ESTADÍSTICA

BLOQUE III

1. Introducción a la Inferencia e Inferencia No paramétrica

Índice

1. INFERENCIA Y CONCEPTOS BÁSICOS

La Inferencia estadística es aquella parte de la Estadística que pretende obtener conclusiones para toda una **población** a partir del estudio de las características de una **muestra.** Utiliza para ello un conjunto de técnicas y herramientas concretas.

Problemas que resuelve la Inferencia:

1. Estimación de parámetros poblacionales (estimación puntual y por intervalos)
2. Contrastes de hipótesis acerca de:
   - Parámetros poblacionales
   - Otras características de la distribución poblacional

3

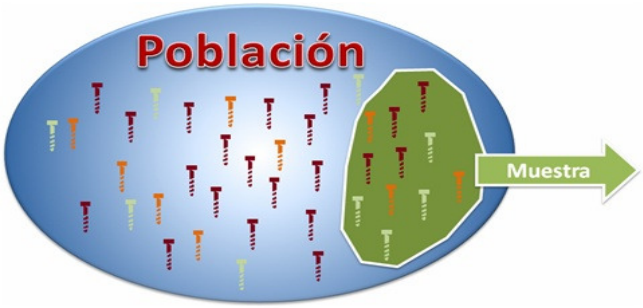
Conceptos básicos

1.- Población o universo: Conjunto de todos los elementos que cumplen ciertas propiedades y entre los cuales se desea estudiar un determinado fenómeno o característica. Su tamaño se denota por "N", y en función de éste pueden ser finitas o infinitas.

Cuando se conocen y pueden enumerarse todos y cada uno de los elementos de la población, a esta población listada se le denomina **censo** (censo electoral, padrón municipal, censo de estudiantes de un centro, ficheros de clientes y proveedores, etc.)

4

2.- Muestra: Subconjunto representativo en cuanto a cantidad y calidad de la población en estudio. Su tamaño se denota por "**n**", y en función de éste se denominan grandes o pequeñas.



5

3.- Tipos de muestreo:

- **Muestreo probabilístico o aleatorio:**
  - La muestra es extraída al azar.
  - Cada elemento de la población tiene igual oportunidad de ser seleccionado y las muestras así obtenidas son equiprobables.
  - El error muestral se puede medir o acotar en términos de probabilidad
- **No probabilísticos:** Se obtienen muestras que no están basadas en un proceso de azar.

Nota: Para que la inferencia estadística sea válida el muestreo debe ser aleatorio o probabilístico.

6

Nota: En Ingeniería es muy frecuente el **Diseño de experimentos:**

Se desarrolla un determinado experimento y se repite varias veces hasta obtener los "n" valores de la variable X en estudio.

(x1 x2 x3 … xn)

¿Qué cuestiones nos planteamos?

7

De dicha variable X, es decir, de todos los posibles valores que puede tomar en la población nos planteamos cuestiones como las siguientes:

(a) ¿Cuál es su distribución (patrón o comportamiento)?

Inferencia No paramétrica

(b) ¿Cuál es su media, μ? ¿Y su varianza o desviación, σ?

Inferencia paramétrica

8

Ejemplo de inferencia NO PARAMÉTRICA:

Lanzamos un dado 120 veces y obtenemos los siguientes resultados. Nos preguntamos si el dado está equilibrado.

Resultados:

X	1	2	3	4	5	6
Nº veces	20	14	23	12	26	25

Una vez definida la población y obtenida la muestra comienza el proceso inferencial.

9

2. INTRODUCCIÓN A LA INF. NO PARAMÉTRICA

El proceso inferencial, ya sea paramétrico o no paramétrico, se recoge en los siguientes pasos.

1. Planteamiento de las hipótesis Ho y H_1
2. Elección del estadístico de contraste
3. Elección de la región de rechazo de Ho
4. Decisión

La hipótesis de nulidad (**hipótesis nula**) se designa por Ho; la hipótesis de investigación (**hipótesis alternativa**), se designa por H_1.

10

- **Ho** corresponde a la situación estándar, de “no diferencia significativa“ entre los resultados observados en la muestra y los de la población que queremos contrastar.

  Es la hipótesis que se rechazará o no, según se disponga de pruebas suficientes para dudar o creer en su validez.

  Cuando no se rechaza una hipótesis nula estamos afirmando que los resultados obtenidos en la muestra no son significativos de un cambio en el comportamiento de la población.

11

- **H_1** es la hipótesis complementaria, que se aceptará o rechazará según rechacemos o no Ho.

