Apéndice A

A.1. Repaso de probabilidad

Las variables económicas tienen un componente sistemático y otro aleatorio, ya que con anterioridad a su observación no podemos predecir con certeza los valores que van a tomar. Este apartado revisa los conceptos de probabilidad que aplicaremos este curso: qué es una variable aleatoria o *estocástica*, cuáles son sus propiedades y, finalmente, se presentan las distribuciones de probabilidad más usuales.

A.1.1. Una variable aleatoria

Una *variable aleatoria*, que denotamos por X, es aquella cuyo valor no es conocido con anterioridad a su observación. La probabilidad es un medio para expresar la incertidumbre sobre el resultado. Se distinguen dos tipos de variables aleatorias: *discretas*, cuando el conjunto de todos sus posibles valores es finito o infinito numerable, y *continuas*, cuando el conjunto de realizaciones es infinitamente divisible y, por tanto, no numerable. Por ejemplo, la superficie de una vivienda es una variable continua mientras que el número de baños es una variable discreta. En general, en este curso nos ocuparemos de variables continuas.

Si X es una variable discreta, podemos asignar una probabilidad $p(x_i) = Prob(X = x_i)$ a cada posible resultado x_i. El conjunto de probabilidades, que se denomina *función de probabilidad*, debe cumplir que $p(x_i) \geq 0$ y $\sum_i p(x_i) = 1$.

Si X es continua, la probabilidad asociada a cualquier punto en particular es cero, por lo que nos referimos a la probabilidad de que X tome valores en un intervalo $[a, b]$. La *función de densidad* $f(x)$ de una variable aleatoria continua X es una función tal que

$$Probabilidad(a \leq X \leq b) = \int_a^b f(x)\, dx$$

Es decir, el área por debajo de la función entre dos puntos a y b es la probabilidad de que la variable tome valores en el intervalo $[a, b]$ (ver panel izquierdo del Gráfico A.1). La función de densidad toma valores no negativos, $f(x) \geq 0$, y el área total por debajo de la función es la unidad, $\int_{-\infty}^{\infty} f(x)dx = 1$.

Un ejemplo de variable aleatoria continua es la **distribución normal**. Su función de densidad tiene forma de campana (ver panel izquierdo del Gráfico A.1). Es muy utilizada en la práctica para modelar variables que se distribuyen simétricamente alrededor de un valor central, con

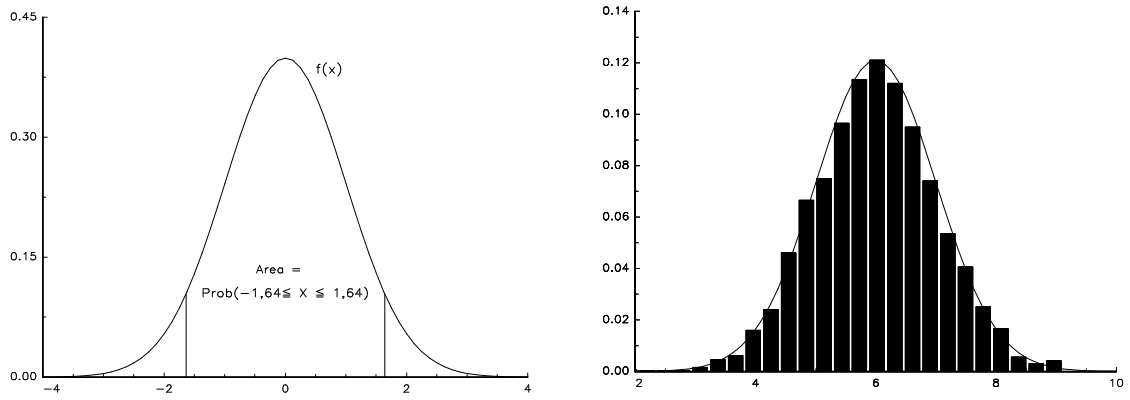


Gráfico A.1: La función de densidad *normal* y el histograma

mucha probabilidad acumulada en valores cercanos a dicho punto central y poca en valores alejados.

El panel derecho del Gráfico A.1 ilustra la relación entre la función de densidad y el histograma de los datos. Tal y como mencionan Peña & Romo (1997): "*La función de densidad constituye una idealización de los histogramas de frecuencia o un* ***modelo*** *del cual suponemos que proceden las observaciones. El histograma representa frecuencias mediante áreas; análogamente, la función de densidad expresa probabilidades por áreas. Además, conserva las propiedades básicas del histograma: es no negativa y el área total que contiene es uno.* "

La distribución de una variable aleatoria puede resumirse utilizando medidas de posición (media, mediana y moda), dispersión (varianza, desviación típica y coeficiente de variación) o forma (coeficiente de asimetría y coeficiente de curtosis). Estos conceptos se definen de forma similar a los utilizados para resumir las características de un conjunto de datos. Definiremos los elementos que utilizaremos a lo largo del curso.

La media o valor esperado, μ, de una variable aleatoria X se define como el promedio ponderado de todos los posibles valores que puede tomar X, donde la ponderación es la probabilidad de cada valor. Si la variable es continua se define:

$$\mu \;=\; E(X) \;=\; \int_{-\infty}^{\infty} x\, f(x) dx$$

donde E se conoce como el operador de esperanzas matemáticas o, simplemente, esperanzas. La media recoge el centro de gravedad sobre el que se distribuye la variable. Así, cuanto mayor sea la media, mayor es el valor que se espera que tomen las realizaciones del experimento (ver panel izquierdo del Gráfico A.2).

La varianza de una variable aleatoria X es su momento central, o respecto a la media, de orden 2. Es decir,

$$var(X) \;=\; \sigma_X^2 \;=\; E[(X-\mu)^2] \geq 0$$

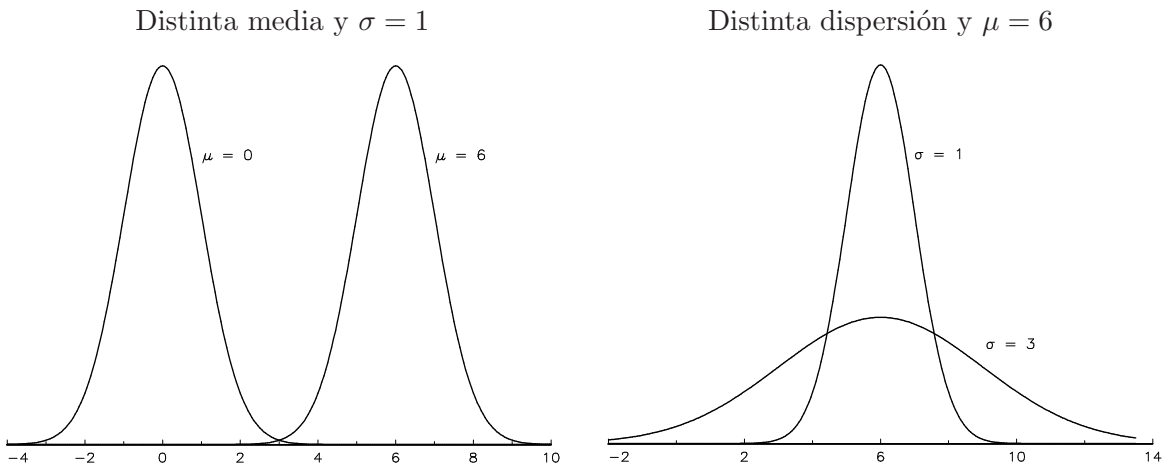


Gráfico A.2: Ejemplos de distribución normal

La varianza es una medida de dispersión de la distribución. Su raíz cuadrada positiva se conoce como **desviación típica** o **desviación estándar** de la variable aleatoria X, es decir:

$$des(X) = \sigma_X = \sqrt{var(X)}$$

El panel derecho del Gráfico A.2 muestra que cuanto menor es la varianza de la variable, mayor es la probabilidad concentrada alrededor de la media.

Distribución normal estándar. La distribución normal se caracteriza por el valor de su media y su varianza. Si Z es una variable aleatoria normal de media igual a 0 y varianza igual a la unidad, se dice que Z es una variable normal estándar y se denota $Z \sim N(0,1)$. Existen tablas de esta distribución que a cada posible resultado z le asigna la probabilidad acumulada hasta ese punto, $Prob(Z \leq z)$.

