INTRODUCCIÓN A LA INFERENCIA ESTADÍSTICA

(INTRODUCCIÓN.-MUESTREO.-DISTRIBUCIONES MUESTRALES)

---

1. Introducción

1.1 Inferencia Estadística

Inferir es, en general, establecer un nuevo conocimiento partiendo de uno ya "dado". La inferencia estadística va a ser una forma especial de realizar este proceso. Consiste, básicamente, en determinar algunas características desconocidas de una población partiendo de datos muestrales conocidos. Estas características poblacionales serán "inferidas" utilizando los recursos de la *TEORÍA MATEMATICA DE LA PROBABILIDAD.*

Fundamentalmente la Inferencia Estadística consiste en la resolución de dos grandes categorías de problemas:

*** ***LA ESTIMACIÓN***: Determinar el valor de una característica poblacional desconocida. Podrá ser:

- **Por punto**: Determinación de un valor poblacional concreto
- **Por intervalo: Determinación** de un intervalo en el que quede incluido el valor de la característica con cierto grado de probabilidad.

*** ***EL CONTRASTE DE HIPOTESIS***: Determinar si es aceptable, partiendo de los datos muestrales, que la característica poblacional estudiada tome un valor determinado o bien que pertenezca a un intervalo de valores determinado. (Es obvio que estos dos

problemas de conocimiento pueden, muy bien, considerarse como dos tipos particulares de problemas de decisión estadística y así, de hecho, lo considera una de las escuelas metodológicas de la Estadística)

1.2 Conceptos básicos.

POBLACION: Colectivo sujeto del estudio .Cabe distinguir entre **Población** (colectivo en el que estamos considerando la magnitud sujeta a estudio) y **Universo** (colectivo de todos los elementos sujetos del estudio, en el que no consideramos la magnitud). El *universo* es, por tanto, el conjunto de individuos que poseen la característica o características sujetas a estudio, y éstas en su conjunto forman la *población*

Así; Analizando las estaturas de los españoles, la población sería el conjunto de todas las estaturas de todos los españoles, siendo el universo el conjunto de todos los españoles.

MUESTRA: Un subconjunto cualquiera de la población. Para que la muestra nos sirva para extraer conclusiones sobre la población deber ser **representativa**, lo que se consigue seleccionando sus elementos al azar, lo que da lugar a una muestra aleatoria

MUESTREO: Procedimiento para la obtención de una muestra

MUESTREO OPINATICO: es aquel procedimiento de selección de los elementos muestrales que se realiza según el criterio del investigador. Es, por tanto, subjetivo y la muestra obtenida puede no ser representativa de la población.

MUESTREO ALEATORIO: es aquel procedimiento de selección de la muestra en el que todos y cada uno de los elementos de la población tiene una cierta probabilidad de resultar elegidos. De esta forma, si tenemos una población de N elementos y estamos interesados en obtener una muestra de n elementos (muestra de tamaño n), cada subconjunto de n elementos de la población tendrá también una cierta probabilidad de resultar la muestra elegida.

Si designamos por M_i a cada uno de estos subconjuntos, con i= 1,2,3,...N;

cada M_i tendrá una cierta probabilidad P(Mi) de resultar elegido.

MUESTREO ALEATORIO SIMPLE: **(M.A S.)** :es aquel muestreo aleatorio en el que la probabilidad de que un elemento resulte seleccionado se mantiene constante a lo largo de todo el proceso de obtención de la misma . La técnica del muestreo puede asimilarse a un modelo de extracción de bolas de una urna con devolución (*reemplazamiento)* de la bola extraída. Un mismo dato puede, en consecuencia, resultar muestreado más de una vez .Cada elección no depender de las anteriores y, por tanto, los datos muestrales serán ***estocásticamente independientes***.