  Al aceptar esta hipótesis afirmamos que los resultados muestrales son significativos de dicho cambio en la población.

12

Sigamos el ejemplo anterior:

1.- PLANTEAMIENTO

Tenemos la variable X de interés "resultado del lanzamiento del dado".

Ho: El dado está equilibrado
(La distribución es la uniforme discreta)
H1: El dado no está equilibrado

13

2.- ESTADÍSTICO DE CONTRASTE

Es una expresión matemática que mide la distancia entre lo que hemos obtenido y lo que debería obtenerse en el caso de que la hipótesis nula fuese cierta.

En nuestro caso, vamos a comparar los resultados obtenidos con los "teóricos"

X	1	2	3	4	5	6
Frec. observada	20	14	23	12	26	25
Frec. teórica	20	20	20	20	20	20

14

Si Ho fuese cierta, ¿no debería haberse obtenido 20 veces cada resultado o muy próximos a 20? ¿Hasta qué punto las frecuencias obtenidas difieren de esas 20 veces teóricas?

Para medir esas diferencias utilizamos el siguiente estadístico, llamado **Chi-cuadrado de Pearson con n-1 grados de libertad.** En este caso Chi-cuadrado con 6-1 g.l.

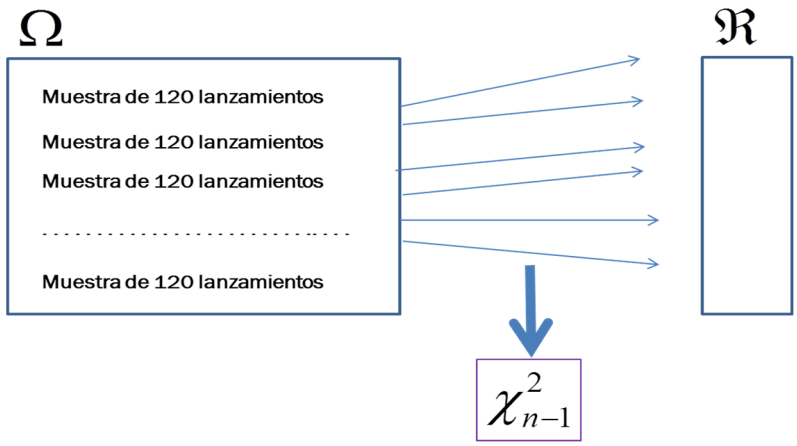
$$\chi^2_{obs} = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$$

Oi= frecuencias observadas

Ei= frecuencias esperadas si Ho es cierta

15

Es importante observar que si repitiésemos los 120 lanzamientos muchas veces, se obtendrían distintos resultados :



Por tanto, se establece que el estadístico es una variable aleatoria continua y su distribución es Chi-cuadrado de Pearson con n-1 g.l.

16

$$\chi^2_{obs} = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = \frac{(20-20)^2}{20} + \ldots + \frac{(25-20)^2}{20} = 8{,}5$$

$Si \quad \chi^2 \approx 0, \quad no \quad rechazaremos \quad Ho$

En este caso frecuencias observadas y teóricas serían prácticamente iguales, lo que confirmaría Ho

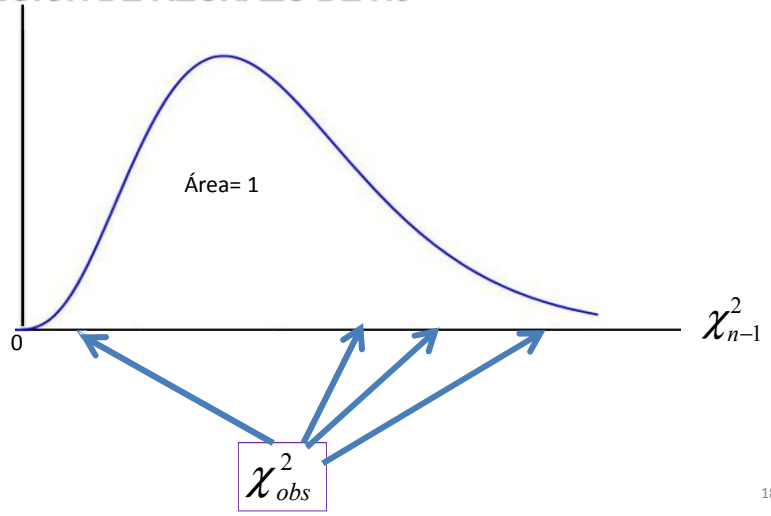
$Si \quad \chi^2 \quad es \quad grande, \quad rechazaremos \quad Ho$

En este caso frecuencias observadas y teóricas serían muy distintas, lo que conllevaría a rechazar Ho

17

¿Pero qué valor del estadístico consideramos pequeño o grande? ¿Hasta qué valor admitimos para No rechazar Ho y a partir de cuál sí rechazamos Ho?