En general, si X es una variable normal con media μ y varianza σ^2 se denota $X \sim N(\mu, \sigma^2)$. Dado que la transformación $Z = (X - \mu)/\sigma$ es una normal estándar, con la tabla de esta distribución normal se obtiene la probabilidad acumulada $Prob(X \leq x)$.

Ejercicio 1: simulación normal estándar. Crea un conjunto de datos artificiales (N=250 observaciones), generados a partir de variables aleatorias normales estándar independientes. El proceso es el siguiente:

1. En Gretl, crea el conjunto de datos siguiendo los pasos: *Archivo →Nuevo conjunto de datos*, en *Número de observaciones:* escribe 250, elige la estructura de datos *de sección cruzada* y pincha en *No* desea empezar a introducir los valores. Se crea un conjunto de datos con dos variables que genera Gretl automáticamente: la constante *const* y la variable índice *index*, que toma valores 1,2,3,...,250.
2. Crea una serie de 250 realizaciones independientes de una variable normal con:

   Añadir →Variable aleatoria →Normal ...

Aparece un cuadro titulado *gretl: variable normal* donde debes indicar el nombre de la variable, su media y su desviación típica σ. Por ejemplo, para generar observaciones de una variable que llamamos z1 y que se distribuye como una N(0,1), escribimos:

z1 0 1

Tras pinchar en *Aceptar*, en la ventana principal de Gretl aparece la variable creada, $z1$, con la nota explicativa $z1 = normal()$.

3. Repitiendo el paso 2, crea una nueva realización de la normal estándar y llámala $z2$.
4. Haz dos gráficos, uno con $z1$ y otro con $z2$, sobre la variable índice con la opción: *Ver →Gráficos →Gráfico X-Y (scatter)*. Observa sus características comunes: los datos oscilan en torno al valor cero, y la mayor parte de ellos se encuentra en el intervalo (-2, 2).
5. Compara el histograma de las frecuencias relativas con la función de densidad normal. Para ello debes situar el cursor sobre una de las variables y seguir la ruta:

   Variable →Gráfico de frecuencias →contra la normal

   El resultado es un gráfico similar (no idéntico) al Gráfico A.3.

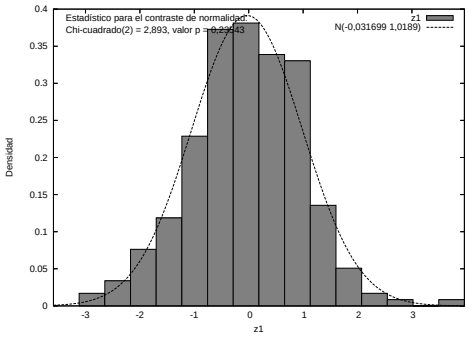


Gráfico A.3: Simulación 1: histograma

En este gráfico aparece el histograma junto con la función de densidad de la distribución normal de media $\mu = 0,1087$ y desviación típica $\sigma = 1,0055$. Estos valores aparecen en la parte superior derecha del gráfico y se eligen en función de la media y varianza de los datos.

Ejercicio 2: simulación normal general. En el mismo fichero crea dos series de datos:

- $x3$= 250 datos generados con una variable normal de media 25 y desviación típica 6 (es decir, $\sigma^2 = 36$). En *Añadir →Variable aleatoria →Normal ...* escribir *x3 25 6*.
- $x4$, generados a partir de una distribución normal de media 50 y desviación típica 0.

Haz el gráfico de los datos sobre la variable *index* y su distribución de frecuencias frente a la normal. ¿Hay algún problema al crear o representar la distribución de

$x4$? ¿Por qué?

Ejercicio 3: transformación lineal. Se trata de construir una nueva serie de datos, que llamaremos $z3$ y que se define a partir de la variable $x3$ del ejercicio anterior:

$$z3 = \frac{x3 - 25}{6}$$

1. Pincha en la opción *Añadir →Definir nueva variable*.
2. En la siguiente ventana escribe el nombre de la nueva serie y su fórmula de cálculo, es decir *z3=(x3-25)/6*.

Si has realizado el proceso correctamente, en la ventana principal de Gretl aparece la variable creada, $z3$. Haz el histograma de $z3$, comparándola con la de la variable inicial $x3$. Compara sus estadísticos descriptivos, en particular, las medias y las varianza. ¿Cambian mucho?

A.1.2. Dos o más variables aleatorias

Para responder a preguntas relativas a dos o más variables aleatorias debemos conocer su función de densidad conjunta. Si las variables aleatorias X e Y son discretas, a cada posible par de resultados (x_i, y_j) podemos asignar una probabilidad $p(x_i, y_j)$. El conjunto de probabilidades es la *función de probabilidad conjunta*, cumpliéndose que $0 \leq p(x_i, y_j) \leq 1$ y $\sum_i \sum_j p(x_i, y_j) = 1$.

Si las variables aleatorias son continuas, su distribución conjunta se recoge mediante la *función de densidad conjunta* $f(x, y)$. Si las dos variables siguen una distribución normal, la forma típica de su función de densidad conjunta se encuentra en el Gráfico A.4.

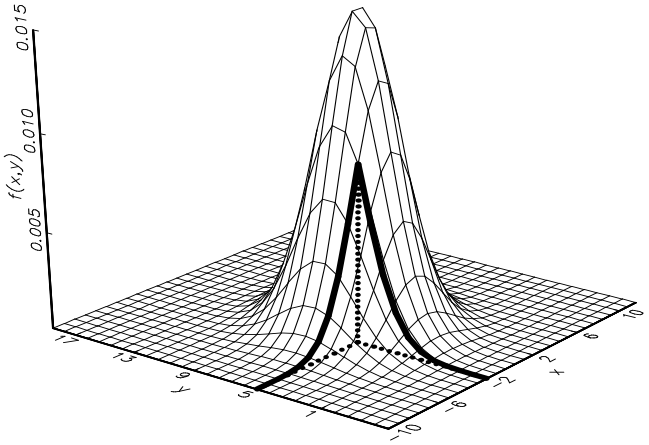


Gráfico A.4: Distribución normal bivariante

El volumen total recogido bajo esta superficie es la masa de probabilidad total que es igual a la unidad, es decir, $\int_x \int_y f(x, y)\, dx\, dy = 1$. Además, la función no toma valores negativos, $f(x, y) \geq 0$. Así, el volumen debajo del rectángulo definido por dos puntos (a, b) mide la probabilidad de que X tome valores por debajo de a e Y por debajo de b. Es decir,

$$Probabilidad(X \leq a,\, Y \leq b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y) dx\, dy$$

Por ejemplo, el volumen recogido bajo la superficie marcada en el Gráfico A.4 es la probabilidad de que $X \leq -2$ e $Y \leq 4{,}5$. La **función de densidad marginal** de cada variable puede obtenerse mediante integración. Así:

$$f(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \qquad f(y) = \int_{-\infty}^{\infty} f(x,y)\,dx \tag{A.1}$$

La distribución conjunta de dos variables aleatorias se puede resumir mediante:

- El centro de gravedad de cada variable, es decir, las medias (μ_X, μ_Y), que se obtienen de las distribuciones marginales (A.1).
- Medidas de dispersión de cada variable alrededor de su media, por ejemplo, las varianzas de X e Y, σ_X^2 y σ_Y^2, que se derivan de las distribuciones marginales (A.1).
- Medida de la relación lineal entre las dos variables aleatorias, para lo que se utiliza la covarianza σ_{XY}:

  $$cov(X,Y) = \sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)]$$

  o bien el coeficiente de correlación entre las variables,

  $$corr(X,Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \in [-1,1]$$

  Covarianza y correlación de las variables aleatorias tienen una interpretación similar a sus homólogas en los datos. Así, si $\sigma_{XY} = \rho_{XY} = 0$ se dice que las variables X e Y están incorrelacionadas.

