INFERENCIA ESTADÍSTICA

DECISIÓN ESTADÍSTICA. CONTRASTE DE HIPÓTESIS.

Introducción

La Inferencia Estadística persigue la obtención de conclusiones sobre un gran número de datos, en base a la observación de una muestra obtenida de ellos; también intenta medir su significación, es decir, la confianza que nos merecen.

Dos son los temas que trata de resolver la Estadística Inferencial en torno a la toma de decisiones:

1. Decidir si un valor obtenido a partir de la muestra es probable que pertenezca a la población.
2. Decidir, en términos de probabilidad, si las diferencias observadas entre dos muestras significan que las poblaciones de las que se han obtenido las dos muestras son realmente diferentes.

(Este año estudiaremos el primer apartado)

Los métodos de decisión estadística están ligados a los de estimación de parámetros mediante los intervalos de confianza, aunque se apoyan en nuevos conceptos como significación de hipótesis y otros.

Recuerda: *En la estimación de parámetros, intentábamos obtener un valor o un intervalo de valores que constituyesen la mejor estimación del parámetro desconocido, a partir de la información muestral.*

Para estimar puntualmente la media poblacional, utilizamos $\bar{x}$ (la media de la muestra).

Para estimar puntualmente la proporción poblacional, utilizamos p (la proporción de la población).

Y si hacemos la estimación por intervalos, utilizamos

para la media,

$$\left(\bar{x} - z_{\frac{c}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{c}{2}}\frac{\sigma}{\sqrt{n}}\right) \text{ y}$$

para la proporción

$$\left(p - z_{\frac{c}{2}}\sqrt{\frac{p.q}{n}}, p + z_{\frac{c}{2}}\sqrt{\frac{p.q}{n}}\right)$$

1. DECISIONES ESTADÍSTICAS. HIPÓTESIS ESTADÍSTICAS

En la Teoría de la Decisión Estadística se trata de utilizar los datos obtenidos a partir de una muestra para tomar una decisión sobre la población.

Por ejemplo, decidir si un nuevo fármaco es efectivo, si un método de aprendizaje es mejor que otro, si una moneda está o no cargada, etc.

Estas decisiones se llaman **decisiones estadísticas**. Hay que tener en cuenta que estas decisiones están tomadas sobre una base probabilística, es decir, el acierto de la decisión se mide en términos de probabilidad.

Tenemos que realizar determinados supuestos o conjeturas acerca de las poblaciones que se estudian. Estos supuestos que pueden ser o no ciertos se denominan **hipótesis estadísticas**.

Hipótesis nula es la denominación que recibe la hipótesis a contrastar, que se considera provisionalmente como verdadera y que es revisada tras haber obtenido información de los datos muestrales. Se designa por **H_0.**

Hipótesis alternativa es la denominación que recibe el conjunto de situaciones restantes posibles, o admitidas como posibles, en una situación experimental dada. Se designa por **H_1.**

En la toma de decisiones estadísticas, toda hipótesis nula H_0 ha de ir acompañada de una alternativa H_1, que es la que aspira a desplazar a la nula.

Todo contraste de hipótesis nos lleva a aceptar o rechazar la hipótesis nula planteada, y pueden ocurrir los siguientes casos:

1. Aceptar la hipótesis nula siendo verdadera. Esta decisión es correcta.
2. Rechazar la hipótesis nula siendo falsa. Esta decisión también es correcta.
3. Rechazar la hipótesis nula siendo verdadera. Cometemos un **error de tipo I.**
4. Aceptar la hipótesis nula siendo falsa. Cometemos un **error de tipo II.**

Está claro que si supiésemos la veracidad o falsedad de la hipótesis no sería necesario hacer el contraste, así que nunca sabremos si estamos cometiendo algún error.

2. CONTRASTES DE HIPÓTESIS PARA LA MEDIA

Se tiene una población $N(\mu,\sigma)$ y a la vista de los resultados obtenidos en una muestra de tamaño n, hay que tomar una decisión, con un nivel de significación $N_s = 1 - N_c$, sobre el valor original de la media de la población. Vamos a considerar tres casos.

a) *Contraste bilateral*

1º) Enunciamos la hipótesis nula y la alternativa

Hipótesis nula H_o: $\mu = \mu_o$ Hipótesis alternativa H_1 : $\mu \neq \mu_o$

2º) Calculamos $z_{c/2}$ que verifique $P\left(z \leq z_{c/2}\right) = \frac{1+N_c}{2}$