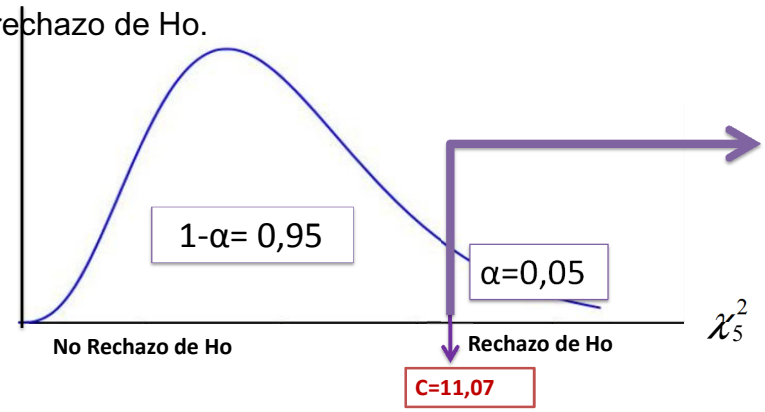
3.- REGIÓN DE RECHAZO DE Ho



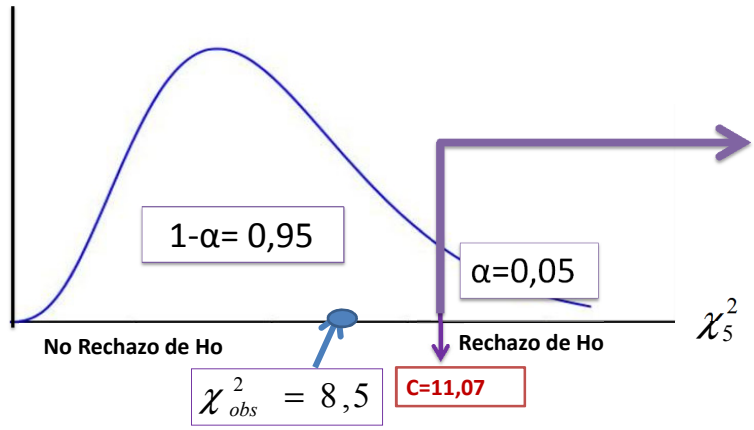
18

Fijamos una probabilidad α=0,05 a la derecha y calcularemos el valor del eje de la χ^2_5

Este punto crítico separa las regiones de rechazo de Ho y la de No rechazo de Ho.



19

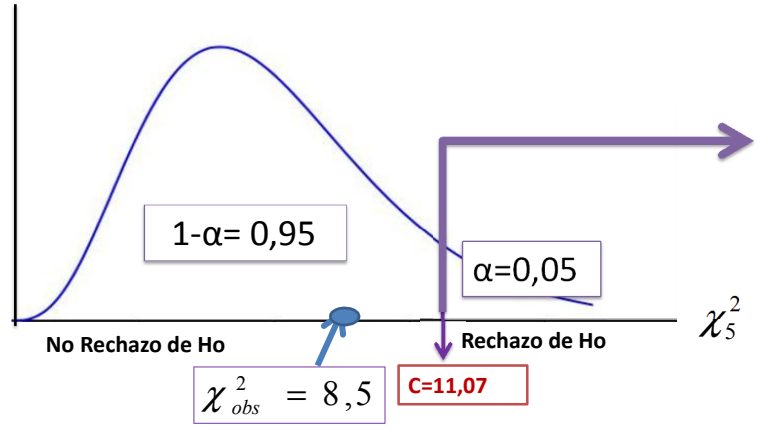


$$Si\ \alpha = 0.05$$
$$P(\chi^2_{obs} > c / Ho) = 0.05 \Rightarrow c = 11.07$$
$$c\ es\ el\ \ percentil\ \ 95$$

20

4.- DECISIÓN

Como el estadístico observado en nuestra muestra es inferior al punto critico, estamos en la zona de No rechazo de Ho. Luego no hay evidencia para rechazar la hipótesis de que el dado está bien construido.