La distribución conjunta se resume en el vector de medias μ y la matriz de varianzas y covarianzas Σ ó V:

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \qquad \Sigma = \begin{pmatrix} var(X) & cov(X,Y) \\ cov(X,Y) & var(Y) \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$$

Distribución condicionada. Al estudiar un conjunto de variables, interesa evaluar la posibilidad de que un suceso ocurra dado que otro suceso ha tenido lugar. Por ejemplo, ¿cuál es la probabilidad de que una mujer casada y con hijos en edad escolar participe en el mercado de trabajo? La **probabilidad condicionada** permite responder este tipo de preguntas. Si las variables son discretas, se define la distribución condicional de Y dado que la variable aleatoria X toma el valor x_i como:

$$Prob(Y = y_j | X = x_i) = \frac{Prob(Y = y_j, X = x_i)}{Prob(X = x_i)} = \frac{p(x_i, y_j)}{\sum_j p(x_i, y_j)} \quad \text{para } Prob(X = x_i) > 0$$

Si las variables son continuas, se define la función de densidad de Y condicionada a que la variable aleatoria X tome el valor x (para $f(x) > 0$):

$$f(y|X = x) = \frac{f(x,y)}{f(x)}$$

De esta forma se obtiene una nueva distribución, con las propiedades ya vistas. Los momentos de interés de esta distribución se denominan media y varianza condicionada de Y para el valor dado de $X = x$, y se denotan $E(Y|X = x)$ y $var(Y|X = x)$.

Independencia. Dos variables aleatorias X y Y son estadísticamente independientes o están independientemente distribuidas si conocido el valor que toma una de ellas, no aporta ninguna información sobre el valor que puede tomar la segunda. Si las variables X e Y son independientes, entonces su función de densidad conjunta puede descomponerse según:

$$f(x,y) = f(x) \times f(y) \qquad -\infty < x, y < \infty$$

Además, se tiene que $f(y|X = x) = f(y)$. Se demuestra que si X e Y son independientes, entonces $Cov(X,Y) = 0$. También se demuestra que, si las variables X e Y se distribuyen conjuntamente según una normal y $Cov(X,Y) = 0$, entonces X e Y son independientes.

Más de dos variables. Los resultados anteriores se pueden generalizar a un conjunto de n variables, $X_1, X_2, \ldots, X_n$, que se recogen en un vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

La distribución conjunta de estas variables se resume en el vector de **medias** $E(\mathbf{X})$ ó $\vec{\mu}$ y la matriz de **varianzas y covarianzas** $V(\mathbf{X})$ ó Σ_X. Así:

$$E(\mathbf{X}) = \vec{\mu} = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \qquad \text{y}$$

$$\Sigma_X = \begin{pmatrix} var(X_1) & cov(X_1,X_2) & \ldots & cov(X_1,X_n) \\ cov(X_1,X_2) & var(X_2) & \ldots & cov(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_1,X_n) & cov(X_2,X_n) & \ldots & var(X_n) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \ldots & \sigma_{1,n} \\ \sigma_{1,2} & \sigma_2^2 & \ldots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,n} & \sigma_{2,n} & \ldots & \sigma_n^2 \end{pmatrix}$$

donde Σ_X es una matriz cuadrada de orden n, simétrica y definida no negativa. Esto implica que los elementos de la diagonal principal son no negativos, $\sigma_i^2 \geq 0$, $\forall i$.

Si las variables son mutuamente independientes, entonces están incorrelacionadas, es decir, $\sigma_{i,j} = 0, \forall i \neq j$, por lo que la matriz Σ_X es diagonal:

$$\Sigma_X = \begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_n^2 \end{pmatrix}$$

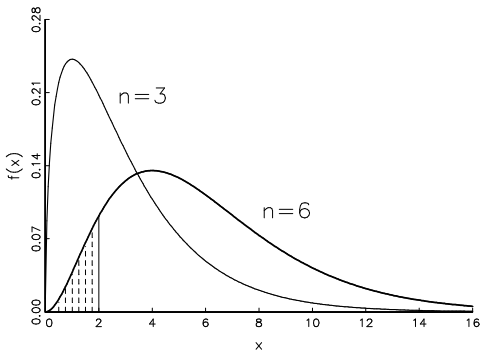


Gráfico A.5: Función de densidad de la distribución Chi-cuadrado

Si, además, $X_1, \ldots, X_n$ siguen la misma distribución, con la misma media y la misma varianza:

$$E(\mathbf{X}) = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} \qquad \Sigma_X = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}$$

entonces se dice que son variables aleatorias idéntica e independientemente distribuidas con media μ y varianza σ^2 y se denota $X_i \sim iid(\mu, \sigma^2), \forall i = 1, \ldots, n$.

Si $X_1, \ldots, X_n$ son variables aleatorias normales, se dice que el vector $\mathbf{X}$ sigue una **distribución normal multivariante**, y queda caracterizada por su vector de medias $\vec{\mu}$ y su matriz de varianzas y covarianzas Σ_X. Se denota $\mathbf{X} \sim N(\vec{\mu}, \Sigma_X)$. Si además las variables son independientes, con media y varianza común, se denota $X_i \sim NID(\mu, \sigma^2)$, $i = 1, \ldots, n$.

Además de la distribución normal, a lo largo del curso utilizaremos otras distribuciones, todas ellas relacionadas con la distribución normal. Veamos sus propiedades.

A.1.3. Algunas distribuciones de probabilidad

La distribución Chi-cuadrado. Si $(Z_1, \ldots, Z_n)$ son variables aleatorias independientes con distribución normal estándar, es decir, $Z_i \sim NID(0, 1)$, se dice que $X = \sum_{i=1}^{n} Z_i^2$ es una variable aleatoria chi-cuadrado de n grados de libertad y se denota $X \sim \chi^2(n)$. Para valores negativos de X, $f(x) = 0$ y la forma general de su función de densidad se recoge en el Gráfico A.5.

Es una distribución asimétrica, con media igual a n y varianza $2n$. Existen tablas que proporcionan la probabilidad acumulada hasta un punto $Prob(X \leq x)$, es decir, el área rayada del gráfico, en función de los grados de libertad, n.

> **Ejercicio 4: transformación no lineal.** Siguiendo el procedimiento del ejercicio 3, crea una nueva serie de datos, $y = z1^2 + z2^2 + z3^2$. En este caso debes escribir:
>
> *y = z1^2 + z2^2 + z3^2*

Haz la representación gráfica de la distribución de frecuencias de esta variable frente a la normal. El histograma que obtengas tendrá un patrón bastante diferente a la distribución normal. ¿Puedes justificar el resultado? ¿Con qué distribución la compararías?

La distribución F de Snedecor. Si $Z_1 \sim \chi^2(n_1)$ y $Z_2 \sim \chi^2(n_2)$ y además se distribuyen independientemente, entonces la distribución $X = (n_2/n_1)(Z_1/Z_2)$ se conoce como distribución F de n_1, n_2 grados de libertad y se escribe:

$$X = \frac{Z_1/n_1}{Z_2/n_2} \sim \mathcal{F}(n_1, n_2)$$

El Gráfico A.6 muestra su función de densidad para distintos grados de libertad.

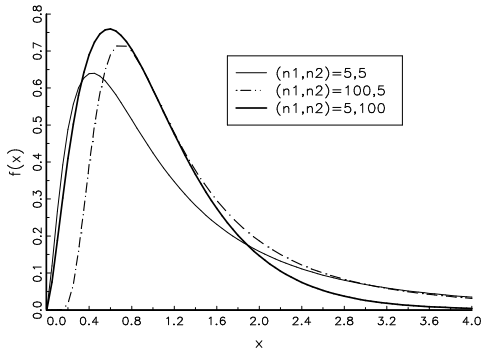


Gráfico A.6: Función de densidad de la distribución F-Snedecor

La probabilidad se acumula en la parte positiva de la recta real, $x > 0$. A medida que aumentan los grados de libertad del denominador, $n_2 \to \infty$, la distribución de $n_1\mathcal{F}(n_1, n_2)$ converge a la distribución $\chi^2(n_1)$.

