Estadística. Práctica 8

Intervalos de Confianza y Contrastes de Hipótesis en poblaciones normales (II)

8.1 C.H. e I.C. para diferencia de medias y cociente de varianzas de dos poblaciones normales independientes

Ejemplo 1: Dividimos aleatoriamente un conjunto de pacientes afectados por cierto virus en dos grupos a los que se administra respectivamente un placebo y un tratamiento en estudio para combatir dicha afección vírica. Después de tratar a los pacientes durante 2 meses se mide la concentración de virus de cada uno de ellos. Los resultados se muestran en la siguiente tabla:

Sujeto	Administración	Nivel	Sujeto	Administración	Nivel
1	placebo	23	2	placebo	25
3	placebo	26	4	placebo	24
5	placebo	26	6	placebo	24
7	placebo	22	8	placebo	25
9	placebo	27	10	placebo	25
11	tratamiento	22	12	tratamiento	23
13	tratamiento	22	14	tratamiento	24
15	tratamiento	19	16	tratamiento	20
17	tratamiento	21	18	tratamiento	23
19	tratamiento	22	20	tratamiento	23

Supuesto que las variables bajo estudio siguen distribuciones normales independientes, ¿puede decirse, al 5% de significación, que el tratamiento hace el efecto deseado en la infección, es decir, que disminuye el nivel de virus?

Para responder a esta cuestión hay que contrastar la igualdad de medias de las distribuciones X (nivel de virus de los pacientes a los que se administró el placebo) e Y (nivel de virus de los pacientes a los que se administró el tratamiento en estudio). Estas variables son independientes al haber sido distribuidos los pacientes aleatoriamente entre los dos grupos, y se supone que se distribuyen normalmente.

Se trata pues de resolver, al 5% de significación, el contraste:

$$\begin{cases} H_0: & \mu_X = \mu_Y \\ H_1: & \mu_X > \mu_Y \end{cases} \qquad \text{que equivale a} \qquad \begin{cases} H_0: & \mu_X - \mu_Y = 0 \\ H_1: & \mu_X - \mu_Y > 0 \end{cases}$$

Las varianzas poblacionales son desconocidas, y lo primero que se hace es testar la igualdad de las mismas resolviendo, también al 5% de significación, el contraste:

$$\begin{cases} H_0: & \sigma_X^2 = \sigma_Y^2 \\ H_1: & \sigma_X^2 \neq \sigma_Y^2 \end{cases} \qquad \text{que equivale a} \qquad \begin{cases} H_0: & \sigma_X^2/\sigma_Y^2 = 1 \\ H_1: & \sigma_X^2/\sigma_Y^2 \neq 1 \end{cases}$$

El procedimiento que debemos seguir para ello es `Estadísticos` → `Varianzas` → `Test F para dos varianzas`[1]. En `grupos` elegiremos la variable de agrupación que debe ser de tipo carácter y tener sólo dos modalidades para así separar los datos en las dos muestras independientes, que en nuestro caso es la variable *Administración*. En `variable explicada` elegiremos la variable *Nivel* que contiene los datos bajo estudio y elegiremos, además, la hipótesis alternativa `Bilateral`.

Al realizar el contraste sobre las varianzas llegamos a que *no hay evidencia para rechazar la igualdad de varianzas* de ambas poblaciones, ya que el p-valor obtenido es 0.9546, que es mayor que 0.05.

Ahora se debe realizar el contraste sobre la diferencia de medias, supuesto que las varianzas son iguales. Para ello seleccionamos `Estadísticos` → `Medias` → `Test t para muestras independientes`. Procedemos entonces a elegir, en el campo `Grupos` la variable *Administración*, en el campo `variable explicada` elegiremos la variable *Nivel*, como hipótesis alternativa elegiremos en este caso `Diferencia >0` y, finalmente, indicaremos que sí suponemos que las varianzas son iguales. Obtenemos que *sí hay evidencia para rechazar la hipótesis nula* pues el p-valor obtenido es 0.0003018, que es menor que 0.05. Es decir, el nivel medio de virus de los pacientes que reciben el placebo es superior al de los pacientes que reciben el tratamiento y, por tanto, el tratamiento surte el efecto deseado.