MUESTREO IRRESTRICTO (**SIN REEMPLAZAMIENTO):** en este tipo de muestreo la probabilidad de obtener un dato en cada selección viene influida por los resultados anteriores , en la medida en que en este muestreo no permitimos que un mismo dato sea seleccionado más de una vez (lo que hace variar las probabilidades en

cada extracción muestral). Se corresponde con un modelo de extracción sin reemplazamiento. Teniendo en cuenta la convergencia de la distribución hipergeométrica a la binomial es fácil intuir que cuando la población sea muy grande ($N \to \infty$) el muestreo irrestricto puede considerarse como muestreo aleatorio simple.

Por tanto, en el estudio de muestras para poblaciones grandes consideraremos sólo el muestreo simple .En el estudio de muestras de poblaciones finitas es, sin embargo, fundamental analizar las distribuciones muestrales que generará su adecuado muestreo irrestricto)

MUESTRA GENERICA DE TAMAÑO n: Es una ***variable aleatoria***

n-dimensional ; $X=[x_1 ,x_2 ,x_3,...,x_n]$ donde cada x_j (con j=1,2,...n)

(cada dato muestral genérico) recorre todos los posibles valores que puede tomar el j-simo elemento de una muestra de n elementos.

Por tanto, una muestra concreta (ya obtenida) será un valor particular (una realización concreta) de la muestra genérica.

En la medida en que en el muestreo aleatorio cada elemento de la población tiene una probabilidad de ser elegido, cada dato muestral genérico será una variable aleatoria que tendrá asociada una función de probabilidad f(x) (de cuantía o de densidad) según una determinada distribución que llamaremos **distribución básica**, madre, o, simplemente, **distribución de la población** y recorrerá todos los posibles valores de la población.

Si trabajamos con un muestreo aleatorio simple (M.A.S.), cada dato muestral genérico será estocásticamente independiente de los demás y por tanto la función de probabilidad (cuantía o densidad) conjunta de la muestra genérica será:

$f(x) = f (x_1 ,x_2 ,x_3 ,x_4 \ldots\ldots ,x_n) = f(x_1)\cdot f(x_2)\cdot f(x_3).....f(x_n)$

por ser las x_j variables aleatorias independientes.

ESTADÍSTICO: Es cualquier función de los valores muestrales que dependa exclusivamente de estos. En la medida en que los valores muestrales son variables aleatorias también lo serán las funciones de éstos: los estadísticos.

A modo de ejemplo podemos decir que son estadísticos la media muestral, la varianza muestral , la cuasivarianza muestral , dado que son funciones de valores muestrales exclusivamente y no sería estadístico la función $\frac{nS^2}{\sigma^2}$ que si bien contiene la varianza muestral , también depende de la poblacional σ^2 y por tanto no es función exclusiva de la muestra.

Como hemos visto, los estadísticos son variables aleatorias por lo que tendrán determinas distribuciones de probabilidad y determinados parámetros (media, varianza, etc.) .Para el desarrollo de la inferencia es imprescindible conocer dichas distribuciones

y parámetros, consiguiendo establecer entonces las relaciones entre éstas y las de la población, pudiendo entonces inferir las características desconocidas de ésta.

Tras un breve recorrido por las técnicas de muestreo pasaremos a desarrollar las distribuciones de probabilidad de los principales estadísticos.

1.3. Técnicas de muestreo

Es evidente que un conocimiento previo por parte del investigador de las características de la realidad de la población mejora o debe mejorar los resultados inferenciales que se pueden obtener de la obtención de una muestra; parace claro que si bien el método de selección aleatoria conlleva los mejores resultados, quizá el adecuar la manera de extraer la muestra a las posibles distintas naturalezas de las poblaciones puede mejorar el rendimiento, aunque sólo fuere a nivel de coste. No es por tanto lo mismo intentar conocer la altura media de los habitantes de un país, que el número de errores en una gran contabilidad, dado que la naturaleza de su universo y por tanto el comportamiento poblacional son distintos. Es por ello, que para distintas "naturalezas" del problema han de plantearse distintas soluciones, si bien todas, o casi todas, pasan por la aleatoriedad; de ahí que se establezcan diversas "técnicas" o "métodos" de muestreo, de los que brevemente enumeramos algunos.