21

3. ERRORES TIPO I Y TIPO II

Para determinar con precisión la regla de actuación en cada caso concreto deberemos analizar los dos errores posibles que podemos cometer al realizar un contraste de hipótesis:

- **Error de tipo I**: Es el error que se cometer al rechazar la hipótesis nula cuando es cierta. Su probabilidad es igual al nivel de significación.

  α = P (rechazar Ho / Ho es cierta)

- **Error de tipo II**: Es el error que se comete al aceptar la hipótesis nula cuando es falsa.

  β = P (aceptar Ho / Ho es falsa)

  = P (rechazar H_1 / H_1 es cierta)

22

- **Potencia de un contraste**:

  $1 - \beta$ = P(aceptar H_1 / H_1 cierta)

	SITUACIÓN REAL	
DECISIÓN	**Ho CIERTA**	**Ho FALSA**
Aceptar Ho	Correcto $1-\alpha$= *nivel de confianza*	***Error tipo II*** = β
Rechazar Ho	***Error tipo I*** = α Nivel de significación	Correcto $1-\beta$ = Potencia

23

Ambos errores son de naturaleza bien distinta, y en la determinación de minimizarlos intervienen muchas veces factores económicos y sociales importantes, que son necesarios tener en cuenta y valorar. Veamos estos errores en el ejemplo de un juicio:

Ho: El acusado es inocente

H1: El acusado es culpable

Solo se condena a un acusado si las pruebas demuestran claramente que es culpable, ya que el perjuicio de condenar a un inocente es muy alto. De este modo, nuestra hipótesis Ho no se rechazará a menos que las pruebas muestrales indiquen evidencias muy fuertes en contra.

24

En este caso:

- **Error de tipo I = Rechazar Ho siendo Ho cierta = Condenar a un inocente**
- **Error de tipo II = Aceptar Ho siendo Ho falsa = Dejar libre a un culpable**
- **Potencia = Rechazar Ho siendo Ho falsa = Condenar a un culpable (No hay error)**

	SITUACIÓN REAL	
DECISIÓN	**Ho CIERTA** Acusado es inocente	**Ho FALSA** Acusado es culpable
Aceptar Ho	Correcto $1-\alpha$ = *nivel de confianza*	***Error tipo II*** $= \beta$ NO Se le condena
Rechazar Ho	***Error tipo I*** $= \alpha$ Se le condena	Correcto $1-\beta$ = Potencia Se le condena por culpable

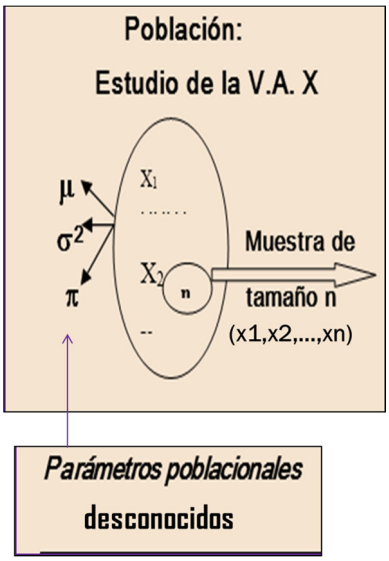
La justicia intenta condenar a un culpable con la mayor probabilidad posible, es decir, se intenta maximizar la potencia.

Por otro lado, se actúa asegurando todo lo posible posible que los inocentes no sean condenados. Es decir, se fija una probabilidad pequeña para el error de tipo I.

En cada contraste, la valoración de estas consecuencias es vital; por ello se hace imprescindible calcular el error de tipo II y la potencia del contraste.

26

4. INTRODUCCIÓN A LA INFERENCIA PARAMÉTRICA

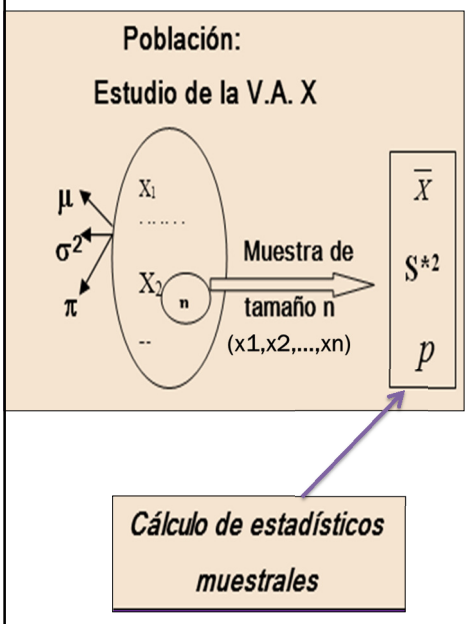


IMPORTANTE:

Las observaciones que vamos a extraer de una muestra $(x_1,x_2,\ldots,x_n)$ son un vector aleatorio.