3º) Calculamos el intervalo de aceptación

$$\left(\mu_0 - z_{c/2}.\frac{\sigma}{\sqrt{n}}, \mu_0 + z_{c/2}.\frac{\sigma}{\sqrt{n}}\right)$$

4º) Tomamos la decisión y la interpretamos:

Si $\bar{x} \in \left(\mu_0 - z_{c/2}.\frac{\sigma}{\sqrt{n}}, \mu_0 + z_{c/2}.\frac{\sigma}{\sqrt{n}}\right)$, aceptamos la hipótesis nula H_o (si nos estamos equivocando, cometeríamos un error de tipo II)

Si $\bar{x} \notin \left(\mu_0 - z_{c/2}.\frac{\sigma}{\sqrt{n}}, \mu_0 + z_{c/2}.\frac{\sigma}{\sqrt{n}}\right)$, rechazamos la hipótesis nula y tomamos la hipótesis alternativa (y podríamos estar cometiendo un error de tipo I)

Ejercicios:

1º) **(Pau 2010)** Se sabe que el precio de los libros de bachiller es una variable aleatoria normal con media 38.2 euros y desviación típica de 5.25 euros. Una muestra aleatoria simple de 16 libros de Química de distintas editoriales tiene un precio medio de 42.3 euros. Se quiere decidir si existe diferencia significativa entre la media del precio de los libros de Química y la media del precio de los libros de bachiller en general con un nivel de significación α = 0 05

2º) **(Pau 2010)** Se sabe que las calificaciones de los alumnos de segundo de bachiller en matemáticas es una variable aleatoria normal de media 5.5 y varianza 1.69. Se extrae una muestra aleatoria de 81 alumnos que cursan el bachiller bilingüe obteniéndose una media muestral de 6.8 puntos en las calificaciones de dichos alumnos en la asignatura de matemáticas. Se quiere decidir si existe una diferencia significativa entre la media de las calificaciones en matemáticas de los alumnos del bachiller bilingüe y la media de las

calificaciones en matemáticas de los alumnos de segundo de bachiller en general con un nivel de significación α = 0 01.

3º) **(Pau 2011)** Se sabe que el ingreso anual por hogar en España es una variable normal de media 29400 euros y desviación típica de 17400 euros. Se extrae una muestra aleatoria simple de 400 hogares de la Comunidad de Murcia obteniéndose un ingreso anual medio por hogar de 26600 euros. Suponiendo que el ingreso anual por hogar en la Comunidad de Murcia es una variable normal con la misma desviación típica, decidir con un nivel de significación del 5% si existe una diferencia significativa entre el ingreso anual medio por hogar en España y el ingreso anual medio por hogar en la Comunidad de Murcia.

4

4º) **(Pau 2011)** Se sabe que la edad de los profesores de una Comunidad Autónoma sigue una distribución normal con varianza de 5 años. Una muestra aleatoria de 200 profesores de dicha Comunidad tiene una media de 45 años. ¿Se puede afirmar con un nivel de significación del 0,05 que la edad media de todos los profesores de la Comunidad es de 46 años?

b) ***Contraste unilateral***

Primer caso: Hipótesis nula H_o: $\mu > \mu_0$ Hipótesis alternativa H_1 : $\mu \leq \mu_0$

Región de aceptación $\left(\mu - z_c.\frac{\sigma}{\sqrt{n}}, +\infty\right)$

Con z_c que verifique $P(z \leq z_c) = N_c$

Ejemplo:

Se quiere contrastar si el nivel de colesterol en sangre de un grupo de enfermos es mayor que el que tiene una población que se ha tomado como referencia, y que es de 160 u. Se sabe que la desviación típica de la cantidad de colesterol en sangre es de 20 u.

a) Establecer el contraste dando la hipótesis nula y la alternativa.
b) Si el tamaño de la muestra es de 50 y la media muestral es de 165 u, se rechazará la hipótesis con un nivel de significación de 0,0025?

1º) Enunciamos las hipótesis nula y alternativa

Hipótesis nula H_o: $\mu > 160$ Hipótesis alternativa H_1 : $\mu \leq 160$

2º) Como se trata de un contraste unilateral, calculamos z_c que verifique $P(z \leq z_c) = N_c$

3º) Calculamos el intervalo de aceptación

$$\left(\mu - z_c.\frac{\sigma}{\sqrt{n}}, +\infty\right) \qquad \left(160 - 1{,}96.\frac{20}{\sqrt{50}}, +\infty\right) = (154{,}5; +\infty)$$

4º) Tomamos la decisión y la interpretamos

Como la media muestral es 165 u, y cae dentro de la zona de aceptación $(154,5;+\infty)$, podemos decir que el nivel de colesterol ha aumentado. Aceptamos la hipótesis nula.

