$. Se define como:

- **varianza** (s^2)

$$s^2 = \frac{S_{xx}}{n-1} = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

- **covarianza** $cov(x, y)$

$$cov(x, y) = \frac{S_{xy}}{n-1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- **coeficiente de correlación de Pearson** (r)

$$r = \frac{cov(x, y)}{s_x s_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \tag{11.1}$$

[1] Repetimos: covariación no implica necesariamente causalidad.
[2] Ahora sólo hablaremos del coeficiente de correlación de Pearson(r)
[3] En el caso de más variables, se hablará de coeficiente de correlación múltiple y de los coeficientes de correlación parcial.

Capítulo 12

Análisis de Varianza (ANOVA)

El análisis de varianza permite cotejar diferencias entre factores y/o subgrupos, pertenecientes a estos factores. Un factor es siempre una variable categórica o "categorizada"; y por lo tanto cualitativa.[1]

Obviamente, el caso más sencillo es cuando se aborda un único factor, limitado o subdividido en grupos que se denominan: niveles del factor. En este caso concreto, se llamará ANOVA de un factor, o también conocidos como ANOVA de una vía, una clasificación o de una entrada. En biomédicina, se emplea también el término de: ANOVA de un tratamiento.[2,3] Por su generalidad, aquí se empleará el término de **ANOVA de un factor**.

El ANOVA, obtiene su máxima potencia analítica en el contexto del diseño experimental. Es decir, aquel en el que el investigador controla el máximo de la variabilidad causal para ver cómo es el comportamiento/reacción de una determinada variable respuesta o de interés. Por lo tanto, los factores y/o sus niveles (subgrupos) están fijados de forma discreta (cualitativa) por el experimentador y la variación se produce tanto dentro (intra) como entre (inter) grupos y/o factores.

Tanto el modelo de regresión lineal simple como el ANOVA de un factor, ya indicado previamente, pertenecen al Modelo Lineal General. Una de las diferencias más evidentes es que mientras en el modelo de regresión la variable explicativa es cuantitativa; en el ANOVA, es cualitativa. Cuando en el modelo intervienen ambos tipos de variables, entonces se habla de análisis de covarianza (ANCOVA).

12.1. En la Población

La denominación de análisis de varianza puede inducir a engaño. La prueba de hipótesis siempre es sobre las medias, aunque se trabaje con las varianzas. En el fondo, no es más que un método de comparación entre medias.

Para un factor, la población está dividida en grupos. A cada grupo se le llama nivel. Por lo tanto, la población de acuerdo a un factor tendrá varios niveles.

En cada nivel, a la variable continua se le llama: $X_1,\ X_2, \cdots,\ X_j, \cdots,\ X_k$

Grupos/ Niveles	1	2	...	j	...	k
Variable	X_1	X_2	...	X_j	...	X_k

Figura 12.1: ANOVA. Grupos o Niveles de un Factor.

[1] En escala de medida nominal u ordinal.
[2] Cuando se habla de tratamiento, se sobreentiende que es en condiciones controladas del laboratorio.
[3] A veces, en inglés, se habla de: complete randomize experimental design.

Es decir, que se tienen k niveles. Cada nivel tiene que ser visto, en este contexto, como una subpoblación de medias y varianzas:

$$\mu_1,\ \mu_2, \cdots,\ \mu_j, \cdots,\ \mu_k$$

$$\sigma_1^2,\ \sigma_2^2, \cdots,\ \sigma_j^2, \cdots,\ \sigma_k^2$$

La media de la población será la media de los niveles:

$$\mu = \frac{\sum_j \mu_j}{k}$$

Suposiciones del modelo

Las suposiciones[4] fundamentales del ANOVA, son:

1. **Independencia**. Inexistencia de autocorrelación entre las puntuaciones.
2. **Homocedasticidad**. Las varianzas de cada nivel son <u>todas</u> iguales:

$$\sigma_1^2 = \sigma_2^2, \cdots, = \sigma_j^2, \cdots, = \sigma_k^2 = \sigma^2\ (la\ total)$$

3. **Normalidad**. Las X en cada nivel se distribuyen normalmente, con la media y varianza correspondiente de su nivel:

$$X_j\ \cap\ N(\mu_j,\ \sigma_j^2)$$

Objetivo del modelo

Formalmente en investigación, el mejor objetivo se expresa a través de una prueba de hipótesis. En el ANOVA, la hipótesis se hace sobre las medias, sí <u>sobre las medias de los niveles (μ_j)</u> y la hipótesis nula es que las medias son todas iguales:

$$\boxed{H_o: \quad \mu_1 = \mu_2 = \cdots = \mu_j = \cdots = \mu_k = \mu}$$

Esto es lo que tiene que probar el ANOVA, aunque lo que "interese"sea rechazar la H_o en el supuesto de que haya diferencias significativas entre las medias de los niveles.