Si el muestreo ha sido el aleatorio simple, cada observación xi se distribuye igual que la variable aleatoria X en la población y, además, son independientes entre sí.

27

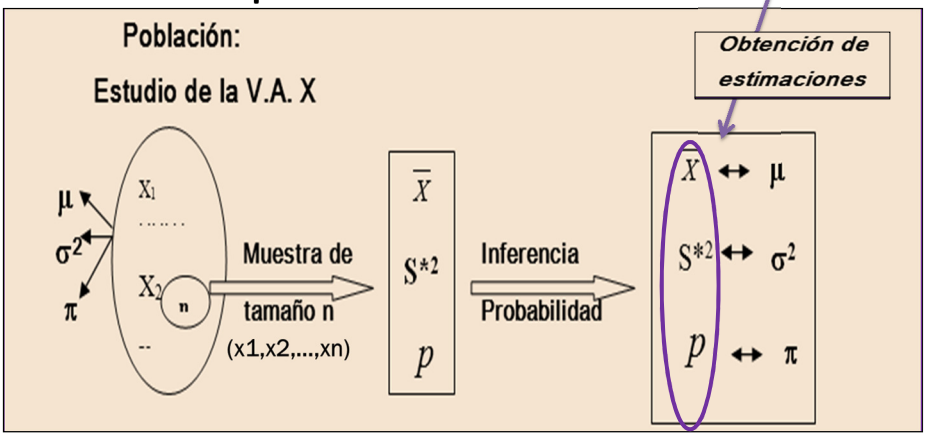


Cualquier transformación que hagamos con la muestra (calcular la media o varianza, por ejemplo) es una V.A denominada ESTADÍSTICO.

$$\vartheta = h(x_1, x_2, \ldots, x_n)$$

28

Una vez que la muestra ha sido seleccionada (el azar ya ha actuado) el estimador ya ha tomado un valor que sirve para estimar el parámetro poblacional. Este valor es $\hat{\vartheta}$ y se llama **Estimador puntual.**



29

Parámetro poblacional a estimar		Estimador muestral (V.A muestral)	
Una población			
Media poblacional	μ	Media muestral	$\overline{X}_n$
Proporción poblacional	Π	Proporción muestral	P_n
Varianza poblacional	σ^2	Cuasivarianza muestral	S^2
Dos poblaciones			
Diferencia de medias poblacionales	$\mu_1 - \mu_2$	Diferencia de medias muestrales	$\overline{X}_1 - \overline{X}_2$
Diferencia de proporciones poblacionales	$\Pi_1 - \Pi_2$	Diferencia de proporciones muestrales	$p_1 - p_2$
Cociente de varianzas poblacionales	σ_1^2/σ_2^2	Cociente de cuasivarianzas muestrales	S_1^2/S_2^2

30

5. DISTRIBUCIONES MUESTRALES

Distribución poblacional:
Es el comportamiento probabilístico de la totalidad de las medidas individuales de una variable en una población.
Así por ejemplo, decimos que la renta media de una CCAA es Normal con una media y desviación concreta o que la duración de un componente electrónico es Exponencial de media 5 años.

Distribución muestral:
Se refiere al comportamiento probabilístico de los diferentes valores que un estimador muestral (por ejemplo la media) podría adoptar en cada una de las muestras del mismo tamaño obtenidas de esa población.