La distribución t de Student. Si $Z \sim N(0, 1)$ e $Y \sim \chi^2(n)$ y además, Z e Y se distribuyen independientemente, entonces la distribución de $X = Z/\sqrt{Y/n}$ se denomina distribución t de Student de n grados de libertad y se denota:

$$X = \frac{Z}{\sqrt{Y/n}} \sim t(n)$$

El Gráfico A.7 incluye ejemplos de la función de densidad de la t-Student comparándolas con la distribución normal estándar:

Se trata de una distribución simétrica alrededor de 0. Para $n > 1$, la media de la distribución es cero y para $n > 2$ su varianza es igual a $n/(n-2)$. Esta distribución tiene las colas más gruesas que la normal, es decir, su exceso de curtosis es positivo, pero, a medida que aumentan sus grados de libertad, la distribución t converge a la normal estándar.

A.2. Repaso de inferencia estadística

Supongamos que interesa conocer cuál es el salario medio de los recién licenciados. Se trata de una población o conjunto de individuos muy amplio, por lo que se recoge la información

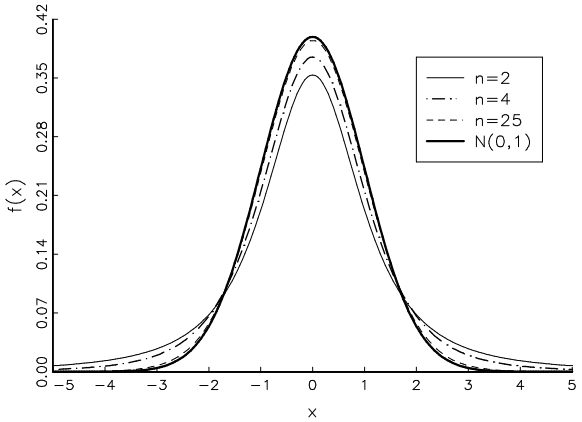


Gráfico A.7: Función de densidad de la distribución t-Student

únicamente de una muestra o un subconjunto de recién licenciados seleccionados al azar. Con esta información, ¿qué es posible inferir del salario esperado de un recién licenciado? Para responder a esta pregunta y, en general, saber usar los datos para examinar conjeturas y relaciones sobre la población repasaremos algunos conceptos de inferencia estadística.

El objetivo de la inferencia estadística es aprender determinadas características de una población a partir del análisis de una muestra. La **población** es un conjunto bien definido de elementos que son el objeto del estudio, por ejemplo, el conjunto de familias de un país, el conjunto de viviendas de una ciudad o los clientes de una empresa de telecomunicaciones. La **muestra** está formada por un subconjunto representativo de elementos de la población.

Una vez definida la población, hay que especificar un modelo para los datos que recoja las características poblacionales que interesan. En Econometría suponemos que los datos $y_1, y_2, \ldots, y_N$ son realizaciones de N variables aleatorias cuya distribución conjunta depende de varios parámetros desconocidos Θ. Un **modelo** para los datos especifica las características generales de la distribución junto con el vector de parámetros desconocidos Θ. Por ejemplo, supongamos que nos interesa conocer el precio *medio* del metro cuadrado de un piso en una ciudad y la muestra está formada por 50 pisos. Suponemos que los valores recogidos del precio por m^2 de los 50 pisos, $y_1, \ldots, y_{50}$, son realizaciones de variables normales idéntica e independientemente distribuidas. Por tanto, el modelo especificado para los datos es:

$$Y_i \sim NID(\mu, \sigma^2)$$

Los parámetros que determinan la distribución son la media y la varianza del precio del m^2, que son desconocidos, es decir, $\Theta = (\mu,\, \sigma^2)$. Además, la media es el parámetro de interés en el estudio y queremos *aprender* sobre ella a partir de los datos.

En grandes líneas, aplicaremos dos herramientas de la estadística, la estimación y el contraste de hipótesis. En la estimación se trata de calcular posibles valores para parámetros de interés, por ejemplo, una elasticidad o el precio medio por metro cuadrado de la vivienda. En el contraste de hipótesis hay que establecer una hipótesis o conjetura específica sobre la población, por ejemplo, que no hay discriminación salarial por sexo o que el estado de un piso es un factor determinante de su precio, y analizar los datos para decidir si la hipótesis es correcta.

A.2.1. Estimación

El objetivo de la estimación es aproximar el valor de un conjunto de parámetros desconocidos de una distribución a partir de las observaciones muestrales de la misma. Denotaremos como θ a un parámetro desconocido y $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)'$ a un vector de K parámetros desconocidos. Un **estadístico** es una función de los datos, $g(y_1, \ldots, y_N)$. Un **estimador puntual** de θ es un estadístico que pretende ser un aproximación al parámetro desconocido y se denota por $\hat{\theta}$. Por ejemplo, la media de los datos puede ser un estimador de la media de una variable aleatoria y la varianza de los datos un estimador de su varianza. Es decir,

$$\hat{\mu} = \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad\qquad \hat{\sigma}^2 = S_y^{*2} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

Un estimador es una regla que está definida antes de que los datos se observen. El valor numérico que se obtiene al aplicarlo a los datos se denomina *estimación*. Por ejemplo, la estimación de la media del precio por metro cuadrado de un piso con la muestra de la Tabla 1.1 es:

$$\hat{\mu} = \frac{3,82 + 5,246 + \ldots + 3,434 + 4,20}{50} = 3,91 \text{ miles de euros}$$

Es decir, se estima que el precio de un piso oscila alrededor de 3910 euros/m^2. Sin embargo, ¿qué confianza podemos tener en este resultado? Por ejemplo, ¿valoraríamos igual esta cantidad si se hubiera calculado con una muestra de 5 observaciones? La respuesta obvia es NO, sino que consideramos más fiables los resultados con 50 datos que con 5. Por tanto, un estimador (y sus estimaciones) deben complementarse con una medida de su fiabilidad o precisión.

Un estimador es una variable aleatoria que depende de las variables Y_i, $i = 1, \ldots, N$. Su distribución de probabilidad se denomina distribución muestral o distribución empírica del estimador. En el ejemplo anterior, si $Y_i \sim NID(\mu, \sigma^2)$, entonces el estimador $\hat{\mu} = \bar{y}$ es una combinación lineal de N variables normales independientes, por lo que su distribución muestral es:

$$\hat{\mu} = \bar{y} \sim N(\mu, \sigma^2/N) \tag{A.2}$$

La media muestral se distribuye alrededor de la media poblacional y se concentra más probabilidad alrededor de μ cuanto mayor es N (es decir, menor es la varianza). Por tanto, hay mayor probabilidad de obtener una estimación cercana a μ con 50 datos que con $N = 5$. En este caso, es sensato utilizar como indicador de la *precisión* la desviación típica $\sigma/\sqrt{N}$: menor desviación típica indica mayor precisión. Normalmente, σ es desconocido, por lo que sustituimos su valor poblacional por el correspondiente muestral, S_y^*. La estimación de la desviación típica de la distribución muestral de $\bar{y}$,

$$\hat{\sigma}_{\bar{y}} = S_{\bar{y}} = S_y^*/\sqrt{N}$$

se conoce como *error típico* de $\bar{y}$. En el ejemplo del precio del m^2, obtenemos que el error típico de estimación es $0,993341/\sqrt{50} = 0,14$. Es fácil comprobar que si obtuviéramos los mismos valores de $\bar{y}$ y S_y con una muestra de 5 observaciones, el error típico se triplicaría, $S_{\bar{y}} = 0,993341/\sqrt{5} = 0,44$ miles de euros.

Ejercicio 5. Estimación de la media y la varianza del precio por m^2 de un piso.

1. Abre el fichero de datos de Gretl pisos.gdt.
2. Crea la variable precio por metro cuadrado, que denotaremos *pr_ m2*:
   a) Usa las opción *definir nueva variable* que está en el menú *Añadir* o en *Variable*.
   b) En la nueva ventana escribe *nombre de la nueva variable = fórmula*, es decir,
   $$pr\_\ m2 = precio/m2$$
3. Una vez creados los nuevos datos, las estimaciones de la media, m, y la desviación típica, S, se obtienen de la tabla de estadísticos descriptivos. La estimación de la varianza es el cuadrado de S. El error típico de estimación es $S/\sqrt{50}$.

