Ideas sobre Inferencia Estadística

(Basado parcialmente en *La Estadistica: Una Orquesta Hecha Instrumento* (1996), Jaime Llopis Pérez, Ed. Ariel)

Hasta ahora se han tratado los conceptos de teoría de la probabilidad y se han construido modelos, es decir, sustitutos matemáticos del mundo real. Así, se ha participado en un juego paralelo al de la realidad, pero totalmente abstracto.

La distribución normal, la uniforme, la Bernoulli, etc., con sus momentos y sus funciones de densidad, son *abstracciones* que sustituyen a los verdaderos objetivos de la investigación: las **poblaciones**.

A partir de ahora empezamos lo que se denomina *Inferencia Estadística*. La inferencia estadística es la ciencia mediante la cual se estudian propiedades de un *todo* a partir del análisis de una pequeña parte de este *todo*. El problema básico de la estadística es hacer afirmaciones acerca de poblaciones, entendido el concepto de población en un sentido mucho más amplio del que se tiene en geografía, por ejemplo.

En estadística, todas las personas de una ciudad, los diabéticos de un país, son posibles poblaciones. Así, una población se puede decir que es una agrupación de entidades reunidas según un determinado criterio. No es algo necesariamente homogéneo, claramente delimitado. Una población puede ser, por ejemplo, todas las personas que llevan gafas en una población. Los que llevan gafas no están separados de los demás. Están mezclados.

Todas las afirmaciones acerca de las poblaciones deben hacerse, de la forma más coherente posible, a partir de la elaboración y la canalización de la información de disponible en una muestra. Hay que elaborar estrategias para dar el salto de la muestra a la población, para responder a preguntas que nos hagamos sobre la población y, además, para responderlas con la única información que es la depositada en la muestra.

Como en todas las empresas complejas, hay distintas *filosofías* sobre inferencia estadística. Hay diferentes opciones para hacer este tránsito de la muestra a la población. Nosotros nos moveremos en una de estas filosofías posibles, que es la filosofía **frecuentista**.

En esta filosofía la palabra clave es la de muestra. Todo lo que puede hacerse en inferencia estadística, según esta filosofía, debe quedar circunscrito a lo que pueda hacerse con lo que

tenemos en la muestra.

Se estudiará el concepto de muestra; o mejor dicho, la doble cara del concepto de muestra. También se verá la doble cara del concepto de estadístico. Esta doble cara de ambos conceptos es, sin lugar a dudas, el núcleo fundamental sobre el que se edifica la inferencia estadística.

La muestra, en el sentido más popular, es un repertorio de n valores. Se diría en cierto modo de n *escasos* valores. Un estadístico, como, por ejemplo, una media muestral, es un valor, un simple número. Si esperamos deducir cosas de la población únicamente con esto realmente no lo tenemos fácil.

Estos n valores de la muestra y el valor único del estadístico adquieren otro aspecto si ponemos en marcha toda una potente teoría que se basa en la concepción de una muestra como n variables aleatorias y la de un estadístico como una variable aleatoria.

Se estudiarán, fundamentalmente, herramientas para la selección de los modelos probabilísticos, puesto que sus usos ya se han visto en la parte de cálculo de probabilidades. Se ha visto cómo manejar los modelos, cómo describirlos, cómo usarlos en tanto que sustitutos de la realidad, pero no se han visto procedimientos para, dada una situación real determinada, elegir el modelo más adecuado.

Si en teoría de la probabilidad la preocupación principal es la elaboración de modelos probabilísticos vistos como maquinarias que pueden sustituir a las poblaciones reales, en estadística la preocupación primordial es el buen uso de estos modelos probabilísticos. La estadística es un proceso o camino que comienza con la percepción de una situación real problemática y la posterior elección de una familia de modelos probabilísticos, mediante la selección de la medida más apropiada, entre los representantes de la familia para responder así a las cuestiones que tengamos planteadas sobre aquella situación real.