31

Distribuciones Muestrales: Ejemplo de simulación

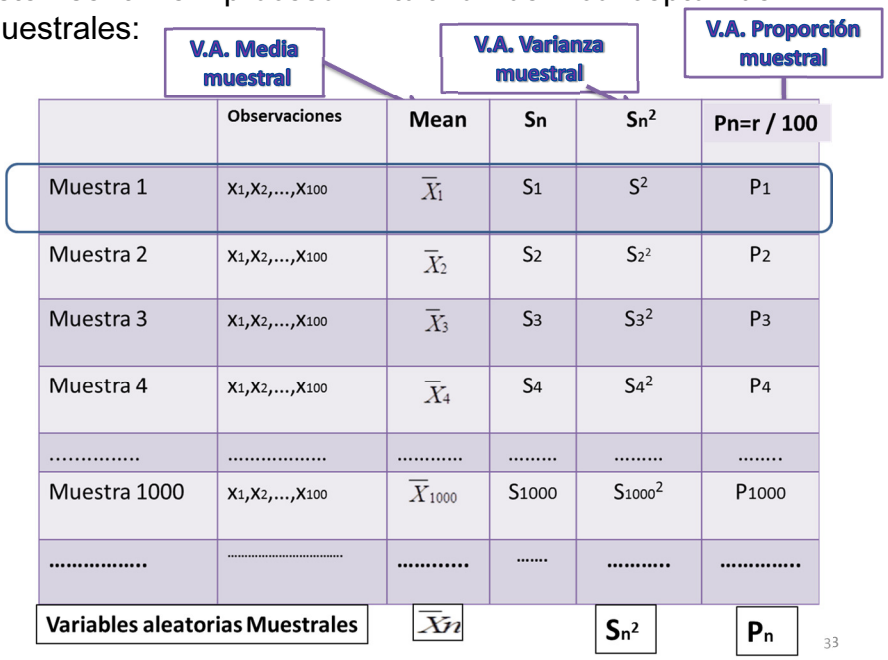
El número de saltamontes en el medio rural de una zona se distribuye según la ley de Poisson de media 2 saltamontes por m_2. De la muestra de 100 observaciones podríamos calcular la media, varianza y proporción de saltamontes machos, por ejemplo.

	Observaciones	Mean	Sn	Sn^2	Pn
Muestra 1	$x_1, x_2, \ldots, x_{100}$	$\bar{X}_1$	S_1	S^2	P_1

Pero supongamos que repetimos el experimento más de 1000 veces…

32

Este sería el proceso intuitivo del concepto de V.A. muestrales:



33

Estudiando gráficamente la variable Mean (media muestral) se concluía que:

1- Su media era igual a 2. E(Mean)= E($\overline{X}_n$)= 2

Y esto ocurre independientemente del tamaño de las muestras.

¿Por qué?. Porque X~Poisson(λ=2)

2- La varianza disminuye cuando se aumenta el tamaño de la muestra. Var (Mean)= Var($\overline{X}_n$) $\longrightarrow$ 0 si n $\longrightarrow \infty$

Var(X)= λ=2 y Var(Mean) = Var(X)/n= 2/n

34

(1) Distribución Muestral de la Media

(1) Si tenemos una variable en la población tal que X ~ **N (μ ,σ)** y extraemos de ella muestras de tamaño n, la distribución de la V.A. Media muestral sigue también una distribución Normal, independientemente del tamaño de la muestra.

$$\overline{X_n} \sim N(\mu, \sigma/\sqrt{n})$$

(2) Si la población no sigue una distribución normal pero la muestra es suficientemente grande, n>30, aplicando el llamado **Teorema Central del Límite** la distribución muestral de medias se aproxima también a la normal anterior.

Sabemos que dada una muestra de tamaño n cada una de las n observaciones X_1, X_2,…, X_n puede considerarse una sucesión de V.A. independientes e idénticamente distribuidas, con media y varianza igual a la variable de procedencia.

- Teorema central del límite

Dadas n v.a. independientes e idénticamente distribuidas, $\{X_1, \ldots, X_n\}$ *tales que*

$$\mathrm{E}[X_k] = \mu < +\infty \quad \text{y} \quad \mathrm{Var}(X_k) = \sigma^2 < +\infty$$

Entonces, si n es suficientemente grande,

$$\sum_{k=1}^{n} X_k \approx \mathrm{N}\left(n\mu, \sigma\sqrt{n}\right) \Rightarrow \frac{\sum_{k=1}^{n} X_k}{n} = \overline{X_n} \approx \mathrm{N}\left(\mu, \sigma/\sqrt{n}\right)$$

36

Ejercicio 1:Distribucion en el muestreo de la media

Una población de un tipo de plantas tiene una talla media de 15 cm y desviación típica de 2.5 cm. Se toma al azar una muestra de 45 plantas.

a) ¿Cuál es la probabilidad de que la media de las tallas de la muestra sea superior a 12.5?

Solución:

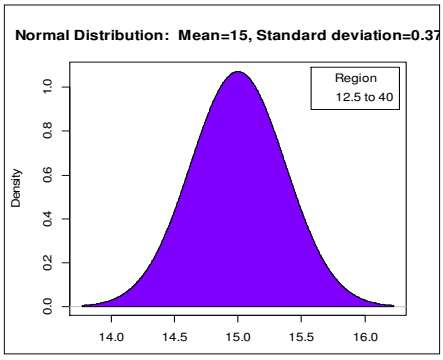
X~ f(x) desconocida, μ= 15cm σ=2,5cm

n=45 (n>30 por lo que se puede aplicar el TCL)

$$\overline{X_n} \approx N(\mu, \sigma/\sqrt{n}) \Rightarrow \overline{X_{45}} \approx N(15; 2.5/\sqrt{45}) = N(15; 0.3727)$$

$$\overline{X_{45}} \approx N(15; 0.3727) \longrightarrow$$

$$P(\overline{X_{45}} > 12.5) = 1$$



b) ¿Qué hipótesis sobre la variable hay que añadir para responder a la misma pregunta si el tamaño muestral fuera de 10 plantas.