1.3.1. Muestreo aleatorio sistemático. Esta técnica consiste en extraer elementos de la población mediante una regla *sistematizadora* que previamente hemos creado (sencillamente cada K elementos). Así; numerada la población, se elige (aleatoriamente) un primer elemento base, partiendo de éste se aplica la regla para conseguir los demás hasta conseguir el tamaño muestral adecuado. Este procedimiento conlleva el riesgo de dar resultados sesgados si en la población se dan periodicidades o rachas.

1.3.2. Muestreo aleatorio estratificado. Consiste en considerar categorías típicas diferentes entre sí (estratos) que poseen una gran homogeneidad interna (poca varianza interna) y no obstante son heterogéneos entre sí (mucha varianza entre estratos) . La muestra se distribuye (se extrae de) entre los estratos predeterminados según la naturaleza de la población (ejemplo: sexo, lugar geográfico, etc.). Dicha distribución-reparto de la muestra se denomina afijación; que puede ser de varias formas:

*afijación simple: a cada estrato le corresponde igual número de elementos (extracciones) muestrales.

*afijación proporcional: la distribución se hace de acuerdo con el peso (tamaño) relativo de cada estrato.

*afijación óptima: Se tiene en cuenta la previsible dispersión de los resultados, de modo que se considera la proporción y la desviación típica.

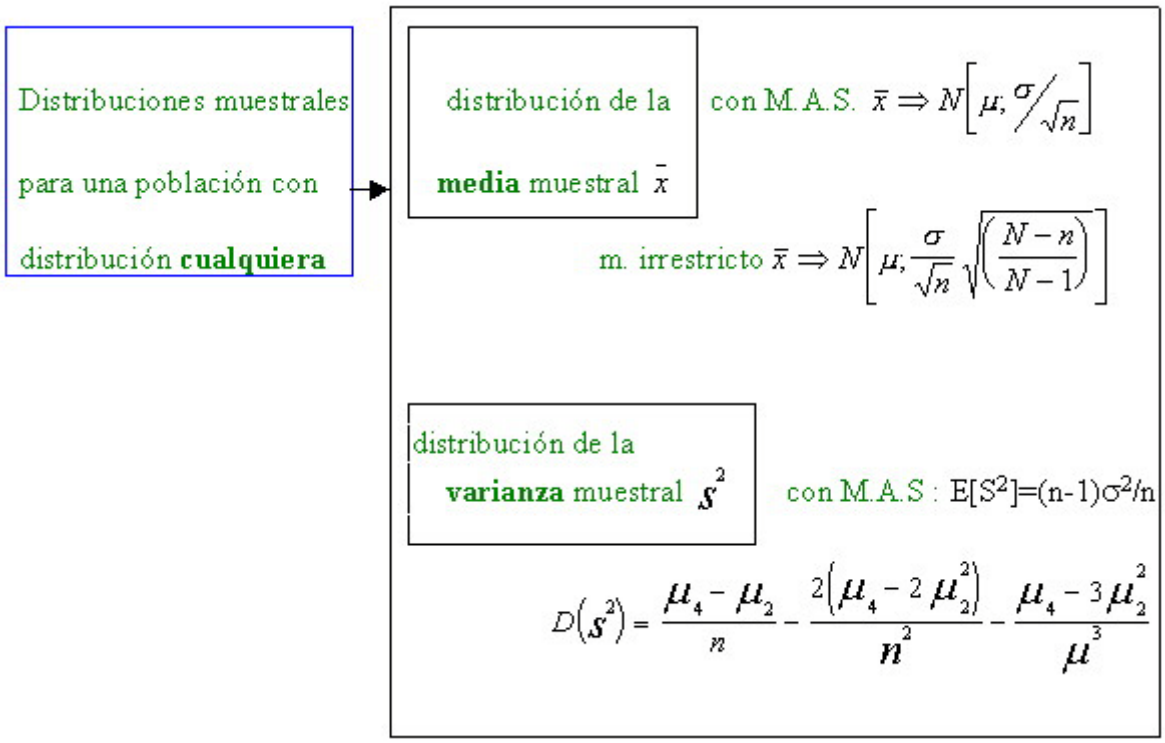
1.3.3Muestreo por conglomerados. La unidad muestral es un grupo de elementos de la población que forman previsiblemente una unidad de comportamiento representativo. Dicha unidad es el conglomerado cuyo comportamiento interno puede ser muy disperso (varianza grande) pero que presumiblemente poseerá un comportamiento próximo a otros conglomerados (varianza entre conglomerados, pequeña). Los conglomerados se estudian en profundidad hasta conseguir el tamaño muestral adecuado.