El concepto de muestra tiene un doble sentido:

1. Por un lado son n observaciones independientes obtenidas de la población de la que se pretende decir cosas.

2. Por otro lado son n variables aleatorias independientes e idénticamente distribuidas.

La muestra, según la primera concepción, son n valores. Son las n observaciones que hace el experimentador de la variable y de la población que pretende conocer.

La muestra, según la segunda concepción, son n variables aleatorias. Es, de hecho, una descripción teórica de todas las muestras posibles.

Ejemplo. Supongamos una población con la distribución siguiente:

$$P(X = 1) = 0{,}5 \qquad P(X = 2) = 0{,}5$$

Si podemos tomar una muestra de tamaño dos, el experimentador obtendrá una de las cuatro que son posibles:

(1,1) (1,2) (2,1) (2,2)

Estas son las únicas observaciones posibles. Para el experimentador, una muestra será una de estas cuatro, la que tenga en sus manos después de la experiencia.

Pensad que él sólo tiene una, no las cuatro. Ésta es la concepción de muestra como n observaciones. Para la otra concepción, la muestra, en realidad, es una descripción de todas las muestras posibles. En este caso es el conjunto formado por las cuatro parejas de valores. Es el listado de todas las muestras posibles. Es el conjunto de las formas en que una población puede aparecer en las muestras. En este caso hemos podido describir todas las muestras posibles porque se trata de una población con poca diversidad. Habrá casos en los que esta descripción sólo podrá hacerse de una forma teórica, mediante las variables aleatorias: n variables aleatorias independientes e idénticamente distribuidas.

Consideramos ahora el concepto de estadístico. Un estadístico es una función muestral. Es decir, es un resumen de una muestra: Es una función de $\mathbb{R}^n$ sobre $\mathbb{R}$. Equivale a asignar un número a n números.

Este concepto tiene una doble cara. Por un lado, puede ser un número, y por otro puede ser una variable. Esta doble concepción va ligada a la doble concepción que tiene la muestra. Si la muestra la vemos como n valores entonces el estadístico será *un número*. Si la vemos como n variables, el estadístico será una variable, para cada valor de la muestra tendremos un valor del estadístico, por lo tanto el estadístico será variable según la muestra y tendrá una distribución que mostrará un esquema de su variabilidad.

Aunque generalmente al hablarse de estadísticos se piensa en los que popularmente más suelen usarse, como la media muestral, la varianza muestral, etc., existen muchos tipos distintos de funciones muestrales que pueden definirse a partir de una muestra.

Es como si, atendiendo a su estructura, los pudiéramos dividir en diferentes familias. Las posibilidades que hay, en cuanto a formas de asignar números a muestras, es enorme. Existen tipos de estadísticos muy sofisticados: cualquier procedimiento mediante el cual, a partir de una muestra, se obtenga un número es un estadístico.

La muestra vista como un vector de n variables aleatorias tiene su distribución. Como se trata de variables independientes e idénticamente distribuidas, si se sabe la familia a la que pertenece la distribución de la población de donde se toma la muestra, es fácil llegar a la distribución muestral. Se trata de calcular el producto de las distribuciones marginales:

$$f(x_1, \ldots, x_n; \theta) = f(x_1; \theta) \ldots f(x_n; \theta)$$

La función de densidad muestral es, pues, por la independencia de las variables implicadas en una muestra, el producto de las funciones de densidad marginales y, en realidad, por ser variables idénticamente distribuidas, es el producto n veces de la función de densidad de la población de donde se ha obtenido la muestra.

De una población nos interesa definir números que la caractericen, números que nos ayuden a establecer algún rasgo peculiar, que resuelvan dudas acerca de la población y que permitan hacer predicciones.

Estos números pueden ser, por ejemplo, la esperanza, la varianza, la proporción de individuos dentro de un intervalo, la proporción de individuos por encima de un cierto valor. El objeto de la inferencia estadística será establecer técnicas para obtener y estudiar números como los anteriores, mediante una estimación puntual, o una estimación por intervalos o mediante contrastes de hipótesis.