Ejercicio 6: Estimación de media y varianza. Utilizando la opción de estadísticos descriptivos o estadísticos principales, obtén las medias y las desviaciones típicas de $z1$, $z2$, $x3$ y $x4$ generados en el ejercicio 1. Completa la siguiente tabla, incluyendo junto con los momentos poblacionales las estimaciones que has obtenido, es decir, correspondientes los momentos muestrales.

Modelo 1	$\mu =$	$\sigma =$
Muestra: $z1$	Estimación =	Estimación =
Modelo 2	$\mu =$	$\sigma =$
Muestra: $z2$	Estimación =	Estimación =
Modelo 3	$\mu =$	$\sigma =$
Muestra: $x3$	Estimación =	Estimación =
Modelo 4	$\mu =$	$\sigma =$
Muestra: $x4$	Estimación =	Estimación =

Criterios para comparar estimadores

Para un problema determinado existen distintos métodos de estimación y, obviamente, unos son mejores que otros. En algunos casos, distintos métodos pueden dar lugar a un mismo estimador de un parámetro. Es posible elegir entre distintos métodos de estimación basándonos en ciertas propiedades de la distribución muestral del estimador. En general, buscamos los estimadores que más se aproximen a los verdaderos valores. Así, exigimos que los estimadores cumplan una serie de propiedades basadas en una medida de la distancia entre θ y $\hat{\theta}$. En este curso nos fijamos en tres propiedades: insesgadez, eficiencia y el error cuadrático medio mínimo.

Insesgadez. Un estimador es insesgado si la media de su distribución empírica es el verdadero valor del parámetro, es decir,

$$E(\hat{\theta}) = \theta$$

Si se pudieran obtener todas las posibles realizaciones muestrales de $\hat{\theta}$, el promedio de todas estas estimaciones sería el valor del parámetro. Es una propiedad deseable porque indica que si un estimador es insesgado, el error de estimación, $\hat{\theta} - \theta$, se anula en promedio. Un ejemplo de estimador insesgado de la media poblacional de una distribución normal es $\bar{y}$, ya que de (A.2) tenemos que $E(\bar{y}) = \mu$. Un estimador insesgado de la varianza de una distribución es la varianza muestral, S^2.

En caso contrario, se dice que el estimador es sesgado. Se define el sesgo de un estimador como $Sesgo(\hat{\theta}) = E(\hat{\theta}) - \theta$. La parte izquierda del Gráfico A.8 representa las distribuciones de 3 estimadores de un mismo parámetro, θ: el estimador $\hat{\theta}_1$ es insesgado; $\hat{\theta}_2$, tiene sesgo negativo, es decir, en promedio subestima el valor del parámetro; finalmente el sesgo de $\hat{\theta}_3$ es positivo, es decir, este estimador en promedio sobrevalora el valor del parámetro.

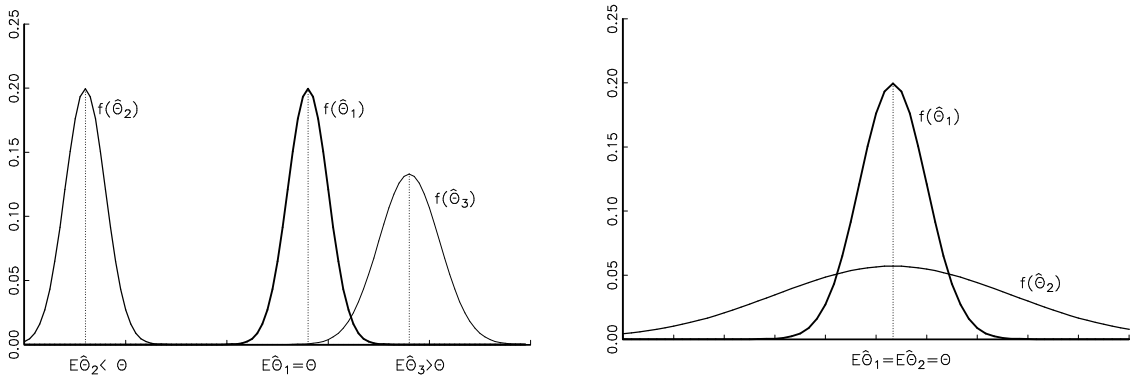


Gráfico A.8: Sesgo y varianza de estimadores

Eficiencia. Si nos fijamos únicamente en los estimadores insesgados, nos interesa establecer un criterio para elegir un estimador dentro de esta clase de estimadores. En la parte derecha del Gráfico A.8 se representa la distribución de dos estimadores, ambos insesgados. Claramente, el estimador con menor varianza, $\hat{\theta}_1$, tiene una probabilidad menor de obtener realizaciones *alejadas* del verdadero valor del parámetro. Por tanto, se considera que $\hat{\theta}_1$ supera al estimador $\hat{\theta}_2$ y se dice que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$.

En general, si un estimador es el que tiene menor varianza dentro de una clase de estimadores se dice que es el estimador *eficiente dentro de esa clase*. Así, se dice que un estimador $\hat{\theta}$ es eficiente dentro de la clase de estimadores insesgados si no hay otro estimador insesgado $\tilde{\theta}$ con una varianza menor:

$$var(\tilde{\theta}) \geq var(\hat{\theta}) \qquad \forall \tilde{\theta} \quad insesgado$$

Por ejemplo, la media de los datos es un estimador eficiente dentro de la clase de estimadores insesgados de la media poblacional μ de una variable normal. Es decir, se demuestra que, si $Y_i \sim NID(\mu, \sigma^2)$, $i = 1, \ldots, N$, entonces para todo estimador insesgado de μ, $\tilde{\mu}$ con $E\tilde{\mu} = \mu$:

$$var(\bar{y}) = \frac{\sigma^2}{N} \leq var(\tilde{\mu})$$

Si se trata de estimar un conjunto de K parámetros Θ, se dice que un estimador insesgado $\widehat{\Theta}$ es más eficiente que otro estimador insesgado $\tilde{\Theta}$ si la diferencia $[V(\tilde{\Theta}) - V(\widehat{\Theta})]$ es una matriz semidefinida positiva. Esto implica que cada elemento de $\widehat{\Theta}$ tiene una varianza menor o igual que el correspondiente elemento de $\tilde{\Theta}$.

Error cuadrático medio Aunque la insesgadez es una propiedad deseable, esto no implica que un estimador insesgado siempre sea preferible a uno sesgado. El Gráfico A.9 ilustra una situación en la que un estimador insesgado $\hat{\theta}_1$ puede descartarse frente a otro sesgado, $\hat{\theta}_2$. El estimador $\hat{\theta}_1$ tiene mucha varianza, por lo que tiene una probabilidad mayor de obtener errores de estimación más grandes que el estimador con menor varianza, $\hat{\theta}_2$, aunque este sea sesgado.

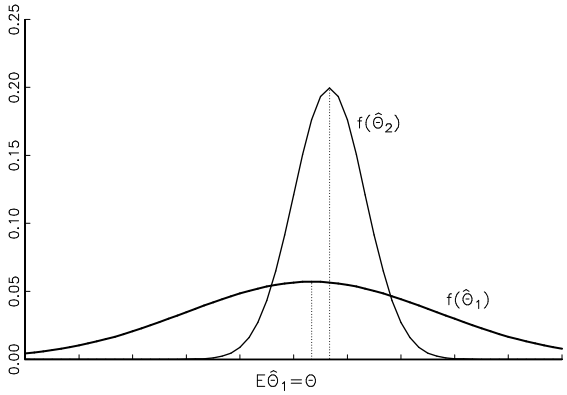


Gráfico A.9: Ejemplos de distribución de estimadores

Esto sugiere utilizar como criterio de elección de estimadores una medida del error del estimador. Se define el error cuadrático medio de un estimador:

$$ECM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = var(\hat{\theta}) + [sesgo(\hat{\theta})]^2$$

que se descompone en un término de varianza y otro de sesgo. Así, entre un conjunto de estimadores se elige aquel que tiene menor error cuadrático medio.

A.2.2. Contraste de hipótesis

Como ya se mencionó, uno de los objetivos de la Econometría es el de *contrastar hipótesis*. Por ejemplo, nos planteamos si los datos del precio del m^2 de la vivienda son compatibles con una determinada distribución con media 3000 euros/m^2. En un contraste de hipótesis se trata de establecer si la diferencia entre la hipotética media poblacional (en el ejemplo, 3000€) y la media muestral (3910€) se debe únicamente a la naturaleza aleatoria de los datos.