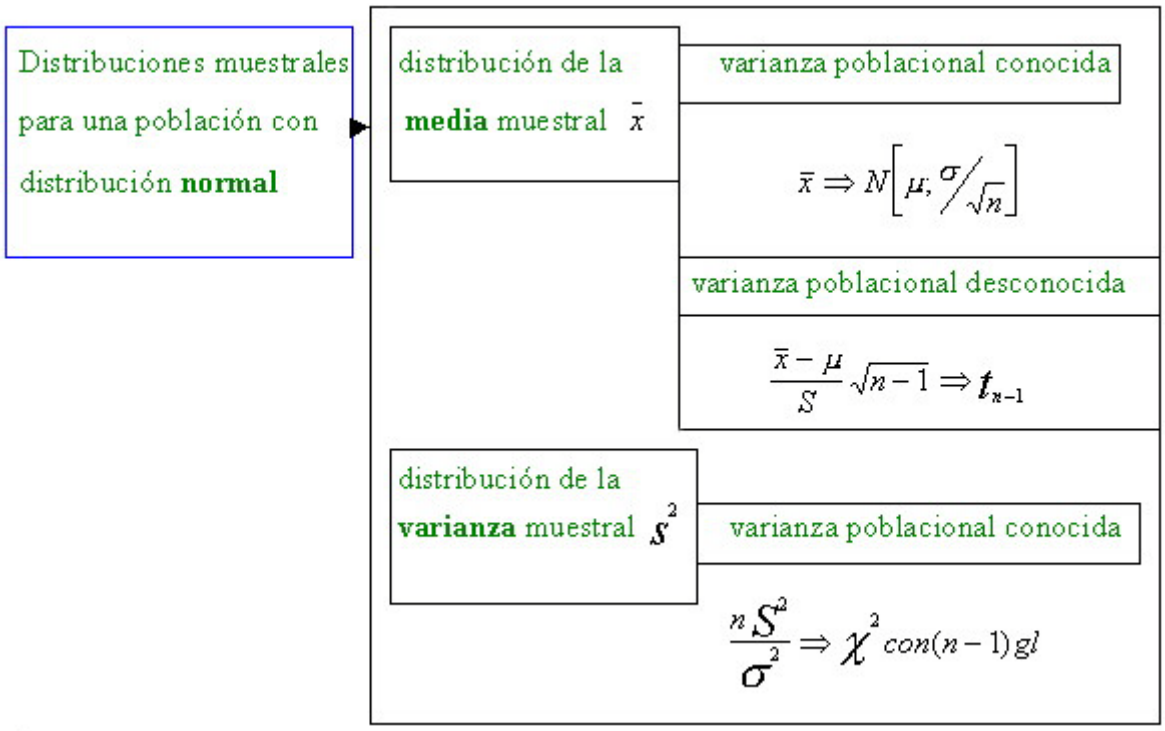
1.3.4.Muestreo por unidades monetarias .Este tipo de muestreo es especifico en auditoría , viene a solucionar el problema que plantea la selección aleatoria de partidas contables que no tienen (evidentemente) el mismo monto económico y por ello en un muestreo estrictamente aleatorio se "primaría" la inspección de las numerosas partidas pequeñas irrelevantes dejando sin inspección las importantes y cuantiosas. Para solucionarlo el M.U.M plantea la selección aleatoria no de asientos o partidas sino de unidades monetarias (ordenadas y numeradas) de tal manera que el defecto anterior se subsana al tener una partida cuantiosa más probabilidades de ser elegida pues contiene más unidades monetarias.