Definiciones

- **Inferencia**: Proceso de obtención de información sobre valores desconocidos de la población a partir de valores muestrales.
- **Parámetro**: Valor desconocido de la población que queremos aproximar a partir de valores de una muestra.
- **Estadístico**: Una función de la información contenida en la muestra.
- **Estimador**: Variable aleatoria que depende de la información de la muestra y cuyos valores aproximan el valor del parámetro de interés.
- **Estimación**: Un valor concreto del estimador para una muestra dada.

Ejemplo

Estimar el gasto familiar medio anual en alimentación en una región a partir de una muestra de 200 familias

- El **parámetro de interés** sería el valor promedio de dicho gasto en la región.
- Un **estadístico relevante** en este caso sería la suma de los gastos de todas las familias en la muestra.
- El **estimador** más razonable sería el promedio del gasto familiar en la muestra.
- Si para una muestra concreta el promedio de gasto en alimentación es de 3.500 euros, la **estimación** del gasto medio anual en la región sería de 3.500 euros.

La inferencia estadística es la ciencia que trata de descubrir, con la máxima coherencia posible, características de poblaciones usando la información de una pequeña parte de ellas (muestras).

En inferencia estadística hay dos pasos fundamentales:

1. La elección de la familia de modelos probabilísticos.
2. La elección de un modelo probabilístico concreto dentro de la familia.

Los tres mecanismos fundamentales de la inferencia estadística son:

Estimación puntual, estimación por intervalos y contrastes de hipótesis.

Dada una situación real, elegir un modelo probabilístico para describir la variabilidad o la incertidumbre de esa situación, es un paso muy importante. Si la aproximación del modelo a la realidad es bueno, podemos sustituir la realidad por el modelo, que se puede manejar de una manera cómoda: es como un juguete que podemos manipular y emplear con la finalidad de conocer mejor la determinada realidad.

Ante una situación de variabilidad debido al azar, nos preguntamos por algún aspecto de dicha variabilidad o algún número que la caracterice. Por ejemplo, el promedio, la dispersión, la proporción de individuos dentro de un intervalo de valores, etc.

Ideas sobre estimación puntual

Mediante la estimación puntual tratamos de aproximarnos a los números que caracterizan a una población, tomando como valores de dichos números el resultado de unos cálculos efectuados a una muestra.

Por tanto, se utiliza un estadístico (función calculada a partir de una muestra) para descubrir los valores poblacionales.

La estimación puntual es un procedimiento mediante el cual, a través de un estadístico, se hace un pronóstico o una estimación de algún número interesante sobre el modelo de una población.

Una característica de la población puede ser una variable aleatoria y puede interesar algún resumen de la misma. Por ejemplo:

$$E(X) \qquad Var(X) \qquad \frac{E(X)}{Var(X)} \qquad P\left(X \geq 0\right) \qquad \cdots$$

Estos números son poblacionales y por lo tanto inaccesibles, pero en realidad, son números relacionados con el modelo que hemos utilizado como sustituto de la población, de modo que calculando valores en el modelo es como si lo estuviéramos calculando en la población. Por supuesto que todos los parámetros de las diferentes distribuciones son números poblacionales que nos interesaran. Incluso, posiblemente, el pronóstico de muchos de estos números que nos interesan pasa por el previo pronostico de los valores poblacionales de los parámetros. Esto ultimo es, sencillamente, porque cuando tenemos los valores de los parámetros lo tenemos todo, probabilísticamente, de aquella poblaci6n.

Mediante la estimación puntual construimos estadísticos, maquinarias, que resumen los valores de la muestra y que lo hacen orientados hacia alguno de los números que nos pueden interesar de la población. Una vez construido el estadístico y una vez comprobado que tiene una calidad aceptable, tomamos como valor poblacional el valor que nos da el estadístico. Aquí aparece, de nuevo, la doble cara del concepto de muestra y de estadístico.

Sabemos que el número buscado, a nivel poblacional, es solo uno: el promedio poblacional (la esperanza) es una, la probabilidad de encontrarnos con individuos por debajo de un cierto valor es solo una, etc. Pero sabemos también que el valor que nos da el estadístico es variable.

