TEMA 2 DISEÑO Y ANALISIS DE DATOS A PARTIR DE LOS APUNTES DADOS POR LA UNED

Recordemos la inferencia estadística sirve para a través de los datos de una muestra conocer las características de la población, así podemos llegar a conocer los parámetros de una población, si existe relación entre las variables, si las variables tienen forma de campana de gauss, si se cumple la hipótesis nula…

En el tema uno se vieron que son los intervalos de confianza y ahora explicará para que sirve y es muy sencillo. El intervalo de confianza de una media por ejemplo sirve para intentar adivinar cuanto valdría esa media en una población ya que nosotros los únicos datos que tenemos son los de la muestra. Tenemos que tener en cuenta que el valor que adivinaremos no es un número real sino un intervalo porque no sabremos exactamente el valor, sino un intervalo porque como mi muestra no es exactamente igual que la población los valores que adivinaremos puede tener un error por arriba y por abajo . Así tendremos intervalos de confianza. Ahora bien y si realmente conocemos el valor de la media de la población , podemos comprobar que el valor de esa media en la población este dentro del intervalo de confianza, sino es que hay un problema en la muestra. También puede que antes de realizar este intervalo hubiese hecho una hipótesis sobre cual sería el posible valor de esa media, y podremos por tanto con los resultados del intervalo saber si se cumple o no se cumple dicha hipótesis

Ejemplo 2.1: En un experimento sobre atención, un psicólogo presenta durante 300 mseg un grupo de 16 letras del alfabeto (con una disposición de 4 filas y 4 columnas). Cada uno de los 12 sujetos que participan en el experimento debe verbalizar tantas letras como recuerde. El promedio de letras bien recordadas es de 7 y la desviación típica insesgada (cuasi-desviación típica) es de 1,3. ¿Entre qué límites se encontrará el verdadero número de palabras bien recordadas, con una probabilidad de 0,95?

$$l_i = \overline{Y} - t_{0,025} \cdot \frac{S_{n-1}}{\sqrt{n}} = 7 - 2.201 \cdot \frac{1,3}{\sqrt{12}} = 6,174$$

$$L_s = \overline{Y} + t_{0,975} \cdot \frac{S_{n-1}}{\sqrt{n}} = 7 + 2.201 \cdot \frac{1,3}{\sqrt{12}} = 7,826$$

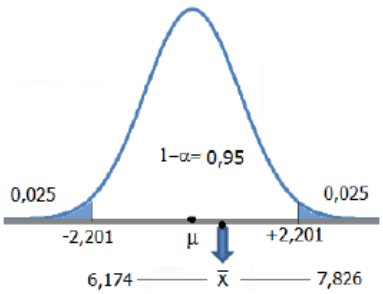


Figura 2.1: Intervalo de confianza par la media

El intervalo de confianza obtenido es (6,174; 7,826). Podemos afirmar al 95% de confianza que la media poblacional (desconocida) para el número de letras recordadas se encuentra entre los valores 6,174 y 7,826. Es decir, el intervalo de confianza de la media nos indica el conjunto de valores que podría tener la media poblacional con el nivel de confianza fijado previamente en el 95%. Por tanto, este intervalo se puede utilizar también para contrastar hipótesis sobre el valor que puede tomar este parámetro en la población. Así, si formulamos las hipótesis:

También el intervalo de confianza sirve para confirmar si se cumple la hipótesis nula, aunque no es el método más habitual, que es por los estadísticos de contraste, que son las pruebas paramétricas y no paramétricas o sea la t y la z, el chi cuadrado… acompañados por su probabilidad asociada.

Un problema que nos podemos encontrar es que muchas veces en la realidad no conocemos la varianza de la población, porque si conociésemos la varianza de la población, tendríamos posibilidad de acceder a todos los datos de la población y no necesitaríamos hacer ningún intervalo, o parámetros. En este caso no podemos utilizar un cálculo de la z para el cual es necesario conocer la varianza de la población, y cuyo objetivo es saber cuanto se aleja la media de la muestra, de la media de la población. Aquí hay un ejemplo conocemos los datos de la población

Ejemplo 2.2: Por estudios previos conocemos que la población masculina de la tercera edad de una determinada Comunidad Autónoma, tiene un gasto medio en medicamentos de 215 euros/año con una desviación típica de 36 euros y queremos saber si la población femenina tiene el mismo gasto. Con tal finalidad analizamos el gasto medio de una muestra de 324 mujeres de la tercera edad de esa misma comunidad observando que la media es de 220 euros/año. Asumimos que esta variable se distribuye normalmente en la población. Fijando un nivel de confianza del 95%, contraste si el gasto de las mujeres es significativamente distinto de 215 euros/año.