Un contraste de hipótesis tiene tres etapas (Ramanathan, 2002): (1) Formulación de dos hipótesis opuestas; (2) derivación de un estadístico de contraste y su distribución muestral; y (3) determinación de un criterio de decisión para elegir una de las dos hipótesis planteadas.

Una **hipótesis** estadística es una afirmación sobre la distribución de una o varias variables aleatorias. En un contraste se trata de decidir cuál, entre dos hipótesis planteadas, es la que mejor se adecúa a los datos. La hipótesis de interés se denomina **hipótesis nula**, H_0, mientras que la hipótesis frente a la que se contrasta se llama **hipótesis alternativa**, H_a. En el

ejemplo, consideramos que el precio del m^2 es una variable aleatoria normal y planteamos la hipótesis nula de que la media de Y sea igual a 3 (miles €) frente a la alternativa de que no lo sea, es decir,

$$H_0\colon \mu = 3 \qquad \text{frente a} \qquad H_a\colon \mu \neq 3$$

Normalmente, la hipótesis nula es una hipótesis simple, es decir, sólo se plantea un valor para μ. La hipótesis alternativa suele ser una hipótesis compuesta, que especifica un intervalo de valores. En el ejemplo, H_a es la negación de H_0 y se dice que es un contraste bilateral o *a dos colas*. Si la hipótesis alternativa se especifica H_a: $\mu < 3$, o bien H_a: $\mu > 3$, se dice que el contraste es unilateral o *a una cola*.

La elección entre las hipótesis se basa en un **estadístico de contraste**, que es una función de los datos que mide la discrepancia entre estos y H_0. Por ejemplo, en el contraste bilateral sobre la media, se define la siguiente medida de la discrepancia:

$$\frac{\bar{y} - 3}{S_{\bar{y}}}$$

Esta discrepancia, que utilizaremos como estadístico de contraste, no depende de las unidades de medida y tiene en cuenta la diferencia entre los datos (resumidos en $\bar{y}$) y el valor establecido en H_0. Además, debe conocerse la distribución de esta variable aleatoria cuando la hipótesis nula es correcta. En el ejemplo, se demuestra que si los datos $y_1, y_2, \ldots, y_N$ son una muestra aleatoria de un conjunto de variables $Y_i \sim NID(\mu, \sigma^2)\, \forall i$, con μ y σ^2 desconocidas, entonces:

$$\frac{\bar{y} - \mu}{S_{\bar{y}}} \sim t(N-1)$$

y sustituyendo $\mu = 3$, tenemos la distribución muestral del estadístico bajo H_0:

$$t = \frac{\bar{y} - 3}{S_{\bar{y}}} \overset{H_0}{\sim} t(N-1) \tag{A.3}$$

Este estadístico se aplica mucho en la práctica y se denomina estadístico t de la media.

Finalmente, para determinar **el criterio de decisión** del contraste se divide el conjunto de posibles resultados del estadístico de contraste en dos zonas, la **región crítica** y su complementaria. Se rechaza H_0 cuando el valor del estadístico obtenido con la muestra t^m pertenece a la región crítica. El punto de partida para establecer la región crítica es que se rechaza H_0 si la discrepancia entre datos y H_0 es *grande*. En el contraste bilateral, se rechazaría H_0 si $\bar{y}$ se alejara *mucho* del valor establecido en H_0, lo que para el estadístico implica que:

$$|t^m| = \left| \frac{\bar{y} - 3}{S_{\bar{y}}} \right| > c \tag{A.4}$$

donde c es la discrepancia máxima que estamos dispuestos a asumir y se denomina *valor crítico*. En caso contrario, si $|t^m| \leq c$, no se rechaza la hipótesis nula. El valor de c depende de la distribución del estadístico de contraste cuando H_0 es cierta y del error que estemos dispuestos a aceptar. En un contraste siempre existe la posibilidad de cometer los siguientes errores:

- Rechazar la hipótesis nula cuando ésta es cierta, que se llama error tipo I. El *nivel de significación* o *tamaño* de un contraste es la probabilidad de incurrir en el error tipo I y se denota por α.

- No rechazar la hipótesis nula cuando ésta es falsa, llamado error tipo II. La *potencia* de un contraste es la probabilidad de no cometer un error tipo II.

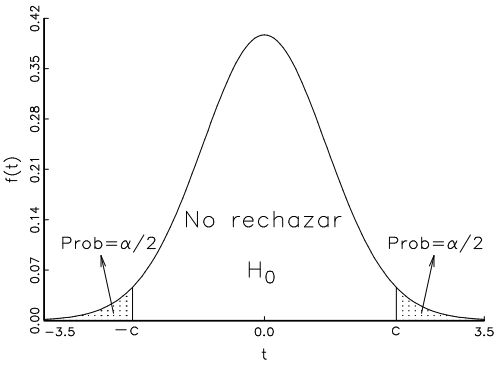
Deseamos cometer el menor error, pero no es posible eliminar los dos errores simultáneamente, es decir, que el tamaño sea 0 y la potencia igual a 1. En general, disminuir el error tipo I lleva consigo un aumento del error tipo II. Por ejemplo, no cometemos error tipo I si decidimos no rechazar nunca la hipótesis nula; pero la potencia del contraste sería 0 porque tampoco rechazaremos H_0 cuando sea falsa. Daremos más importancia al error tipo I, por lo que elegiremos el tamaño del contraste; los niveles más habituales son 10 %, 5 % y 1 %. Para el tamaño elegido, trataremos de utilizar el contraste con mayor potencia.

Ejemplo: zona crítica en un contraste bilateral sobre la media de una distribución normal. Veamos cómo se determina el valor crítico c en el ejemplo sobre la media del precio. El tamaño α es la probabilidad de rechazar H_0 cuando ésta es cierta. Como (A.4) es la condición para rechazar y (A.3) es la distribución del estadístico cuando H_0 es cierta, esto implica que:

$$\alpha = Prob(|t| > c) \qquad \text{cuando el estadístico } t \sim t(N-1)$$

En este caso, rechazaremos H_0 si el valor del estadístico t obtenido con los datos es un valor *poco probable* en la distribución del estadístico bajo H_0.

Este gráfico muestra la distribución del estadístico si H_0: $\mu = 3$ es cierta. La región crítica es la zona punteada en las **dos** colas de la distribución, de modo que en cada cola se acumula una probabilidad $\alpha/2$. Así, c es la ordenada de la distribución $t(N-1)$ que deja en la cola derecha una probabilidad $\alpha/2$. Por ejemplo, para $\alpha = 0,05$ y $N = 50$, entonces, $c = 2,01$ y se rechaza H_0 al nivel de significación del 5 % si $|t^m| > 2,01$.



Ejemplo 1: Contraste sobre la media del precio por m^2 en Gretl.
Suponiendo que la variable precio por metro cuadrado *pr_m2* sigue una distribución normal, contrasta H_0: $\mu = 3$ frente a H_a: $\mu \neq 3$. Los pasos son los siguientes:

1. Cálculo del valor muestral del estadístico $t = (\bar{y} - 3)/S_{\bar{y}}$, siendo $\bar{y}$ la media muestral de *pr_m2*:

$$t^m = \sqrt{50}(3,9144 - 3)/0,99341 = 6,51$$

Se obtiene con la siguiente opción de Gretl:

Herramientas →Calculadora de estadísticos de contraste

En la siguiente ventana elige la pestaña *media* y en ella:

- Marca la opción *Utilice una variable del conjunto de datos.*
- Selecciona la variable *pr_m2*. Aparecerán los estadísticos descriptivos que intervienen en el cálculo de t^m. En este caso:

media muestral:	3,9144
desv. típica:	0,99341
tamaño muestral:	50

- Escribe la hipótesis nula a contrastar: *H0: media =* 3.
- Comprueba que la opción *Suponer que la desv. típica es un valor poblacional* no está activada y pincha en *Aplicar*.