Variable según la muestra. Muestras distintas nos darían valores distintos del estadístico. Por lo tanto, vamos a dar como valor poblacional el valor que nos da el estadístico. Nos creeremos que el valor poblacional buscado es el que sale de la máquina que es el estadístico. Antes de usar, pues, un estadístico, necesitamos comprobar la calidad de esta maquinaria en tanto que estimador. Si sabemos que tiene buena calidad como estimador, nuestra apuesta por él la podemos hacer con tranquilidad, porque esto es como una auténtica *apuesta*.

Pero, ¿qué criterios hay que seguir para valorar la esperanza de un estadístico usado como estimador?

En primer lugar, el *sesgo*. Un estimador es insesgado si su esperanza, vista como variable aleatoria, es igual al valor buscado. Si no coincide, se dice que el estimador es sesgado y a lo que excede del valor buscado se Ie denomina *sesgo*. Algo que tenga sesgo significa que está desviado de algún tipo de objetivo. Un estimador tiene sesgo si su esperanza no es el valor que está tratando de estimar, si está apuntando mal a su objetivo.

Pensad que lo estimado es un número y lo estimamos mediante una variable. La variable tiene variabilidad, tiene su distribución, por lo tanto parece coherente, parece un buen criterio de calidad, pedir que el promedio del estadístico, la esperanza del estadístico, sea, justo, el número que pretendemos estimar, el número al que pretendemos aproximarnos.

Por ejemplo, supongamos que una población sigue la siguiente distribución normal $N(\mu, 1)$ Supongamos, también, que mediante una muestra de tamaño n, pretendemos estimar el parámetro que nos falta por conocer: μ.

Es muy importante tener en cuenta que cualquier estadístico es candidato a ser estimador de cualquier número. Cualquiera. Otra cosa es que lo haga con buena o mala calidad. Veamos algunos candidatos:

$$T_1 = \bar{X}_n$$

$$T_2 = X_1$$

$$T_3 = \frac{X_1 + X_2}{2}$$

$$T_4 = 5$$

$$T_5 = \max_{0 \leq i \leq n} X$$

La esperanza de los tres primeros es el valor que buscamos.

La esperanza de T_1 (la media muestral) es μ. La esperanza de T_2 es la esperanza de una variable aleatoria que siga la distribución $N(\mu, 1)$, por lo tanto es μ.

La esperanza de T_3 es la esperanza del promedio de dos variables normales $N(\mu, 1)$, que es también μ.

La esperanza de T_4 es el valor que buscamos únicamente si ese valor es 5.

0 sea, si μ es igual a 5, entonces la esperanza de T_4 coincide con μ; en caso contrario, no. Por lo tanto, tiene sesgo, en general.

En el caso de T_5 no puede ser su esperanza el valor buscado, puesto que el promedio del máximo de la muestra siempre estará sensiblemente por encima del valor promedio.

Parece coherente pedirle a un estadístico, usado como estimador, que apunte bien, que en promedio nos proporcione el valor buscado. Así, este es un primer criterio de calidad para valorar un estimador. Es lógico exigirle eso a una maquinaria que la vas a tomar como guía para afirmar cosas sobre una población.

En el ejemplo anterior, pues, siguiendo este criterio, los tres primeros estadísticos son buenos como estimadores de μ. Pero ¿son los tres iguales de buenos? Si tuvierais que elegir, ¿cual elegiríais?

En principio, ninguno de los tres tiene sesgo, son los tres insesgados, sería mas económico usar el estadístico T_2 porque no necesitas hacer cálculos, incluso con muestras de tamaño uno te bastaría, no es necesario coger n observaciones de la población. Pero la varianza no será la misma en los tres estimadores y éste es un segundo y fundamental criterio de calidad: la varianza de un estimador. La varianza de un estimador es su variabilidad visto como variable aleatoria.

En el ejemplo que venimos analizando, está claro que la variabilidad será mayor cuanto menos información se utilice de la muestra.

En este caso el estimador elegido será la media muestral porque la varianza de T_2 es igual a 1, la de T_3 es igual a 1/2 y la de T_1 es $1/n$, que es la menor para $n \geq 2$.