Solución:

n=10 es muestra pequeña, luego el TCL no puede aplicarse. La hipótesis que faltaría es la de Normalidad en la variable X.

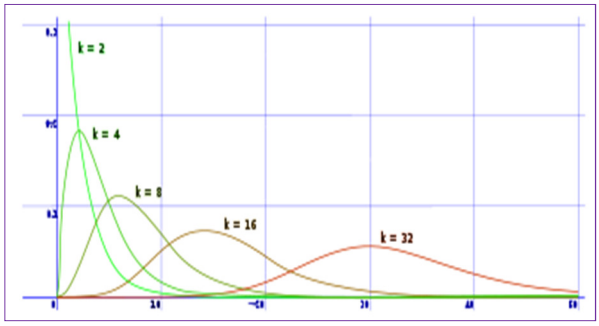
Si X~Normal, la variable $\overline{X_{45}}$ es Normal.

38

(2) Distribución Muestral de la Varianza

Sea una variable aleatoria X ~ N(μ, σ), donde μ es conocida, y una m.a.s. (x_1,...,x_n).La distribución muestral de la varianza nos lleva a la distribución χ^2_{n-1} donde n-1 son los grados de libertad, correspondientes al tamaño de muestra -1.

Función de densidad de la distribución χ^2_n para n g.l



χ^2_n

39

Realizando transformaciones en la varianza de una muestra se llega al siguiente estadístico:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Ejercicio 7: Distribución en el muestreo de la desviación típica muestral

Si la altura en cm de un grupo de población sigue una distribución normal N(176,12), calcular la probabilidad de que la desviación típica muestral sea menor que 10 para una muestra de tamaño 8.

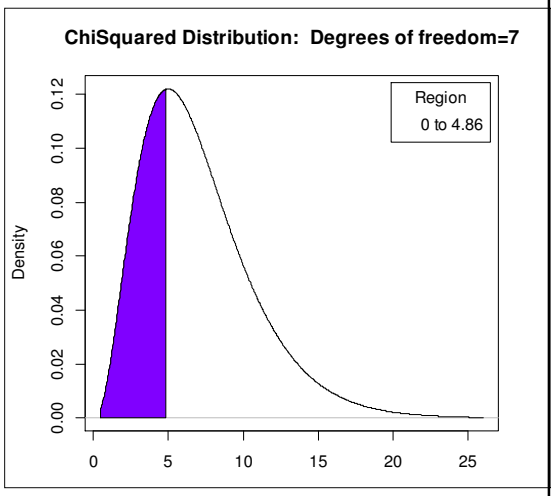
40

$$X = altura \;; \quad X \sim N(176\;;12\,); \quad n = 8$$

¿ **P(S< 10)?** Transformando la anterior expresión llegaremos a la distribución Chi-cuadrado con 7 g.l.

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\begin{aligned} &P(S < 10) \\ &= P(S^2 < 100) \\ &= P\left(\frac{7S^2}{12^2} < \frac{7x100}{12^2}\right) \\ &= P(\chi^2_7 < 4.8611) \\ &= 0.3231 \end{aligned}$$

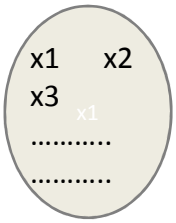


41

(3) Distribución muestral de la Proporción

Para estudiar una proporción poblacional π es razonable establecer un modelo poblacional de Bernoulli para la variable dicotómica.

Población: X~B(1, π)



Al extraer un muestra **grande, de tamaño n>100,** (x1,x2, …,xn) podemos identificar con 1 los elementos que cumplan la característica y con 0 los que no la cumplan.

