Práctica 7: Intervalos de confianza y contrastes de hipótesis para dos poblaciones.

Nota: Para los estudios de proporciones es muy útil tener cargado el plugin "IPSUR". Se carga desde el menú "Herramientas"de Rcommander. Cuando se carga un plugin Rcommander se reinicia (es mejor cargarlo primero antes de empezar la práctica).

Desde el punto de vista práctico, es más habitual que al tomar una muestra se desconozca no sólo la media sino también la varianza de las variables en estudio. Por esa razón, R Commander NO TIENE menús para los problemas en los que la VARIANZA de la variable se asume CONOCIDA.

En esta práctica reduciremos los ejercicios de inferencia a este tipo de problemas por lo que los intervalos y contrastes en los que se supone la varianza conocida deberán construirse de modo análogo a como se indicó en el último apartado de la práctica 6, en donde R-commander se utiliza como herramienta para calcular los valores muestrales y los percentiles de la distribución del estadístico.

Nota: En todos los ejercicios que asumen Normalidad se requiere de **comprobación la hipótesis de normalidad** de ambas variables.

1. Diferencia de medias y cociente de varianzas en dos muestras independientes.

1. Se investiga la temperatura de deflexión bajo carga de dos tipos diferentes de tubos de plástico empleados para conducciones de riego. Se prueban dos muestras aleatorias de 14 y 16 ejemplares de tipo 1 y tipo 2 respectivamente; las temperaturas de deflexión observadas en grados Fahrenheit:

Tipo 1	206	188	205	187	194	193	207	
	189	213	192	210	194	178	205	
Tipo 2	177	197	206	201	180	176	185	190
	197	192	198	188	189	203	192	200

Nota: en la tabla de datos deben aparecer tantas filas como tubos muestrados, es decir 30(14+16) filas y dos variables: una numérica para la temperatura y otra variable factor que asigna el tipo correspondiente a cada elemento.

a) Comprueba las hipótesis necesarias para el estadístico que corresponde

b) Bajo el supuesto de igualdad de varianzas, efectúa el contraste, con $\alpha = 0{,}05$, para discutir la afirmación de que la temperatura de de flexión bajo carga en los tubos de tipo 1 supera a la de los tubos de tipo 2.

c) Construye también el intervalo de confianza al 95 %.

Solución:

* Para comprobar normalidad, debe realizarse el gráfico de comparación de cuantiles con la normal para cada una de las dos muestras.

Ruta de menús: *Estadisticos–test t para muestras independientes...*

* Para obtener el intervalo de confianza, hay que seleccionar el menú hipótesis alternativa bilateral y el nivel de confianza deseado.
t_{OBS}= 1.4823 y p-valor=0.0741. Por lo que se acepta H_0.

* Posteriormente, para resolver el contraste pedido, se sigue la misma ruta y se selecciona como hipótesis alternativa > 0.
$IC_{95\,\%}(\mu_1 - \mu_2)$=(-2.02, 12,87). Como el 0 está dentro del intervalo, se concluye que las medias poblacionales son iguales con seguridad 95 %

2. Comparación de varianzas. Estudio de la igualdad de varianzas. Nos planteamos el siguiente contraste:

$$\begin{cases} \mathrm{H}_0: & \sigma_1^2 = \sigma_2^2 \Leftrightarrow \sigma_1^2/\sigma_2^2 = 1 \\ \mathrm{H}_1: & \sigma_1^2 \neq \sigma_2^2 \Leftrightarrow \sigma_1^2/\sigma_2^2 \neq 1 \end{cases}$$

El estadístico del contraste se construye a partir del cociente $\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \dfrac{S_1^2}{S_2^2}\ \dfrac{1}{\sigma_1^2/\sigma_2^2}$, cuya distribución es F de Snedecor.

En la ruta *Estadísticos–Varianzas–Test F* para dos distribuciones se plantea y se obtiene la resolución del contraste del modo habitual.
Solución: F_{OBS}= 1.2898 y p-valor=0.6311. Por lo que se acepta H_0, es decir, la igualdad de varianzas.

2. Diferencia de dos medias con muestras emparejadas

3. Un dietista quiere estudiar si una dieta de adelgazamiento es efectiva. Para ello, anota el peso de cada uno de los 15 pacientes, antes y dos meses después de aplicar la dieta.

Paciente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Antes	90	94	105	103	85	90	89	78	87	88	91	92	87	88	90
Después	87	90	99	95	81	85	82	75	80	92	93	90	85	82	83

   a) Proporciona un intervalo de confianza al 99 % para la diferencia de medias (antes-despues)

   b) Efectúa el contraste apropiado para validar la eficiencia de la dieta al 95 %

   Nota: en la tabla de datos deben aparecer tantas filas como elementos muestrados, es decir 15 filas y dos variables numéricas una para el peso anterior a la dieta y otra para el posterior.