Criterio de selección de la muestra

El caso más sencillo es extraer, para cada nivel, una muestra aleatoria simple (m.a.s) de tamaño n_j que idealmente será igual para todos los niveles. Esto es lo que denomina como muestreo o diseño *balanceado*. También idealmente, el criterio de extracción de las muestras aleatorias será por niveles. Es decir: una muestra aleatoria para el nivel 1, luego otra para el nivel 2 etc. Aunque se sabe que muchas veces esto no es posible y la asignación de puntuaciones a niveles es *a posteriori*.

[4] Estas son las suposiciones básicas para intergrupos. Para intragrupos, habría que añadir la esfericidad (autocorrelación intra) y la aditividad (interacción sujeto:tratamiento).

12.2. En la muestra

Nivel	1	2	...	j	...	k
x	$x_{i,1}$	$x_{i,2}$	...	$x_{i,j}$	...	$x_{i,k}$
n	n_1	n_2	...	n_j	...	n_k
$\bar{x}$	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_j$	...	$\bar{x}_k$
S_{xx}	$(S_{xx})_1$	$(S_{xx})_2$	...	$(S_{xx})_j$	...	$(S_{xx})_k$
s^2	s^2_1	s^2_2	...	s^2_j	...	s^2_k

Figura 12.2: ANOVA. Estadísticos en la muestra de los niveles para un Factor.

Para cada nivel, los estadísticos fundamentales serán:

- Media:

$$\boxed{\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}}$$

- Suma de cuadrados:

$$\boxed{(S_{xx})_j = \sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2 = \sum_{i=1}^{n_j} x_{ij}^2 - \bar{x}_j(\sum_{i=1}^{n_j} x_{ij})}$$

- Varianza:

$$\boxed{s_j^2 = \frac{(S_{xx})_j}{n_j - 1}}$$

Dentro del mundo de la muestra, la media total $\bar{x}$ será:

$$\bar{x} = \frac{\sum x_{ij}}{n} = \frac{\sum n_j \bar{x}_j}{n} \qquad ; \; n = \sum n_j$$

Que la media total, no es más que la suma de todos las puntuaciones, dividido entre el total de elementos de la muestra.

12.2.1. Suma de Cuadrados. Desarrollo.

En función de las distancias observadas con respecto a diferentes valores medios, tendremos las desviaciones de interés en el ANOVA. Medidas a través de las diferentes **Suma de Cuadrados**.

1. **Suma de Cuadrados Total**, para las desviaciones $\boxed{(x_{ij} - \bar{x})}$:

$$\boxed{S^2_{Total} = \sum_{i,j}(x_{ij} - \bar{x})^2 = \sum_{i,j} x^2_{ij} - \bar{x}\sum_{i,j} x_{ij}}$$

2. **Suma de Cuadrados Dentro o Intra**[5], para las desviaciones $\boxed{(x_{ij} - \bar{x}_j)}$:

$$\boxed{S^2_{Dentro} = \sum_j [\sum_i (x_{ij} - \bar{x}_j)^2] = \sum_j (S_{xx})_j}$$

3. **Suma de Cuadrados Entre o Inter**[6], para las desviaciones $\boxed{(\bar{x}_j - \bar{x})}$:

$$\boxed{S^2_{Entre} = \sum_j n_j(\bar{x}_j - \bar{x})^2}$$

   El n_j es una ponderación, ya que cada nivel puede tener diferente número de elementos. En el caso de ser estrictamente balanceado, todos son iguales.

Para las desviaciones

En resumen, y viéndolo desde la perspectiva de las desviaciones, se tiene que:

$$(x_{ij} - \bar{x}) = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})$$

La desviación total, con respecto a las correspondientes medias, es la suma de las desviación de dentro y la desviación entre, de los grupos o niveles.

$$\boxed{S^2_{Total} = S^2_{Dentro} + S^2_{Entre}}$$

Con lo cual, así se descompone la variación total: dentro y entre los niveles.

12.3. En el Muestreo

¿Y qué dice la TM a este respecto?