En este caso conocemos los datos de la población aunque sea la de los hombres, pero es fácil de ver en el enunciado del problema oímos hablar de un valor en relación a la población , entonces usaremos la z

$$Z = \frac{\overline{Y} - \mu}{\sigma_{\overline{Y}}} = \frac{220 - 215}{2} = 2{,}5$$

$$(p = 0{,}0124)$$

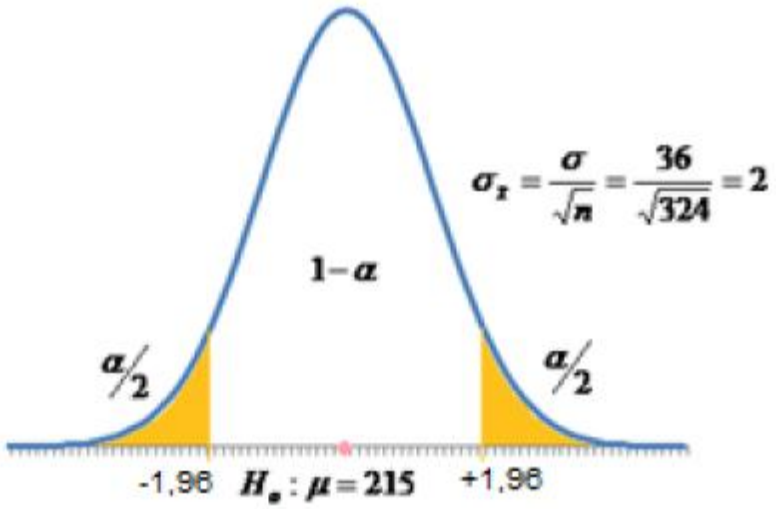


Figura 2.2: Distribución muestral de la media y regiones de decisión de la H_0: μ=215 para un nivel de confianza del 95%

Si se desconocen los valores de la población y por tanto hay que adivinarlos solo a través de las muestras. Entonces al desconocer la varianza poblacional y si la variable se distribuye normalmente, entonces debemos usar el cálculo de la t de student para saber si se cumple o no

se cumple la hipótesis nula,los resultados de esta t de student se parece a la z cuanto más

$$t = \frac{\overline{Y} - \mu_o}{\sigma_{\overline{Y}}} = \frac{\overline{Y} - \mu_o}{\hat{\sigma}/\sqrt{n}}$$

grande es la muestra.

Ejemplo 2.3: En un experimento sobre atención, un psicólogo presenta durante 300 mseg un grupo de 16 letras del alfabeto (con una disposición de 4 filas y 4 columnas). Cada uno de los 12 sujetos que participan en el experimento debe verbalizar tantas letras como recuerde. El promedio obtenido de letras bien recordadas es de 7 y la desviación típica insesgada (cuasi-desviación típica) de la muestra es de 1,3. Sabiendo que el recuerdo es una variable que se distribuye normalmente en la población y fijando el nivel de significación en 0,05, ¿Puede ser 8 la media de letras recordadas?.

En este ejemplo no llegamos a conocer en el enunciado ningún valor de la población todos los datos que tenemos son de un experimento y por tanto de una muestra y por tanto simplemente buscamos ver si un valor para la población que nosotros presuponemos sin motivo como es el 8 de media, que da en la pregunta podría funcionar o no con los datos de la muestra, para ello debemos contrastar la hipótesis, y así si se cumple la hipótesis nula porque los cálculos que hemos obtenido al calcular la t son más pequeños que los de la t teórica que obtenemos de la tabla,

El problema se resolvería así

$$t = \frac{\overline{Y} - \mu_o}{\sigma_{\overline{Y}}} = \frac{7-8}{1{,}3/\sqrt{12}} = \frac{-1}{0{,}375} = -2{,}66$$

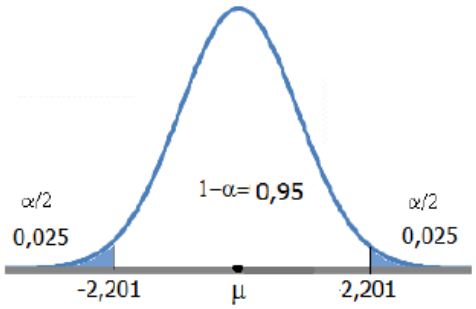


Figura 2.4: Valores críticos de la distribución muestral para un nivel de confianza del 95%

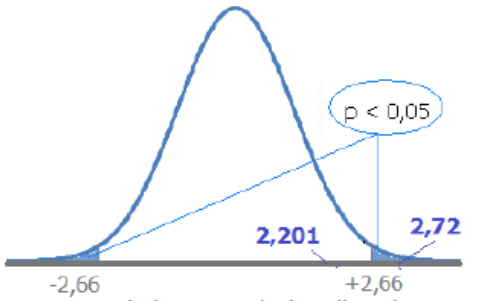
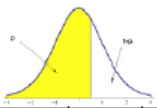