   Solución: Puesto que los individuos muestreados son los mismos en ambas muestras, las muestras no son independientes. El análisis inferencial correspondiente se denomina de muestras pareadas. Para llevarlo a cabo, se trabaja con la variable DIFERENCIA de peso de cada individuo, y se trabaja con UNA SOLA VARIABLE aunque originalmente se tienen dos variables. El modo automatizado para este proceso es la Ruta de menús: *Estadisticos–test t para datos relacionados...*

   * $IC_{99\,\%}(\mu_D)$=(1.25,6.48)

   * Consideramos $H_1 : \mu_D > 0$. Se obtiene t_{OBS}= 4.406 y p-valor=0.0003. Por lo que se rechaza H_0. Conforme a los datos obtenidos se tiene evidencia de que la dieta de adelgazamiento es efectiva con seguridad 95 %.

3. Diferencia de proporciones.

Si los datos se tienen resumidos se introducen con la ayuda del plugin IPSUR, siguiendo la ruta *estadisticos–proporciones-enter table for independent samples...*

Si los datos están incluidos uno a uno en una tabla de datos se sigue la ruta *estadisticos–proporciones-test de proporciones para dos muestras...*

4. Un productor decide cultivar dos variedades de tomate, valencia y perita. De la variedad valencia siembre 230 semillas y de la variedad perita 358. Pasadas de tres semanas el productor recorre el campo y registra que cantidad de semillas emergieron para cada variedad; 126 de valencia y 293 de perita. ¿Es razonable concluir que ambas variedades tienen el mismo éxito en la germinación?

   Haz el estudio mediante un intervalo y también mediante contrastes de hipótesis, ambos al al 99 %.

   Solución: utilizando el plugin IPSUR para Valencia (exito 126 y fracaso 104), para Perita (exito=293 y fracaso 65). Se tiene $IC_{99\,\%}(\pi_{Val} - \pi_{Per})$=(-0.37,-0.17) y en el contraste considerando como alternativa la diferencia en la proporcion se obtiene χ^2_{OBS}=50.0696 y p-valor=$1.484e^{-12}$ por lo que se rechaza H_0, es decir los datos evidencian una diferencia en la proporcion de germinación con seguridad 99 %

4. Ejercicios para entregar 2,3 y 4

1. Hojas de problemas de clase (ejercicio 1). Un fabricante de envases para alimentos congelados está valorando utilizar un nuevo tipo de material. La resistencia del material es muy importante, y se sabe que sigue una distribución normal. Se sabe que desviación de la resistencia es 1 psi(libras/pulgada^2)para los dos tipos de material. A partir de la muestra aleatoria de 10 envases fabricados con el plástico nuevo y otra de 12 con el clásico, se obtiene $\overline{x}_1 = 162'5$ y $\overline{x}_2 = 155'0$ respectivamente. La compañía no adoptará el material nuevo a menos que su resistencia media supere a la del material clásico en al menos 7 psi. Con base en la información muestral, se quiere decidir si debería utilizarse el nuevo material:

   a) Construye un intervalo de confianza y establece conclusiones al 95 %.

   b) Efectúa el contraste con $\alpha = 0{,}05$ que plantea como hipótesis alternativa que la diferencia entre resistencias medias del material nuevo y el clásico no llega a ser 7.

2. Se está haciendo un estudio sobre la cantidad de mercurio en los peces de un río. En esta parte del estudio quiere compararse lacantidad de mercurio en dos zonas del río; el curso medio y la desembocadura.

   Para ello se ha tomado una muestra de 35 peces en cada zona de forma independiente; los resultados son:

Curso medio	1,64 1,67 1,85 1,57 1,59 1,61 1,53 1,40 1,70 1,48 1,46 1,74 1,67 1,57 1,65 1,48 1,47 1,64 1,79 1,69 1,54 1,71 1,57 1,51 1,54 1,52 1,57 1,67 1,47 1,64 1,74 1,62 1,69 1,59 1,85
Desembocadura	1,56 1,55 1,69 1,67 1,60 1,68 1,65 1,59 1,75 1,49 1,69 1,48 1,62 1,48 1,70 1,65 1,67 1,69 1,76 1,59 1,61 1,67 1,69 1,53 1,57 1,62 1,42 1,71 1,54 1,71 1,56 1,67 1,68 1,60 1,66

   a) Realiza un estudio de normalidad.

   b) Estudia mediante un contraste si hay igualdad de varianzas.

   c) Construye un intervalo de confianza poara la diferencia de medias teniendo en cu enta el apartado anterior y establece conclusiones al 95 %.

   d) Efectúa el contraste con $\alpha = 0{,}05$ que plantea como hipótesis alternativa que no hay diferencia entre de la cantidad media de mercurio.