Es coherente pedirle a un estimador que su variabilidad sea pequeña. Vamos a creernos que el valor poblacional es el valor que nos proporciona el estadístico que es un estimador insesgado, y además con poca variabilidad (varianza). Esto no dará lugar, en general, a grandes errores, es decir, no se alejará excesivamente del objetivo.

En el caso de T_4 su varianza es cero lo cual parecería perfecto pero la esperanza no es siempre el valor buscado, lo es solo en un caso: si la μ es igual a 5. Por lo tanto, un estimador debe tener como esperanza el valor buscado la varianza lo más pequeña posible.

En ocasiones los estimadores no son insesgados. pero si asintóticamente insesgados. Esto significa que cuando el tamaño de la muestra aumenta, el sesgo va haciéndose cada vez mas pequeño, tendiendo a cero cuando la n tiende a infinito. Esto sucede, por ejemplo, con la distribución uniforme continua $U(a, b)$. Para estimar el parámetro a es coherente usar el mínimo de la muestra. Pero este estimador **no** es insesgado. Para que su esperanza fuera a deberían darse valores muestrales por debajo de a, y esto es imposible. Sin embargo, sí es asintóticamente insesgado. Si el tamaño de la muestra aumenta, el promedio de valores del estadístico se irá aproximando a a. En el límite sera igual a a.

NOTA: ¿Cómo puede valorarse si tiene o no sesgo un estimador de μ si no sabes el valor de μ? En el caso de la normal, no sabíamos el valor de μ, ni falta que nos hacía. Lo importante no es cuanto valga la μ, sino que valga lo que valga se obtenga la μ de la población normal.

Un estimador sin varianza sería lo ideal, pero, sin embargo, eso no es posible entre los estimadores insesgados, porque existe una varianza mínima para un estimador insesgado de un parámetro. Es un peaje mínimo que hay que pagar si uno quiere aproximarse mediante un estadístico a un parámetro de una distribución. Se trata de la llamada cota de *Cramer-Rao*. La cota de *Cramer-Rao* es el mínimo de los ruidos que se generara si uno quiere estimar un parámetro a partir de la maquinaria de un estadístico.

Un estadístico para el que se alcance la cota será un estadístico **eficiente**: un estadístico

muy interesante, puesto que al hecho de ser insesgado se le habrá de sumar el hecho de tener la menor de las varianzas posibles. Por ejemplo, mediante la cota de Cramer-Rao se puede afirmar que el mejor estimador de la μ de una normal es la media muestral. Es decir, hay mecanismos establecidos para averiguar cuál es la varianza mínima de un estadístico sin sesgo.

En resumen, para estimar un valor desconocido de una población se trata de escoger un estadístico sin sesgo y con la menor varianza posible. Esto nos hace esperar que será una elección correcta que no nos desviará del valor verdadero.

Una manera razonable de medir la efectividad de un estimador es usando el *error cuadrático medio*, que mide tanto el sesgo como la variabilidad del estimador:

$$E\left[\left(\hat{\theta}-\theta\right)^2\right] = E\left[\left(\hat{\theta}-E\left(\hat{\theta}\right)+E\left(\hat{\theta}\right)-\theta\right)^2\right] =$$

$$E\left[\left(\hat{\theta}-E\left(\hat{\theta}\right)\right)^2\right]+\left(E\left(\hat{\theta}\right)-\theta\right)^2 =$$

$$Var\left(\hat{\theta}\right)+\left(Sesgo\left(\hat{\theta}\right)\right)^2$$

Un estimador será tanto más efectivo cuanto menor sea el error cuadrático medio.

Si no se conoce un estimador con buenas propiedades existen procedimientos generales para definir estimadores con propiedades razonables:

- Método de máxima verosimilitud
- Método de los momentos

Ideas sobre Intervalos de Confianza

Mediante la estimación puntual tratábamos de aproximamos a ciertos números que caracterizan a una población, arriesgándonos a dar como valores de dichos números el resultado de unos cálculos efectuados a unas muestras mediante unos estadísticos. En la estimación por intervalos nuestra intención será proporcionar un intervalo donde sea muy probable que pueda encontrarse el verdadero valor del número buscado.