Para los siguientes ejercicios, cargar el fichero **Rcars**, tras descargarlo de moodle.

Ejercicios:

1. Plantear y resolver contrastes de hipótesis adecuados para estudiar si, para un nivel de significación del 5%, existen evidencias significativas para afirmar que el consumo medio de los coches con 8 cilindros es superior al consumo medio de los coches con 4 cilindros.(Suponer normalidad)

   La variable *cilindr* recoge el número de cilindros de los coches pero no es de tipo cadena y tampoco toma sólo dos valroes, por lo que no puede elegirse en el campo `grupos` para resolver los contrastes que nos interesan. Debemos proceder como sigue:

   - Filtrar el conjunto de datos activo para construir un nuevo conjunto de datos que contenga, únicamente, los casos correspondientes a los coches que tienen 4 u 8 cilindros. Para ello seleccionar `Datos` → `Conjunto de datos activo` → `Filtrar el conjunto de datos activo`. Desmarcamos `Incluir todas las variables`, seleccionamos la variables *cilindr* y *consumo*, en el apartado `Expresión de selección` escribimos `cilindr==4|cilindr==8` y le damos nombre al nuevo conjunto de datos, por ejemplo, **Ejercicio1**.
   - Convertir en factor la variable *cilindr* del nuevo conjunto de datos mediante `Datos` → `Modificar variables del conjunto de datos activo` → `Convertir variable numérica en factor`.
   - Realizar los contrastes de hipótesis necesarios para resolver el problema.

[1] Hemos elegido el test F porque es el que se corresponde con el test descrito en las clases teóricas.

2. Plantear y resolver contrastes de hipótesis adecuados para estudiar si, para un nivel de significación del 5%, existen evidencias significativas para afirmar que el consumo medio de los coches europeos es distinto del consumo medio de los coches japoneses. (Suponer normalidad)

   En este caso tenemos que construir un nuevo conjunto de datos que contenga únicamente los casos correspondientes a los coches europeos y japoneses. El primer paso consistiría pues en filtrar el conjunto de datos original para seleccionar estos casos. En este caso se seleccionan las variables *consumo* y *origen* y en `Expresión de selección` escribimos `origen=="Europa"|origen=="Japón"`. Si solicitamos un resumen del conjunto de datos activo observamos que, en el nuevo conjunto de datos, la variable origen sigue presentando tres modalidades: Europa, Japón y E.E.U.U., si bien la modalidad E.E.U.U. tiene frecuencia 0. Para que en el nuevo conjunto de datos no aparezca la modalidad E.E.U.U. en la variable *origen*, seleccionamos `Datos` → `Modificar variables del conjunto de datos activo` → `Decartar niveles sin uso` y lo aplicamos a la variable *origen*. Una vez hecho esto podemos observar que la variable creada sólo presenta las dos modalidades deseadas y ya podemos resolver adecuadamente los contrastes.

8.2 Análisis de la varianza de un factor

El procedimiento `Estadísticos` → `Medias` → `ANOVA de un factor` genera un análisis de la varianza para una variable dependiente cuantitativa respecto a un única variable factor utilizando un modelo de efectos fijos.

Ejemplo 2: Se sometió a un grupo de estudiantes a varias técnicas de enseñanza (1, 2, 3 y 4) y se les examinó al final de un periodo específico. Las puntuaciones obtenidas (sobre 100) son:

Técnica 1	65, 87, 73, 79, 81, 69
Técnica 2	75, 69, 83, 81, 72, 79, 90
Técnica 3	69, 78, 67, 62, 83, 76
Técnica 4	94, 89, 80, 88

Supuesto que se verifican las hipótesis del modelo de análisis de la varianza, nos preguntamos si existen diferencias significativas entre las puntuaciones medias obtenidas con distintas técnicas (tomaremos α =0.05). Para el análisis necesitamos un archivo de datos con las variables *Puntuación* y *Técnica* y seguimos el procedimiento `Estadísticos` → `Medias` → `ANOVA de un factor`. A continuación seleccionamos *Técnica* en `Grupos` y *Puntuación* como `variable explicada`. Marcamos también la opción `Comparaciones dos a dos de las medias` para comparar las técnicas dos a dos en caso de rechazar en el contraste de igualdad de medias.