3. Se quiere estudiar si en los embalses de agua el nivel de Chl es el mismo en la cabecera que en la cola.

   Para ello se han tomado muestras en 10 embalses; el resultado ha sido:

Embalse	1	2	3	4	5	6	7	8	9	10
Cabecera	11,6	29,2	16,7	37,8	33,9	12,2	5,9	18,5	20,4	10,1
Cola	34,5	11,8	21,6	40,2	34,1	12,5	12,9	24,5	21,7	18,6

   a) Construye un intervalo de confianza y establece conclusiones al 99 %.

   b) Efectúa el contraste con $\alpha = 0{,}05$ que plantea como hipótesis alternativa que la diferencia entre el nivel medio de Clh en la cola del embalse es superior en al menos 2 unidades al nivel en la cabecera.

4. Con los datos de galtonR.RData, crea un nuevo factor de modo que si la edad es menor o igual que 23 se asocie a una nivel del factor denominado "joven" y para el resto de edades en otro nivel de factor denominado "mayor". Se quiere comparar la proporción de casados entre los jovenes y mayores.

a) Construye un intervalo de confianza para la diferencia de proporciones y establece conclusiones al 99 %.

b) Efectúa el contraste con $\alpha = 0{,}05$ que plantea que no hay fdiferencia en la proporción de casados entre los jovenes y mayores.

5. ANÁLISIS DE LA VARIANZA (ANOVA)

Posiblemente la técnica de análisis de la varianza sea una de las más aplicadas en el campo de la agronomía. Generaliza el estudio de diferencia de medias para dos muestras independientes para cuando se tienen más muestras.

Requisitos teóricos:

a) Muestras independientes.

b) Cada muestra debe evidenciar que procede de una variable con distribución normal.

c) Las muestras deben evidenciar que las varianzas (σ^2) de todas variables son iguales.

Ejemplo:

5. Algunas variedades de nematodos (gusanos redondos que viven en el suelo y que con frecuencia son tan pequeños que no se ven a simple vista) se alimentas de las raíces de los pastos y otras plantas. Esta plaga, que es especialmente peligrosa en climas templados como el nuestro, se puede combatir con nematicidas. Se reunieron datos del porcentaje de nematodos muertos para cuatro concentraciones de nematicidas (estas cantidades se dan en kg de ingrediente por hectárea)

Concentración nematicida

C20	C30	C50	C70
86	87	94	90
82	93	99	85
76	89	97	86
		91	

Deberemos introducir la información en dos variables, la primera con el porcentaje de nematodos y la segunda debe ser una variable factor que identifique la concentración y que tendrá 4 niveles.

Luego, X_i: se define como " % de nematodos muertos para la concentración i"

El contraste ANOVA con un factor fija en la hipótesis nula la igualdad de medias de las cuatro variables:

$$\begin{cases} \mathrm{H}_0: & \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ \mathrm{H}_1: & \text{Al menos una } \mu_i \text{ es distinta} \end{cases}$$

La salida que proporciona R-commander siguiendo la ruta:

Estadísticos- medias-ANOVA de un factor

```
            Df Sum Sq Mean Sq F value  Pr(>F)
Concent      3  345.6  115.20   8.634 0.00516 **
Residuals    9  120.1   13.34
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA	gradosLibertad	SumaCuadrados	MediaCuadrados	estadis F
entre niveles	niveles-1	SC(entre)	MC(entre)	$\frac{MC(entre)}{MC(dentro)}$
dentro niveles	numDatos-niveles	SC(dentro)	MC(dentro)	
TOTAL	numDatos-1	SC(total)		

El CONTRASTE ES UNILATERAL y el estadístico F es el cociente de dos varianzas. Una de ellas, el numerador, mide la parte de la variabilidad que se debe a la diferencia entre los centros de las diferentes muestras. La otra, el denominador, mide la parte de la variabilidad que se debe al hecho de que son muestras de variables aleatorias, es decir la variabilidad inevitable.

Si el valor de F_{OBS} es suficientemente grande (mayor que $F_{niveles-1,numDatos-1,\,\alpha}$) significa que entre las medias de los grupos hay diferencia, por lo que hay una evidencia muestral en contra de la hipótesis nula y por lo tanto se rechaza.

Además el p-valor (cola de la derecha) se interpreta como en el resto de los contrastes. Cuando el p-valor es menor que α (nivel de significación), hay evidencia muestral suficiente en contra de la hipótesis nula, por lo tanto se rechazaría H_0.