Figura 2.5: Nivel crítico p asociado al estadístico de contraste t=2,66 en un contraste bilateral



g.l.	0,550	0,600	0,650	0,700	0,750	0,800	0,850	0,900	0,950	0,975	0,990	0,995
10	0,1289	0,2602	0,3966	0,5415	0,6998	0,8791	1,093	1,372	1,812	2,228	2,764	3,169
11	0,1286	0,2596	0,3956	0,5399	0,6974	0,8755	1,088	1,363	1,796	2,201	2,718	3,106
12	0,1283	0,2590	0,3947	0,5386	0,6955	0,8726	1,083	1,356	1,782	2,179	2,681	3,055

Con 11 gl, el valor 2,66 se encuentra entre 2,201 y 2,718, (fig: 2.5 y tabla de t) por tanto: 0,025>p>0,01 en una cola de la distribución y 0,05>p>0,02 utilizando las dos colas de la distribución

Os acordáis en el tema anterior que una vez explicado en calculo del intervalo de confianza para la media y la t de student o la z para la media, comenzamos ha hacer los mismos cálculos para la varianza y las proporciones (que no son más que medias de variables dicotómicas)

Contraste sobre proporción

Recordamos la proporción p no es más que la media de una variable dicotómica, y por tanto los cálculos que se realizan con el son los mismos que en el caso anterior , pero en vez de aparecer los símbolos de la media aparecerán los de la proporción . P es la frecuencia relativa de casos positivos o casos que nos interesa, mientras que q es la frecuencia relativa de casos que no nos interesa o negativos , mientras que su varianza se calcula por la multiplicación de p por q, y cuando necesitemos para los cálculos de la media de la distribución muestral y la desviación típica de la distribución muestral y se calculará con esta fórmulas

$$\mu_p = \pi$$

$$\sigma_p = \sqrt{\frac{\pi_o \cdot (1-\pi_o)}{n}}$$

Teniendo estos valores ya podemos saber si se cumple la hipótesis nula y por tanto la proporción de la variable es aceptada o no como real y se corresponde con la población. Por medio de esta formulita

$$Z = \frac{p-\pi_o}{\sigma_p} = \frac{p-\pi_o}{\sqrt{\frac{\pi_o \cdot (1-\pi_o)}{n}}}$$

Así podemos verlo en este ejemplo

observando que 39 de ellos afirman haber cambiado de móvil en el último año. Con un nivel de confianza del 99%, ¿podemos admitir la hipótesis del investigador?

Ejemplo 2.4: Un investigador de estudios de mercado cree que más del 20% de los adolescentes cambian de móvil cada año. Con esta finalidad realiza una encuesta sobre una muestra de 150 adolescentes

Siguiendo la fórmula anterior podemos resolver el problema

$$Z = \frac{p-\pi_o}{\sigma_p} = \frac{p-\pi_o}{\sqrt{\frac{\pi_o \cdot (1-\pi_o)}{n}}} = \frac{0{,}26-0{,}20}{\sqrt{\frac{0{,}20 \cdot 0{,}80}{150}}} = 1{,}837$$

y con este valor compararlo con el valor teórico y si este teórico es mayor que el práctico calculado se acepta la

hipótesis nula.

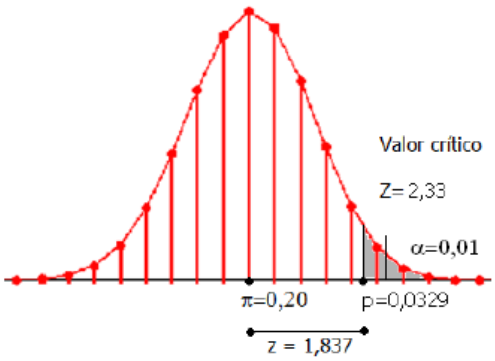


Figura 2.7: Estadístico de contraste, nivel de significación y nivel crítico p para el contraste unilateral derecho de H_0: π=0,20

Conclusión. Como el estadístico de contraste -o discrepancia encontrada entre los valores p=0,26 y $\pi = 0{,}20$ de 1,837 no supera la máxima diferencia que puede esperarse por simple azar (el valor crítico 2,33), no tenemos evidencia suficiente para rechazar la hipótesis nula. De otra forma, el nivel crítico p de 0,0329 es mayor que el nivel de significación $\alpha = 0{,}01$ por lo que no podemos rechazar la hipótesis nula.

Contraste de hipótesis sobre la varianza poblacional