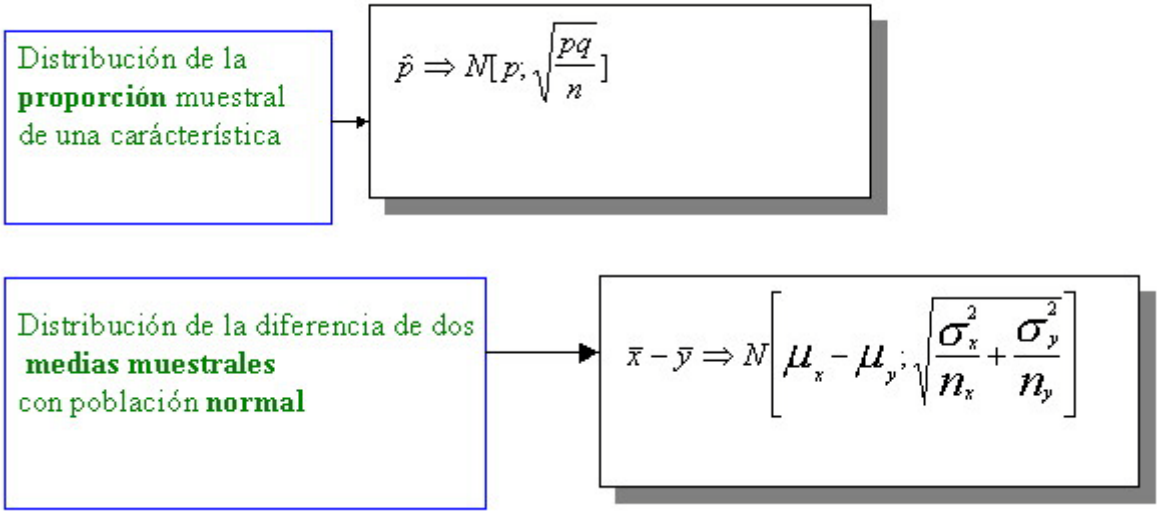
1.3.5. Otros tipos de muestreo. Es evidente que los planteados no son las únicas técnicas de muestreo. Existen otras como las no aleatorias: Cuotas, Intencional, Incidental, bola de nieve, etc. Y otras aleatorias y complicadas como el muestreo por superpoblaciones, y que en este curso no podemos desarrollar.

2. Distribuciones en el Muestreo de los principales Estadísticos.

Esquemáticamente pueden plantearse algunos de los posibles escenarios que darán origen a las diversas distribuciones muestrales que después desarrollaremos.







Una vez esquematizadas las distribuciones de algunos de los principales estadísticos muestrales pasamos a desarrollar algunas de ellas

2.1 Distribuciones muestrales para una población *cualquiera*

Si desconocemos la distribución de la población no podemos, evidentemente, conocer la distribución de la muestra genérica de tamaño n y, en consecuencia, no podremos llegar

a conocer la distribución de los estadísticos. Pero sí se podrá, en cualquier caso, determinar los principales parámetros (esperanza y varianza) de las principales distribuciones maestrales (de estadísticos) en función de los parámetros de la distribución de la población (media poblacional, varianza poblacional.)

2.1.1. Distribución de la media muestral

La esperanza de la media muestral coincide con la media de la población, en definitiva; la media de la media muestral es la media de la población:

$$E\left[\bar{x}\right]=\mu$$

en efecto :

$$\bar{x}=\sum_{i=1}^{n}x_1/n=\frac{x_1}{n}+\frac{x_2}{n}+\ldots\ldots\ldots+\frac{x_n}{n}$$

así :

$$E[\bar{x}]=1/n\,E[x_1]+1/n\,E[x_2]+\ldots\ldots\ldots\ldots+1/n\,E[x_n]$$

dado que para cualquier valor de i $E[\bar{x}_i]=\mu$ ya que x_i pertenece a la población