El resultado es la tabla y el Gráfico A.10. En el gráfico se representa la distribución del estadístico bajo H_0, en este caso $t(49)$, junto con el valor muestral del estadístico (la línea verde).

```
Hipótesis nula: media poblacional = 3    Tamaño muestral: n = 50
Media muestral = 3,91439, desv. típica = 0,993407
Estadístico de contraste: t(49) = (3,91439 - 3)/0,140489 = 6,50864
valor p a dos colas = 3,83e-008 (a una cola = 1,915e-008)
```

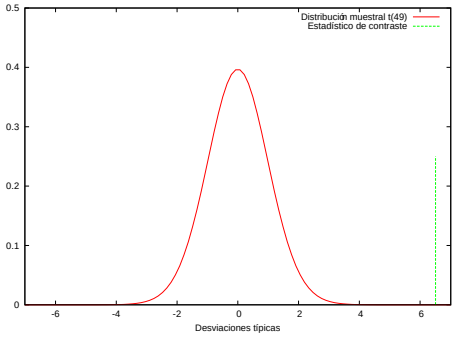


Gráfico A.10: Ejemplo 1: Resultado y distribución del estadístico bajo H_0

En este caso tenemos que el valor muestral del estadístico cae en la cola superior, en una intervalo de valores poco probable si H_0 es cierta. Por tanto, rechazaremos la hipótesis nula. Pero calcularemos exactamente la región crítica.

2. Región crítica o zona de rechazo. El valor crítico c se obtiene con la opción de Gretl *Herramientas* →*Tablas estadísticas*.

   En la nueva ventana hay que elegir la pestaña de la variable t y en el siguiente cuadro hay que rellenar:
   - *gl*= grados de libertad n, en este caso 49
   - probabilidad en la cola derecha = $\alpha/2$. Fijamos un nivel de significación del 5 %, por lo que escribimos 0,025.

   Tras pinchar en *Aceptar*, obtenemos el siguiente resultado:

```
t(49)     probabilidad en la cola derecha = 0,025
          probabilidad complementaria = 0,975
          probabilidad a dos colas = 0,05

          Valor crítico = 2,00958
```

   Interpretación: $Prob(t > 2,00958) = 0,025$ o bien $Prob(X < 2,00958) = 0,975$. Por tanto, el valor crítico con $alpha = 5\,\%$ es igual a $c = 2,00958$.

3. Aplicación de la regla de decisión. Como $|6,51| > c$, al nivel de significación del 5 %, se rechaza la hipótesis de que el precio medio sea igual a 3000€ frente a la alternativa. Cierra las ventanas de *calculadora de estadísticos* y *tablas estadísticas*.

Ejemplo: región crítica en el contraste unilateral sobre la media de una distribución normal. En los estudios econométricos a veces se plantean contrastes a una cola. Por ejemplo, en estudios sociales interesa analizar si hay discriminación salarial, de modo que las mujeres perciben salarios más bajos que los hombres. Habitualmente, se contrasta la hipótesis nula de que la media del salario que perciben las mujeres es igual al salario medio de los hombres frente a la hipótesis alternativa de que la media del salario es mayor en el grupo de hombres.

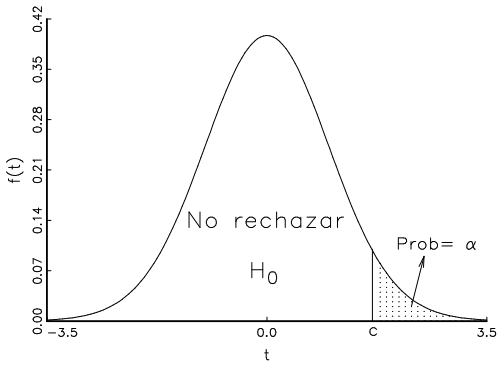
En el estudio del precio del m^2, supongamos que interesa contrastar si la media es tres o mayor, por lo que planteamos las hipótesis:

$$H_0 : \mu = 3 \qquad \text{frente a} \qquad H_a : \mu > 3$$

Al mantenerse la misma hipótesis nula, el estadístico de contraste es (A.3), $t = \sqrt{N}(\bar{y}-3)/S_y$, que bajo H_0 sigue una distribución $t(N-1)$. La hipótesis alternativa determina el criterio de decisión. Rechazaremos H_0 cuando la discrepancia tome valores alejados de H_0 y compatibles con H_a, es decir, cuando t tome valores positivos *grandes*. La región crítica está definida por la condición $t > c$. El valor crítico c se determina por:

$$\alpha = Prob(t > c) \qquad \text{cuando el estadístico } t \sim t(N-1)$$

La región crítica del contraste es la zona punteada en **una** cola de la distribución, la derecha. Así, c es la ordenada de la distribución $t(N-1)$ que acumula en la cola derecha una probabilidad α.
Por ejemplo, si $\alpha = 0,05$ y $N = 50$, entonces el nivel crítico es $c = 1,67655$ (usar herramienta de tabla estadística de Gretl) y no se rechaza H_0 al nivel de significación del 5 % si $t^m < 1,67655$.



En general, se usan las expresiones *rechazar* o *no rechazar* H_0. Esto es así porque en un contraste mantenemos la H_0 mientras no haya suficiente evidencia en contra. Los datos pueden rechazar la hipótesis, pero no pueden probar que H_0 sea correcta, por lo que no se dice que *se acepta* H_0. No rechazar H_0 significa que los datos no son capaces de mostrar su falsedad.

Ejemplo 2: Contraste de igualdad de varianzas. Los datos que estamos analizando sobre precio de la vivienda incluye dos tipos de viviendas:

- Viviendas a reformar, es decir, es necesario realizar un gasto adicional para acondicionar la vivienda.
- Viviendas acondicionadas para entrar a vivir.

Es posible que el precio medio de las viviendas a reformar y reformadas sigan

patrones diferentes. Esto implica que la distribución del precio de los dos tipos de vivienda es distinta. Por tanto, consideramos el siguiente modelo:

- El precio por metro cuadrado de la vivienda que no necesita reforma, Y_1 sigue una distribución normal de media μ_1 y varianza σ_1^2.
- El precio por metro cuadrado de la vivienda a reformar, Y_2 sigue una distribución normal de media μ_2 y varianza σ_2^2.
- Ambas variables Y_1 e Y_2 son independientes.

Vamos a contrastar si la varianza es la misma en ambas distribuciones frente a que sea menor en el grupo de pisos a reformar. Por tanto, planteamos el contraste de hipótesis:

$$H_0 \colon \sigma_1^2 = \sigma_2^2 \qquad \text{frente a} \qquad H_a \colon \sigma_1^2 > \sigma_2^2$$

El procedimiento de contraste consiste en comparar las dos varianzas muestrales, S_1^{*2} y S_2^{*2}, que son estimadores insesgados de las respectivas varianzas poblacionales. Valores cercanos de S_1^{*2} y S_2^{*2}, o ratios $S_1^{*2}/S_2^{*2} \simeq 1$, apoyan H_0. El estadístico de contraste y su distribución bajo H_0 son:

$$F = \frac{S_1^{*2}}{S_2^{*2}} \overset{H_0}{\sim} \mathcal{F}(N_1 - 1, N_2 - 1)$$

donde N_1 es el número de pisos que no necesita reforma y N_2 el número de pisos a reformar. Dada H_a, rechazamos H_0 si el ratio S_1^{*2}/S_2^{*2} está muy por encima de 1. La región crítica, por tanto, está definida por $S_1^{*2}/S_2^{*2} > c$, siendo c el valor crítico. Los pasos para realizar el contraste con Gretl son:

1. Seleccionar el subconjunto de pisos que no necesitan reforma. En el fichero de datos *pisos.gdt* son las observaciones para las que la variable *Reforma = 1*. En Gretl, seleccionamos la submuestra que cumple esta condición si:
   a) Vamos a *Muestra →Definir a partir de v. ficticia*.
   b) En la nueva ventana aparece como opción *Reforma* y pinchamos en *Aceptar*

   Si el proceso es correcto, en la parte inferior de la pantalla de *Gretl* aparece el mensaje *Sin fecha: rango completo n=50; muestra actual n=31*. Ahora sólo trabajamos con los datos de pisos que no necesitan reforma: si consultamos los datos en *Datos →Mostrar valores* ahora sólo aparece la información de los 31 pisos que pertenecen a esta clase.
2. Crear la serie de datos $y1$ que incluye únicamente los precios por m^2 de los pisos reformados: en *Añadir →Definir nueva variable...* escribimos *y1 = pr_m2*.
3. Seleccionar el subconjunto formado por los pisos que necesitan reforma, es decir, caracterizados por *Reforma = 0*:
   a) Vamos a *Muestra →Restringir, a partir de criterio*.
   b) En la nueva ventana escribimos el criterio de selección: *Reforma =0*
   c) Pinchamos en *Reemplazar restricción actual* y luego en *Aceptar*.