42

X1	X2	X3	X4	X4	X6	…	…	…	Xn
0	1	1	0	0	1	…	…	…	1

Para calcular la proporción muestral contamos el número de "unos"(sumamos las observaciones)y dividimos entre n. Resulta:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + ... + X_n \quad \sim B(n, \pi)$$

Sabemos que :

$$E(\sum_{i=1}^{n} X_i) = n\pi; \quad Var(\sum_{i=1}^{n} X_i) = n\pi(1-\pi)$$

43

Si dividimos la suma anterior entre el total de elementos de la muestra, n, podemos calcular la proporción muestral:

$$P_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(0 + 1 + 1 + .^{(n)} .. + 1)$$

Calculamos la media y varianza de esta variable:

$$E(P_n) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}E(\sum_{i=1}^{n} X_i) = \frac{1}{n}(n\pi) = \pi$$

$$Var(P_n) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}[n\pi(1-\pi)]$$

$$= \frac{\pi(1-\pi)}{n}$$

44

Conclusión:

La distribución de la V.A. Proporción muestral tiene la media y varianza anteriores y, además, **distribución Normal.**

$$P_n \sim N(\pi, \sqrt{\frac{\pi(1-\pi)}{n}})$$

Ejercicio 4: Distribución en el muestreo de la proporción

El 4% de las piezas que produce una maquina son defectuosas. Se toma una muestra aleatoria de 180 piezas. Calcula la probabilidad de que en la muestra existan menos de 10 piezas defectuosas.

45

Solución 1: Utilizando la V.A Suma

X=Nº de piezas defectuosas en la población. π= 0,04. Sabemos que el total de defectuosas en la muestra es la V.A Suma y su distribución es:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + ... + X_n \quad \sim B(n,\pi) = B(180;0.04)$$

Sabemos que :

$$E(\sum_{i=1}^{n} X_i) = n\pi; \quad Var(\sum_{i=1}^{n} X_i) = n\pi(1-\pi)$$

46

$$\sum_{i=1}^{n} X_i \sim N(n\pi, \sqrt{n\pi(1-\pi)})$$

$$= N(180x0.04; \sqrt{180x0.04x0.96}) = N(7.2;\ 2.6291)$$

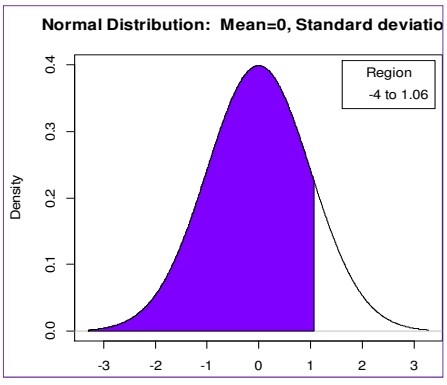
Por tanto:

$$P(\sum_{i=1}^{180} X_i < 10) = P(Z < \frac{10 - 7.2}{2.6291})$$

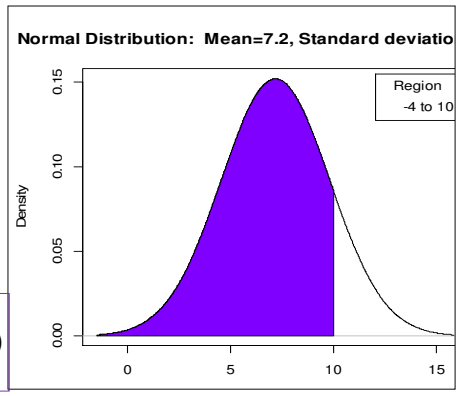
$$= P(Z < 1.065) = 0.8566$$

Hay una probabilidad de 0, 8566 de que en muestras de tamaño 180 haya menos de 10 piezas defectuosas.

47



$$Z \sim N(0,1)$$



$$\sum_{i=1}^{180} X_i \sim N(7.2;\ 2.6291)$$

48

Solución 2: Usando la V.A Proporción muestral

X=Nº de piezas defectuosas en la población; π=0,04

n=180

Que en la muestra existan menos de 10 piezas defectuosas es lo mismo que decir que exista una proporción menor que 10/180=0,0556. Es decir, tenemos que calcular:

$$P\,(P_{180} < 0.0556\;)$$

$$P_n \sim N(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}) = N(0.04;\, 0.0146\,)$$

Por tanto:

49

$$P(P_{180} < 0.0556\,) = P(Z < \frac{0.0556 - 0.04}{0.0146})$$
$$= P(Z < 1.068\,) = 0.8572$$

Hay una probabilidad de 0, 8572 de que en muestras de tamaño 180 haya menos de 10 piezas defectuosas, equivalentemente, una proporción de defectuosas inferior a 5,56%.

Análogamente: el porcentaje de muestras de tamaño 180 con esos valores de defectuosas es un 85,72% en la población.

50