tendríamos que

$$E[\bar{x}]=1/n\,\mu+1/n\,\mu+\ldots\ldots\ldots\ldots+1/n\,\mu=n/n\,\mu=\mu$$

En cuanto a la varianza de la media muestral , tendremos que si el muestreo utilizado es aleatorio simple se cumple que

$$D^2[\bar{x}]=\sigma^2/n$$

en efecto :

$$D^2[\bar{x}]=(1/n^2)\,D^2[x_1]+(1/n^2)\,D^2[x_2]+\ldots\ldots\ldots\ldots+(1/n^2)\,D^2[x_n]$$

dado que en el muestreo aleatorio simple las observaciones o elementos son independientes tendremos covarianzas iguales a cero y dado que :

para todo i

$$D^2[x_i]=\sigma^2$$

tendremos

$$D^2[\bar{x}]=(1/n^2)\,\sigma^2+(1/n^2)\,\sigma^2+\ldots\ldots\ldots\ldots+(1/n^2)\,\sigma^2=n/n^2\,\sigma^2=\sigma^2/n$$

evidentemente la desviación típica será

$$D[\bar{x}] = \sigma/\sqrt{n}$$

En el caso de que el muestreo que hayamos realizado no sea aleatorio simple y que sea irrestricto y por tanto se plantee que no hay reemplazamiento siendo la población finita la media de la media muestral no sufrirá variaciones, pero no así la varianza de la media muestral que se verá afectada por el "**coeficiente corrector de poblaciones finitas**" (C .C .P.F.), o "**coeficiente de exhaustividad**", ya conocido del estudio de la distribución hipergeométrica. Así la varianza de la media muestral quedaría:

$$D^2[\bar{x}] = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

siendo N el tamaño de la población (finita) y el resto lo habitual

Dado que ya conocemos la media y la varianza de la media muestral , y dado que podríamos tomar la muestra genérica como una sucesión de variables aleatorias independientes de media y varianza conocida , aunque con distribución desconocida , y en aplicación del T.C.L. , tendremos que la ley de la media muestral sea cual sea la distribución poblacional viene dada por :

$$\bar{x} \Rightarrow N[\mu, \frac{\sigma}{\sqrt{n}}]$$

2.1.2. Distribución de la varianza muestral

Al igual que la media muestral la varianza muestral tendrá media y varianza dado que se trata también de una variable aleatoria.

La media de la varianza muestral con M.A.S. es:

$$E[s^2] = \left(\frac{n-1}{n}\right)\sigma^2 = \left(1-\frac{1}{n}\right)\sigma^2$$

la varianza muestral tiene de expresión

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

si cambiamos de origen a la media de la población la varianza muestral no varía así :

$$s^2 = \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n} - (\bar{x}-\mu)^2$$

aplicando el operador esperanza tendremos :

$$E[S^2] = E[\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}] - E[(\bar{x} - \mu)^2] =$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2}{n} + \cdots + \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n} = \left(1 - \frac{1}{n}\right)\sigma^2$$

dado que

$$E[(x_i - \mu)^2] = \sigma^2$$

y también

$$E[(\bar{x} - \mu)^2] = D^2[\bar{x}] = \frac{\sigma^2}{n}$$

En cuanto a la varianza de la varianza muestral diremos que su expresión es la siguiente, que admitimos sin demostrar que:

$$D(s^2) = \frac{\mu_4 - \mu_2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} - \frac{\mu_4 - 3\mu_2^2}{\mu^3}$$

2.2 Distribuciones muestrales para una población *normal*

De todas las posibles distribuciones básicas es, sin duda, la distribución normal la más importante por el gran número de poblaciones que se distribuyen así, real o asintóticamente, (en virtud de los Teoremas Límite).

Así pues, en los subapartados siguientes, consideraremos que conocemos la distribución de la población y que , ésta, es normal. Consideraremos igualmente muestreo aleatorio simple (m.a.s.)