Es una postura más conservadora que la estimación puntual. No se aventura a dar un número. Da muchos. La clave aquí será dar un intervalo lo más pequeño posible, claro. Al menos, lo importante será dar un intervalo de tamaño equilibrado entre lo que desconocemos y lo que conocemos de la variabilidad de las estimaciones de lo que buscamos. Cuando hay mucha variabilidad, es más difícil hacer pronósticos.

Dado un modelo, la estimación por intervalos de algún número de interés del modelo es la formulación de un intervalo donde, al afirmar que en su interior está el verdadero valor de dicho número, habremos acertado con una probabilidad fijada de antemano. Construir intervalos es tarea aparentemente sencilla. Si queremos crear un intervalo para la media de alturas de la población adulta de españoles, haciendo un intervalo, por ejemplo, como el siguiente: (1,50, 1,80), tenemos muchas probabilidades, si no todas, de acertar. Pero no se trata de hacer cualquier intervalo, sino de hacer un intervalo lo más pequeño posible. Nos interesa construir una expresión del siguiente estilo:

$$P\left(T_1 \leq g(\theta) \leq T_2\right) = 1 - \alpha$$

donde la función paramétrica sea la expresión, en función de los parámetros de la distribución, del número buscado a nivel poblacional, y donde T_1 y T_2 son dos estadísticos y la α un número pequeño elegido por nosotros. Este número acostumbra a ser 0,01, 0,05 ó 0,1.

Antes de ver cómo se construyen estos intervalos, se presenta una definición importante: **Pivote**.

Un *pivote* es una función de la muestra y del valor poblacional que buscamos (al cual podemos escribir perfectamente como una función del parámetro), cuya distribución no depende del parámetro de la distribución. O sea, un pivote será una función:

$$h\left(X_1, \ldots, X_n, g(\theta)\right)$$

cuya distribución, como variable aleatoria que es, no dependa del parámetro.

Un pivote es, en realidad, una función muestra1 donde, además, tenemos incorporada la expresión de un número poblacional que nos interesa. Como función muestral que es, es una variable aleatoria y por lo tanto tiene su distribución. Si la conocemos y es independiente del parámetro de la distribución de la población, entonces tenemos, en una expresión muestral, incorporada, la función paramétrica de la que queremos decir cosas y donde su distribución no depende de cosas que no sepamos. De esta forma, nos será posible conseguir aislar lo desconocido entre dos valores conocidos.

Vamos a ver cómo se construyen intervalos de confianza a través de pivotes, y lo veremos a través de un par de ejemplos. Dos ejemplos, además, típicos, que nos ayudarán a captar la esencia de este procedimiento. Veamos el primero:

Intervalo de confianza de la media de una población con distribución normal

Si la población es normal, entonces, a partir de una muestra de tamaño n, podemos conseguir el siguiente pivote:

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{S_n}$$

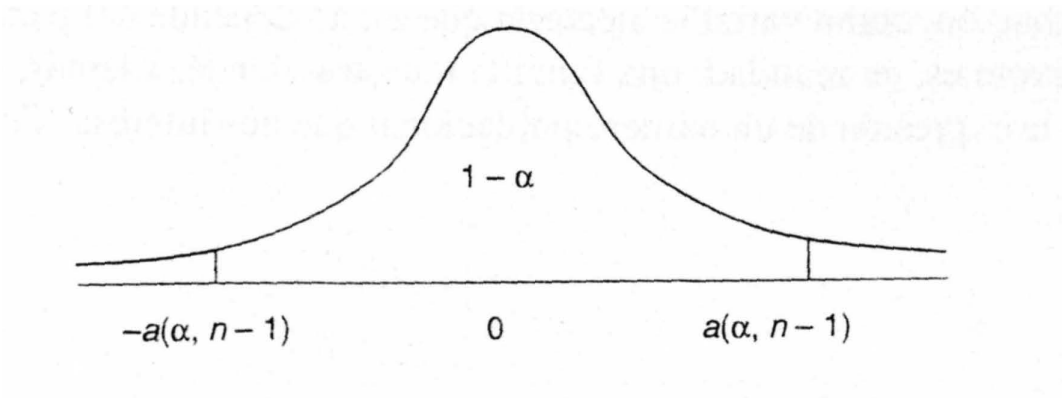
La distribución de esta función muestral es una *t de Student* con $n-1$ grados de libertad, valga lo que valga la media poblacional. Por eso es un pivote. Esto se puede deducir a través del teorema de *Fisher*. El parámetro de la distribución normal está aquí como de *espectador* únicamente. No influye en la distribución. Por eso es un pivote. Se observa que es una función de la muestra y del parámetro μ, pero que su distribución no depende de μ.