Interpretación del resultado:

La tabla ANOVA muestra el resultado del contraste de la igualdad de puntuaciones medias en los distintos niveles del factor. En este caso, para un nivel de significación α =0.05, se rechaza la hipótesis de igualdad de medias al obtenerse un p-valor=0.0323.

En el apartado de comparaciones múltiples se ha utilizado el test de Tukey. En una tabla aparecen los contrastes de las comparaciones múltiples, donde se puede observar cuáles son las medias que difieren. Se encuentran diferencias significativas, al 5% de significación, entre las técnicas 3 y 4, ya que el p-valor correspondiente a esta comparación es menor que 0.05. En otra tabla aparecen intervalos de confianza al 95% para cada pareja de medias y finalmente se muestra una tabla donde se clasifican los niveles de factor en "grupos homogéneos" para un nivel de significación α =0.05.

Ejercicio: Un agricultor desea comparar tres nuevas clases de trigo. Para ello, se toman 6 fincas al azar, plantando en cada una de ellas y en partes distintas las tres clases. La producción, en miles de kilos, de las 6 fincas fue la siguiente:

Clase A	47, 49, 60, 55, 57, 48
Clase B	37, 41, 52, 56, 44, 65
Clase C	45, 28, 32, 20, 31, 23

Supuesto que se verifican las hipótesis del modelo de análisis de la varianza,

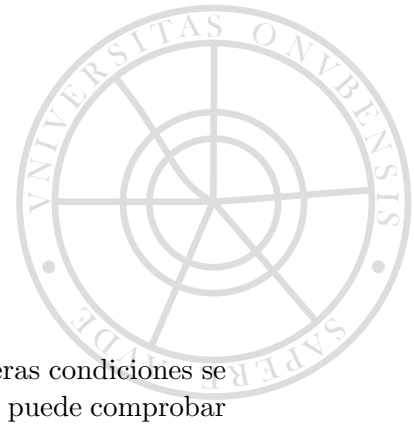
(a) ¿Se puede afirmar, al 5% de significación, que la producción depende significativamente de la clase de trigo?

(b) ¿Qué niveles del factor difieren entre sí al 5% de significación?

8.2.1 Validación del modelo de análisis de la varianza

El modelo de análisis de la varianza se fundamenta en tres hipótesis:

- Normalidad de la variable respuesta en cada grupo.
- Igualdad de varianzas de la variable respuesta entre los grupos.
- Independencia.



Describimos a continuación procedimientos para averiguar si las dos primeras condiciones se cumplen. Si no se está seguro de la independencia de las observaciones, ésta se puede comprobar contrastando si los residuos (errores) son aleatorios aplicando algún test de aleatoriedad, como el test de rachas, que no describimos aquí.

Para comprobar la hipótesis de normalidad de la variable respuesta en cada nivel de factor, se puede aplicar el test de Shapiro-Wilk, que ya describimos en la práctica anterior.

Para comprobar la igualdad de varianzas de la variable respuesta entre los niveles del factor se puede aplicar el **test de Bartlett** o el **test de Levene**, que se encuentra en `Estadísticos` → `Varianzas`. El test de Bartlett es más sensible que el test de Levene a las desviaciones de la normalidad de los datos por lo que se utiliza cuando se tiene asegurada la hipótesis de normalidad.

Ejercicio: Comprobar que los datos del ejemplo y ejercicio anteriores verifican las hipótesis de normalidad de cada grupo y de igualdad de varianzas entre los grupos.