2.2.1. Distribución de la media muestral

Si la población se distribuye N[μ ;σ] entonces

$$\bar{x} \Rightarrow N\left[\mu, \sigma/\sqrt{n}\right]$$

en efecto si

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

y dado que

$$\forall i \ldots x_i \Rightarrow N[\mu, \sigma]$$ siendo independientes pues

realizamos m.a.s. y en aplicación del teorema fundamental de las distribuciones normales obtendremos

$$\bar{x} \Rightarrow N\left[\mu, \sigma/\sqrt{n}\right]$$

2.2.2. Distribución de la varianza muestral

En lugar de obtener la distribución muestral del estadístico varianza muestral

L [S^2] que nos llevaría a conclusiones próximas a las anteriormente descritas en el apartado en el que la población no era normal, es más conveniente la utilización de la variable aleatoria $\frac{nS^2}{\sigma^2}$ que recordemos, no es un estadístico, y que contiene en su expresión a la varianza muestral y a la poblacional, de ahí su utilidad dado que ambas quedan relacionadas con una distribución conocida; la jhi-dos.

No demostramos la relación pero la recordamos dada su importancia posterior.

$$\frac{nS^2}{\sigma^2} \Rightarrow \chi^2 con(n-1)gl$$

2.2.3 Distribución de la media muestral con varianza desconocida.

En apartados anteriores estudiamos el comportamiento de la media muestral y vimos que ésta dependía tanto del valor de la media poblacional, como de la varianza poblacional, parece lógico pensar que si nuestro interés radica en inferir comportamientos de la población partiendo de la muestra parece ilógico pensar que conozcamos la varianza. De ahí la importancia de establecer una distribución para la media muestral que la relacione únicamente con la poblacional, lo que hará que conocida la muestral concreta podamos aventurar el comportamiento de la poblacional.

Así tendríamos:

$\bar{x} \Rightarrow N\left[\mu, \sigma/\sqrt{n}\right]$ lo que da lugar a :

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \Rightarrow N[0;1]$$

hemos visto sin demostrar que

$$\frac{nS^2}{\sigma^2} \Rightarrow \chi^2 con(n-1)gl$$

conocemos que

$$t_n = \frac{N[0;1]}{\sqrt{\frac{\chi_n^2}{n}}}$$

luego

$$\frac{\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{nS^2}{\sigma^2}}{n-1}}} \Rightarrow t_{n-1}$$

simplificando tendríamos

$$\frac{\bar{x} - \mu}{S}\sqrt{n-1} \Rightarrow t_{n-1}$$

expresión que relaciona ambas medias y la varianza muestral con una distribución conocida

2.3. Distribución de la proporción muestral de una característica.

Por su importancia incluimos esta distribución si bien podría considerarse un caso particular de la media muestral con distribución poblacional desconocida.

Así dada una población que por sus características consideramos binomial pues nuestra intención es inferir la proporción de éxitos p y por tanto la población sería B (N,p) podemos considerar cada realización muestral x_i una D(p) cuya media sería p y su varianza pq

la variable

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

dado que las x_i son independientes y en aplicación del TCL tendríamos

que

$$\hat{p} \Rightarrow N[p; \sqrt{\frac{pq}{n}}]$$

2.4. Distribución de la diferencia de dos medias muestrales de dos poblaciones normales.

Por último y dado el gran número de intervalos y contrastes que emanan de la utilización de esta distribución creemos necesario incluirla. Así:

Dadas dos variables X e Y que se distribuyen normalmente y por tanto:

$$X \Rightarrow N[\mu_x; \sigma_x] \quad \text{e} \quad Y \Rightarrow N[\mu_y; \sigma_y]$$

realizado un MA.S de tamaños n_x y n_y respectivamente tendremos por conocido que

$$\bar{x} \Rightarrow \left[\mu_x; \sigma_x/\sqrt{n_x}\right] \quad \bar{y} \Rightarrow \left[\mu_y; \sigma_y/\sqrt{n_y}\right]$$

por lo que la variable

$$Z = \bar{x} - \bar{y}$$

n aplicación del teorema fundamental de las distribuciones normales será :

$$\bar{x} - \bar{y} \Rightarrow N\left[\mu_x - \mu_y;; \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right]$$