   Ahora debe aparecer *Sin fecha: rango completo n=50; muestra actual n=19*.
4. Crear la serie de datos $y2$ de precios por m^2 de pisos no reformados: en *Añadir →Definir nueva variable...* escribimos *y2 = pr_m2*.

5. Recuperar la muestra completa en *Muestra →Recuperar rango el completo*. Comprobamos que las series $y1$ e $y2$ no tienen errores editando los datos de estas series. Las celdas de $y1$ estarán vacías en pisos no reformados y lo recíproco para $y2$.

6. Calcular el valor muestral del estadístico F^m en *Herramientas →Calculadora de estadísticos de contraste →2 varianzas*. En la siguiente ventana rellenamos los datos:
   - Marcar *Utilice una variable del conjunto de datos* y seleccionar $y1$. Aparecen los estadísticos necesarios de $y1$: $S_1^{*2} = 0,77702$ y $N_1 = 31$
   - Marcar *Utilice una variable del conjunto de datos* y seleccionar $y2$. Aparecen los estadísticos necesarios de $y2$: $S_2^{*2} = 0,70340$ y $N_2 = 19$
   - Comprobar la marca en *Mostrar el gráfico de la distribución muestral* y *Aplicar*.

   El resultado es una tabla y un gráfico con la distribución del estadístico bajo H_0, $\mathcal{F}(30, 18)$ y el valor muestral del estadístico.

```
Hipótesis nula: Las varianzas poblacionales son iguales
Muestra 1: n = 31, varianza = 0,777054
Muestra 2:  n = 19, varianza = 0,703402
Estadístico de contraste: F(30, 18) = 1,10471
valor p a dos colas = 0,8436 (a una cola = 0,4218)
```

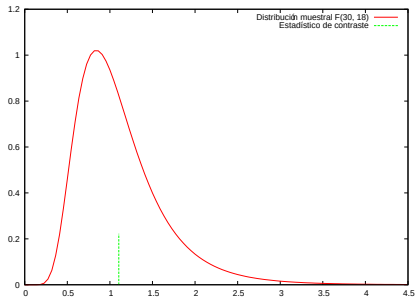


Gráfico A.11: Ejemplo 2: Resultado y distribución del estadístico bajo H_0

7. El gráfico anterior sugiere que no rechazaremos H_0. Calculamos la región crítica: se trata de un contraste a una cola, por tanto, buscamos c tal que $0,05 = Prob(F > c)$. Vamos a *Herramientas →Tablas estadísticas →F*.

   Los grados de libertad del numerador son *gln 30* y los del denominador, *gld 18*. Finalmente, la *probabilidad en la cola derecha* es 0,05. El resultado es:

```
F(30, 18)     probabilidad en la cola derecha = 0.05
              probabilidad complementaria = 0.95
              Valor crítico = 2.10714
```

   Por tanto, si $\alpha = 5\,\%$, entonces $c = 2,107$.

8. Conclusión del contraste: $F^m = 1,10 < 2,11$, por tanto, al nivel de significación del $5\,\%$ no rechazamos la hipótesis de igualdad de varianzas entre los dos tipos de viviendas.

Ejemplo 3: Contraste de igualdad de medias. Vamos a contrastar la hipótesis de que el precio medio del piso es mayor en los pisos reformados. Suponiendo que el precio por m^2 de los dos tipos de pisos son variables independientes, ambas con distribución normal de igual varianza, σ^2 y medias diferentes, μ_1 y μ_2.

Para contrastar la hipótesis anterior, planteamos H_0: $\mu_1 = \mu_2$ frente a H_a: $\mu_1 > \mu_2$.

El procedimiento de contraste se basa en la comparación de las dos medias muestrales, $\bar{y}_1$ y $\bar{y}_2$. Pequeñas diferencias entre ellas apoyan la H_0. El estadístico de contraste y su distribución bajo H_0 son:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{S\sqrt{1/N_1 + 1/N_2}} \overset{H_0}{\sim} t(N_1 + N_2 - 2)$$

donde S^2 es el estimador de la varianza común utilizando todos los datos:

$$S = \frac{1}{N_1 + N_2 - 2}\left(\sum_{i=1}^{N_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{N_2}(y_{2i} - \bar{y}_2)^2\right)$$

Dada H_a, rechazamos H_0 si la diferencia $\bar{y}_1 - \bar{y}_2$ es *grande*. La región crítica, por tanto, está definida por $t > c$, siendo c el valor crítico.

Aplicamos el procedimiento de contraste a los datos en Gretl. Las dos series de datos $y1$ e $y2$ se crean según lo descrito en el ejemplo 2. A continuación debemos:

1. Calcular el valor muestral del estadístico t^m en *Herramientas →Calculadora de estadísticos de contraste →2 medias*. En la siguiente ventana rellenamos los datos:
   - Marcar *Utilice una variable del conjunto de datos* y seleccionar $y1$. Aparecen los estadísticos de $y1$: $\bar{y}_1 = 4,3040$, $S_1^* = 0,88150675$ y $N_1 = 31$
   - Marcar *Utilice una variable del conjunto de datos* y seleccionar $y2$. Aparecen los estadísticos de $y2$: $\bar{y}_2 = 3,278717$, $S_2^* = 0,83869$ y $N_2 = 19$
   - Marcar *Suponer desviación típica poblacional común.*
   - Marcar *Mostrar el gráfico de la distribución muestral* y pinchar en *Aplicar.*

   El resultado es una tabla y un gráfico con la distribución $t(50-2)$ y el valor muestral del estadístico.

```
Hipótesis nula: Diferencia de medias = 0
Muestra 1: n = 31, media = 4,304, d.t. = 0,881507
        desviación típica de la media = 0,158323
        Intervalo de confianza 95% para la media: 3,98066 a 4,62734

Muestra 2: n = 19, media = 3,27872, d.t. = 0,838691
        desviación típica de la media = 0,192409
        Intervalo de confianza 95% para la media: 2,87448 a 3,68295

Estadístico de contraste: t(48)= (4,304-3,27872)/0,252229 =4,0649
valor p a dos colas = 0,0001774 (a una cola = 8,871e-005)
```

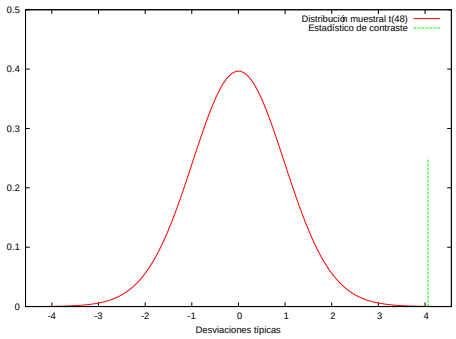


Gráfico A.12: Ejemplo 3: Resultado y distribución del estadístico bajo H_0

2. Definir la región crítica: se trata de un contraste a una cola, por tanto, buscamos c tal que $0,05 = Prob(t > c)$. Vamos a *Herramientas* $\rightarrow$*Tablas estadísticas* $\rightarrow t$, grados de libertad *gl 48* y para $\alpha = 5\,\%$, obtenemos $c = 1,229$.

3. Resultado del contraste: $4,06496 > 1,229$, por tanto, al nivel de significación del $5\,\%$ rechazamos la hipótesis nula de igualdad de medias. Es decir, los datos apoyan la hipótesis de que el precio del m^2 es mayor en los pisos reformados.

Bibliografía

Peña, D. y J. Romo (1997), *Introducción a la Estadística para las Ciencias Sociales*, McGraw-Hill.