Como sabemos la distribución de este pivote, podemos encontrar, sin problemas mediante unas tablas, los números que nos permitan escribir la siguiente expresión

$$P\left(-t_{n-1,\alpha/2} \leq \frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{S_n} \leq t_{n-1,\alpha/2}\right) = 1-\alpha$$

Escribo $t_{n-1,\alpha/2}$ indicando que es un número que, en la distribución t de Student, entre él y su negativo, nuestro pivote, como variable aleatoria que es, tiene una probabilidad igual a $1-\alpha$.

Mirad el dibujo



La expresión $t_{n-1,\alpha/2}$ nos está indicando que este número depende de $\alpha/2$ y del tamaño muestral, en concreto de $n-1$, que es el valor del parámetro de la t de Student.

Observad que en esta expresión todo es conocido excepto la media poblacional, por lo tanto, para poder obtener la expresión buscada se trata simplemente de ir aislando dicha media entre las desigualdades. Al final se obtiene la importante expresión:

$$P\left(\bar{X}_n - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\right) = 1-\alpha$$

Por lo tanto, tenemos que los dos estadísticos que buscábamos son:

$$T_1\left(X_1,\ldots,X_n\right) = \bar{X}_n - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}$$

$$T_2\left(X_1,\ldots,X_n\right) = \bar{X}_n + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}$$

Observaciones:

1. E1 intervalo se construye partiendo de la media muestral como punto de referencia. El intervalo se abre a partir de la media muestral. A ésta se le suma y se le resta el mismo valor.

2. Cuanto mayor sea el tamaño de la muestra, de más información disponemos y por lo tacto más estrecho será el intervalo. Observad que la n está en el denominador de lo que

determina el radio del intervalo, o sea la expresión:

$$t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}$$

por lo que, si la n es grande, entonces el intervalo se hace más estrecho.

3. Cuanta mayor variabilidad haya en la población, más imprecisión, lógicamente, a la hora de decir cosas acerca de la propia población, lo que conllevará intervalos más amplios. Observad que S_n, que es una estimación de la desviación típica de la población, está en el numerador de la expresión que es el radio del intervalo. Por lo tanto. cuanta mayor variabilidad poblacional, menor precisión en la estimación por intervalos: o sea, mas grandes los intervalos.

Fijaos que son completamente lógicas estas observaciones que se pueden hacer analizando, por dentro, la fórmula del intervalo.

Observad que usar estos intervalos es muy sencillo. Si tenemos una muestra de una población que hemos comprobado que es normal, simplemente tenemos que encontrar el valor de $t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}$ en las tablas y calcular la media muestral y la varianza. Si los valores los introducimos en un software estadístico él hará lo mismo que os acabo de explicar.

Veamos la construcción del intervalo del segundo ejemplo:

Intervalo de confianza de la varianza de una población con distribución normal

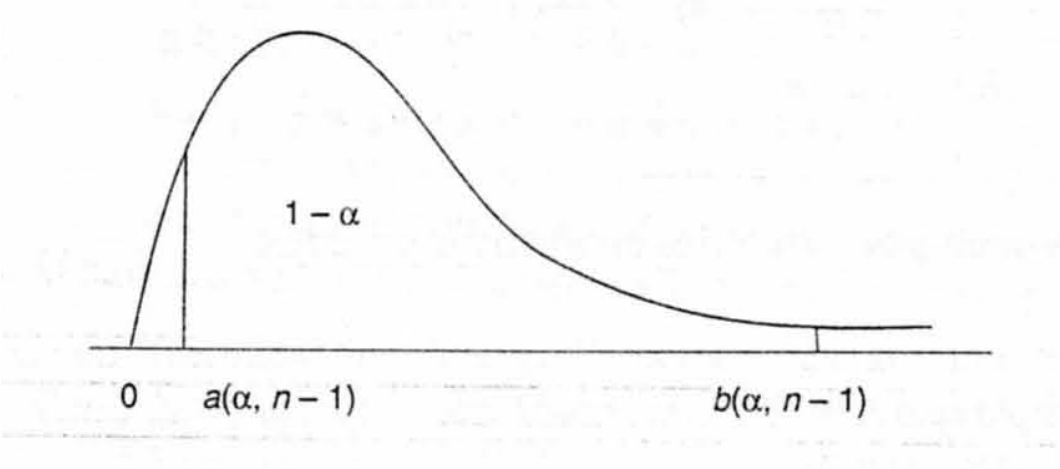
A partir de una muestra de tamaño n de una población normal podemos definir el siguiente pivote:

$$\frac{(n-1)S_n^2}{\sigma^2}$$

Esta expresión sigue una distribución ji-cuadrado, sea la que sea la varianza de la población normal. Esto se puede obtener también del teorema de Fisher. Podemos, como antes lo hemos hecho con la distribución t de Student, encontrar dos valores que permitan escribir la siguiente igualdad:

$$P\left(\chi_{n-1,1-\alpha/2} \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi_{n-1,\alpha/2}\right) = 1-\alpha$$

Mirad el dibujo



Escribo ahora $\chi_{n-1,1-\alpha/2}$ y $\chi_{n-1,\alpha/2}$, indicando que son dos números que en la distribución ji-cuadrado, entre los dos, nuestro pivote, como variable aleatoria que es, tiene una probabilidad de $1-\alpha$. Se observa que, como ocurría antes, las expresiones $\chi_{n-1,1-\alpha/2}$ y $\chi_{n-1,\alpha/2}$ nos están indicando que estos números dependen de α y del tamaño muestral, en concreto de $n-1$, que es el valor del parámetro de la ji-cuadrado.

Observad que aquí todo es conocido excepto la varianza poblacional. por lo que podemos iniciar, como hemos hecho para la media. un proceso de aislamiento de la varianza, hasta llegar a la expresión:

$$P\left(\frac{(n-1)S_n^2}{\chi_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{\chi_{n-1,1-\alpha/2}}\right) = 1-\alpha$$

A la hora de usar este intervalo en un caso real, basta, como en el caso de la μ, buscar los valores de $\chi_{n-1,1-\alpha/2}$ y $\chi_{n-1,\alpha/2}$ en la tabla de la ji-cuadrado, y luego calcular la varianza muestral y sustituir, finalmente, en la fórmula. O bien entrarlos en un software estadístico, que hará todo lo que acabo de deciros para daros los dos valores entre los que está el parámetro, con una probabilidad $1-\alpha$.

En ocasiones no es posible construir intervalos como los anteriores. simplemente porque se desconoce la distribución de la población o de los estadísticos utilizados para la construcción de los intervalos. Es posible que puedan construirse unas aproximaciones, usando el teorema central del límite (TCL). Es decir, todo lo que hemos hecho ha sido suponiendo que la población de

donde se coge la muestra puede modelarse mediante una distribución normal. Si esto no sucede, ¿qué ocurre? Pues que los pivotes, los dos pivotes que hemos visto, no siguen una distribución ni t de Student el primero, ni ji-cuadrado el segundo. Esto quiere decir que aquellos números que buscábamos para la construcción de los intervalos ya no son válidos. Sin embargo, si la distribución de la población no es la distribución normal y si el tamaño de la muestra es grande, entonces, por el *TCL*, al hacer muchas operaciones (tanto la media muestral como la varianza muestral son el fruto de bastantes operaciones con variables) acabas estando en unas condiciones de normalidad, al menos aproximada. Es decir, según el *TCL*, sumar muchas variables distintas acostumbra a dar como resultado una variable que sigue distribución normal.