1. Para cada nivel j:

   - $E[\bar{x}_j] = \mu_j$
   - $V[\bar{x}_j] = \dfrac{\sigma^2}{n_j} \quad \rightarrow \quad var[\bar{x}_j] = \dfrac{s^2}{n_j}$
   - $\bar{x}_j \cap N(E, V)$
   - También dice que:

   $$\frac{(S_{xx})_j}{\sigma^2} \cap \chi^2_{(n_j-1)} \quad \Rightarrow E[\frac{(S_{xx})_j}{n_j - 1}] = \sigma^2 \quad \Rightarrow E[s^2_j] = \sigma^2$$

2. Para todos los (j) niveles será:

   $$\frac{\sum_j (S_{xx})_j}{\sigma^2} \cap \chi^2_{(n-k)}$$

   que no es más que: $\dfrac{S^2_{Dentro}}{\sigma^2} \cap \chi^2_{(n-k)}$

[5] También Dentro o Intragrupos. En inglés within (S^2_W).
[6] También Entre o Intergrupos. En inglés between (S^2_B).

Al final se obtienen dos[7] resultados fundamentales para el ANOVA:

1. Para Dentro:

$$\boxed{\frac{S^2_{Dentro}}{\sigma^2} \cap \chi^2_{(n-k)}}$$

2. Para Entre:

$$\boxed{\frac{S^2_{Entre}}{\sigma^2} \cap \chi^2_{(k-1)}}$$

Las relaciones anteriores son verdaderas sí y sólo sí, la H_o es verdadera. Esta propiedad fundamental del ANOVA es la que se usa para el contraste de hipótesis. Dicho de otra forma, si las medias son distintas esto no es verdadero parcialmente, ya que: $\frac{S^2_{Dentro}}{\sigma^2} \cap \chi^2_{(n-k)}$ siempre es cierto; pero **no** que $\frac{S^2_{Entre}}{\sigma^2} \cap \chi^2_{(k-1)}$

Relacionando ambos (Dentro vs Entre) entornos se tiene formalmente que:

Si H_o es verdadero, entonces:

$$\boxed{\frac{S^2_{Entre}/(\text{k - 1})}{S^2_{Dentro}/(\text{n - k})} \cap F_{(k-1,n-k)}}$$

Y esta es la base fundamental del ANOVA. Si la relación anterior tuviera una distribución F con esos grados de libertad, entonces la H_o es verdadera y si no: es falsa[8].

Prueba de hipótesis y Fuente de Variación

Como ya se mencionó anteriormente la hipótesis será:

$$H_o: \quad \mu_j = \mu$$
$$H_1: \quad \mu_j \neq \mu$$

Con todo lo desarrollado anteriormente, se construye la tabla de la **Fuente de Variación** o también conocido como *cuadro de ANOVA*:

Fuente de Variación	g.l.	Suma de Cuadrados	Cuadrado Medio	Fobs
ENTRE $(\bar{x}_j - \bar{x})$	k - 1	$S^{\wedge}2_{ENTRE}$	$S^{\wedge}2_{ENTRE}$ / (k - 1)	cociente
DENTRO $(x_{ij} - \bar{x}_j)$	n - k	$S^{\wedge}2_{DENTRO}$	$S^{\wedge}2_{DENTRO}$ / (n - k)	
TOTAL $(x_{ij} - \bar{x})$	n - 1	$S^{\wedge}2_{TOTAL}$		

Figura 12.3: Cuadro de ANOVA para un Factor.

[7] Para S^2_{Entre} no se presenta su desarrollo.
[8] Cuidado con el orden de los grados de libertad: primero numerador y luego denominador.

La prueba se puede hacer de dos maneras:

- De forma clásica o histórica:

  Establecer la región crítica con un determinado nivel de confianza (e.g. 0.95) para la aceptación/rechazo de la H_o mediante tablas estadísticas de la distribución F.

$$Prob\{F'_{(n-1,n-k)} \leq \ ?\} = 0{,}95 \quad ; \ ? \ es \ el \ punto \ crítico$$

$$\text{Si el } F_{obs} > \ ? \ \text{ entonces se rechaza la } H_o$$

- Por ordenador:

  Lo que se hace por esta vía es calcular la probabilidad del F_{obs}, estableciendo *a priori* el nivel de significación para la aceptación/rechazo de la H_o:

$$Prob\{X \leq F_{obs(n-1,n-k)}\} = P \quad \Rightarrow \quad (1-P)$$
$$Si(1-P) \leq \alpha \quad \Rightarrow \ Se \ rechaza \ la \ H_o$$
$$Si(1-P) > \alpha \quad \Rightarrow \ No \ se \ rechaza \ la \ H_o$$

  En realidad hoy en día, el ordenador ya nos da directamente el valor de (1-P)