Ejercicios :

5º) **Pau 2013**. Hace veinte años la edad en que la mujer tenía su primer hijo seguía una distribución normal con media 29 años y desviación típica de 2 años. Recientemente en una muestra aleatoria de 144 mujeres se ha obtenido, para dicha edad, una media de 31 años. Con un nivel de significación de 0,05 ¿se puede afirmar que la edad media en la que la mujer tiene su primer hijo es mayor actualmente que hace veinte años?

<u>Segundo caso</u>: Hipótesis nula H_0: $\mu < \mu_0$ Hipótesis alternativa H_1 : $\mu \geq \mu_0$

Región de aceptación $\left(-\infty; \mu + z_c.\frac{\sigma}{\sqrt{n}}\right)$

Con z_c que verifique $P(z \leq z_c) = N_c$

Ejemplo:

La estatura de los varones que estudian 20 de bachillerato en una localidad murciana sigue una distribución normal de media desconocida y desviación típica igual a 8 cm. Se toma al azar una muestra de 100 alumnos y se observa que su estatura media es de 175 cm. ¿Se puede afirmar, con un nivel de confianza del 95% que la estatura media de dicha población de varones es como máximo de 176 cm?

1º) Hipótesis nula H_0: $\mu \leq 176$ Hipótesis alternativa H_1 : $\mu > 176$

2º) $P(z < z_c) = 0,95$ por lo tanto $z_c = 1,65$

3º) La región de aceptación es

$$\left(-\infty; \mu + z_\alpha.\frac{\sigma}{\sqrt{n}}\right) = \left(-\infty; 176 + 1,65.\frac{8}{\sqrt{100}}\right) = (-\infty; 177)$$

4º) Como $\bar{x} = 175 \in (-\infty; 177)$ aceptamos la hipótesis de que la estatura media de la población de varones es como máximo de 176 cm

Ejercicios :

6º) **Pau 2013**. Se sabe que en una población el nivel de colesterol en la sangre se distribuye normalmente con una media de 160 u. y una desviación típica de 20 u. Si una muestra de 120 individuos de esa población que siguen una determinada dieta, supuestamente adecuada para bajar el nivel de colesterol, tiene una media de 158 u. ¿Se puede afirmar que el nivel medio de colesterol de los que siguen la dieta es menor que el nivel medio de la población en general, para un nivel de significación de 0,01?

CONTRASTE DE HIPÓTESIS PARA LA PROPORCIÓN

Se quiere contrastar una hipótesis mediante los resultados de una proporción p de elementos de una muestra de tamaño n que poseen una característica determinada.

También aquí tenemos tres casos

a) Bilateral

* Hipótesis nula . $H_0 : P = p_0$

* Región de aceptación $\left(p_0 - z_c \cdot \sqrt{\frac{p_0 . q_0}{n}}, p_0 + z_c \cdot \sqrt{\frac{p_0 . q_0}{n}} \right)$

* z_c / $P(z < z_c) = \frac{1 + N_c}{2}$

Ejercicios :

7º) **Pau 2012**. Hace 10 años, el 65% de los habitantes de cierta Comunidad Autónoma estaba en contra de la instalación de una central nuclear. Recientemente, se ha realizado una encuesta a 300 habitantes y 190 se mostraron contrarios a la instalación. Con estos datos y con un nivel de significación de 0,01 ¿se puede afirmar que la proporción de contrarios a la central sigue siendo la misma?

Unilateral

Primer caso

- Hipótesis nula $H_0 : P \leq p_0$
- Región de aceptación $\left(-\infty; p_0 + z_c \cdot \sqrt{\frac{p_0 . q_0}{n}} \right)$

Con z_c que verifique $P(z \leq z_c) = N_c$

Segundo caso

- Hipótesis nula $H_0 : P \geq p_0$
- Región de aceptación $\left(p_0 - z_c \cdot \sqrt{\frac{p_0 \cdot q_0}{n}}; +\infty \right)$

- Con z_c que verifique $P(z \leq z_c) = N_c$

Ejercicios :

7º) **Pau 2013** Según un estudio realizado en el año 2000, en una población la proporción de personas que tenía sobrepeso era del 24%. En los últimos años ha disminuido la actividad física que realizan los individuos, lo que hace sospechar que dicha proporción ha aumentado. Para contrastarlo, se ha tomado recientemente una muestra aleatoria de 1195 individuos, de los cuales 310 tienen sobrepeso. Con un nivel de significación del 1%, ¿se puede rechazar que la proporción sigue siendo del 24% e inclinarnos por que dicha proporción ha aumentado?

8º) Una empresa dedicada a la fabricación de luminosos publicitarios anuncia que, como máximo, hay un 1% de luminosos defectuosos. Se selecciona una muestra de 100 rótulos publicitarios y se observa que aparecen 3 defectuosos. Se pide:

a) Con un nivel de significación del 5%, ¿podemos aceptar la hipótesis del fabricante?
b) ¿ Y con un nivel de confianza del 99%?

Soluc: a) Ho : $p \leq \frac{1}{100}$; $0{,}03 \notin (-\infty, 0{,}026)$ rechazamos la hipótesis del fabricante

b) $0{,}03 \in (-\infty; 0{,}033)$ para este nivel de confianza aceptamos la hipótesis.

9º) Según los datos de un censo de 1970, el analfabetismo en cierto país alcanzaba el 40%. Una reciente encuesta realizada sobre una muestra aleatoria de 800 personas, arroja el dato de que 300 de ellas son analfabetas. ¿Puede considerarse, con un nivel de confianza del 95% que se ha reducido el analfabetismo?

Soluc: H_o: $p \leq 0{,}4$; $0{,}37 \in (-\infty; 0{,}43)$ aceptamos la hipótesis.

10º) Una encuesta a 64 profesionales de una institución reveló que el tiempo medio de empleo en dicho campo era de 5 años, con una desviación típica de 4. Considerando un nivel de significación del 0,05, ¿sirven esto datos de soporte de que el tiempo medio de empleo de los profesionales de esta institución está por debajo de los 6 años?. Suponemos que la población de profesionales se distribuye normalmente.

11º) Según un estudio realizado durante el año 2000 en un hospital, la distribución de los pesos de los recién nacidos fue N (3,450; 0,52): A lo largo de este año se ha analizado el peso

de 36 recién nacidos tomados al azar, obteniéndose una media de 3,300 Kg. ¿Podemos afirmar que esta diferencia es debida al azar con una confianza del 95%? Con el mismo nivel de confianza, ¿Cambiaría la respuesta si la media de 3,300 kg se hubiera obtenido al analizar el peso de 81 recién nacidos tomados al azar? (PAU 2001)

12º) El salario medio correspondiente a una muestra de 1600 personas de cierta población es de 935 €. Se sabe que la desviación típica de los salarios en la población es de 200 €. ¿Se puede afirmar, con un nivel de confianza del 99%, que el salario medio en dicha población es de 950 €?

13º) Un laboratorio afirma que un calmante quita la jaqueca como máximo en 14 minutos en los casos corrientes. Con el fin de comprobar esta información, se eligen al azar 30 pacientes con jaqueca y se toma como variable en el experimento el tiempo que transcurre entre la administración del calmante y el momento en que desaparece la jaqueca. Los resultados obtenidos en esta muestra fueron, media 17 minutos y desviación típica 7 minutos. ¿Podemos admitir como cierta la afirmación del laboratorio a un nivel de confianza del 95%?

14º) En una comunidad autónoma se estudia el número medio de hijos por mujer a partir de los datos disponibles en cada municipio. Se supone que este número sigue una distribución normal con desviación típica igual 0,08. El valor medio de estos datos para 36 municipios resulta ser igual a 1,17 hijos por mujer. Se desea contrastar, con un nivel de significación de 0,01, si el número medio de hijos por mujer en la comunidad es de 1,25.

15º) El partido X ha hecho una encuesta a 100 personas de las que 35 les han comunicado su intención de votarles. Suponiendo que no les han engañado, se atreven a decir en sus mítines que van a obtener como mínimo el 40% de los votos. ¿Qué puedes decir de tal afirmación con un nivel de riesgo del 0,04?

16º) Se sabe que la renta anual en Marbella sigue una distribución normal de media desconocida y desviación típica 0,24 millones de €. Se ha observado la renta anual de 16 vecinos de esa localidad escogidos al azar, y se ha obtenido un valor medio de 1,6 millones de €. Contrasta, a un nivel de significación del 5%, si la media de la distribución es al menos de 1,45 millones de €.

a) ¿cuáles son las hipótesis nula y alternativa del contraste?
b) Determina la forma de la región crítica.
c) ¿Se aceptará la hipótesis nula con el nivel de significación indicado?