Práctica 6. Contrastes de hipótesis

Estadística

Grado en Ingeniería Informática

En esta práctica veremos contenidos relativos al Tema 6, esto es, resolveremos contrastes de hipótesis en poblaciones Normales, para proporciones y también el contraste χ^2 para asociación entre variables categóricas.

Comenzaremos recordando los pasos a seguir en la realización de cualquier contraste.

1. Establecer la hipótesis nula, H_0, y la hipótesis alternativa, H_a.
2. Fijar un nivel de significación: $\alpha = 0.05$, $\alpha = 0.10$ o $\alpha = 0.01$, los más usuales. Especificar el tamaño muestral.
3. Escoger adecuadamente un estadístico de contraste y establecer su distribución bajo la hipótesis nula, H_0.
4. Construir la región crítica o de rechazo, teniendo en cuenta si se trata de un contraste unilateral o bilateral.
5. Evaluar el estadístico en la muestra observada y vemos a qué región pertenece el valor observado del estadístico de contraste o, alternativamente, calcular el p-valor.
6. Obtener la conclusión: aceptar (es decir, no rechazar) o rechazar H_0, con la significación fijada (se requiere escribir una frase a modo de conclusión del contraste, ver ejercicios resueltos a modo de ejemplo).

1. Contrastes de hipótesis para una proporción

Supongamos que $X \in Ber(p)$ y que tenemos una m.a.s. $X_1, \ldots, X_n$ de tamaño n. El estadístico de contraste está basado en la proporción muestral $\widehat{p}$, y bajo la hipótesis nula verifica:

$$\frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1).$$

Ejercicio. El profesorado de la materia de Estadística del Grado en Ingeniería Informática han decidido hacer una encuesta entre su alumnado permitiéndoles escoger si preferirían ser evaluados a través de una prueba final de carácter oral o escrita. De los/las matriculados/as, 72 han respondido a la encuesta y de ellos/as 41 han manifestado su preferencia por la prueba escrita. A la vista de los resultados anteriores, ¿se puede afirmar que más de la mitad de los/las alumnos/as prefieren una prueba escrita? Redacta la conclusión del contraste suponiendo que el nivel de significación es $\alpha = 0.1$.

```
library(LearningStats)
n=72
exitos=41
proportion.test(x=exitos,n=n,p0=0.5,alternative = "greater",alpha=0.1)
```

2. Contrastes de hipótesis para una población Normal

Supongamos que $X \in N(\mu, \sigma^2)$, con σ^2 conocida y tenemos $X_1, \ldots, X_n$ una m.a.s. de X. Para realizar inferencia sobre μ a partir de una muestra, nos basaremos en la media muestral, $\overline{X}$. En este caso, el estadístico de contraste que utilizaremos será:

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \in N(0,1).$$

Si nos encontramos en el escenario de varianza desconocida, cambia el estadístico de contraste y consideramos:

$$\frac{\overline{X} - \mu_0}{S_c/\sqrt{n}} \in T_{n-1},$$

donde S_c^2 denota la cuasivarianza.

Si el objetivo es realizar un contraste sobre σ^2, el estadístico de contraste a considerar en ese caso es:

$$\frac{(n-1)S_c^2}{\sigma_0^2} \in \chi^2_{n-1}.$$

Ejercicio. Con objeto de mitigar el daño ocular de las pantallas digitales, se ha realizado un estudio sobre la capacidad visual de los/as trabajadores/as que usan estos dispositivos de manera continuada durante su jornada laboral. Se ha medido la agudez visual de 10 trabajadores/as, con los resultados que figuran a continuación en la escala Snellen que sigue una distribución Normal:

$$0.71 \quad 0.66 \quad 0.64 \quad 0.49 \quad 0.80 \quad 0.67 \quad 0.52 \quad 0.81 \quad 0.55 \quad 0.65$$

A la vista de los resultados anteriores, ¿se puede afirmar que la puntuación media de los/as trabajadores/as que usan estos dispositivos de manera continuada es superior a 0.55 puntos? Redacta la conclusión del contraste suponiendo que el nivel de significación es $\alpha = 0.1$.

```
library(LearningStats)
x=c(0.71, 0.66, 0.64, 0.49, 0.80, 0.67, 0.52, 0.81, 0.55, 0.65)
Mean.test(x=x,mu0=0.55,alternative = "greater",alpha=0.1)
```

Ejercicio. El tiempo de vida de los sistemas de ventilación empleados para los supercomputadores del CESGA sigue una distribución Normal. El CESGA ha recogido datos del tiempo de vida de 9 de sus sistemas de ventilación obteniendo los siguientes valores (en años):

$$9.9 \quad 8.7 \quad 10.2 \quad 10.5 \quad 9.6 \quad 9.2 \quad 9.8 \quad 10.9 \quad 9.8$$

A la vista de los resultados anteriores, ¿se puede afirmar que la varianza poblacional del tiempo de vida de los sistemas de ventilación difiere de 0.5 años^2? Redacta la conclusión del contraste suponiendo que el nivel de significación es $\alpha = 0.01$.

```
library(LearningStats)
tiempo=c(9.9,8.7,10.2,10.5,9.6,9.2,9.8,10.9,9.8)
variance.test(x=tiempo,sigma02=0.5,alternative = "two.sided",alpha=0.01)
```

3. Contrastes de hipótesis para dos poblaciones Normales

Supongamos que tenemos dos poblaciones independientes $X \in N(\mu_X, \sigma_X^2)$, $Y \in N(\mu_Y, \sigma_Y^2)$, de las que disponemos de m.a.s. de tamaños n y m, $X_1, \ldots, X_n$ y $Y_1, \ldots, Y_m$. Si las varianzas poblacionales son conocidas, el estadístico de contraste es:

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \in N(0,1).$$

En el caso de que las varianzas poblacionales sean desconocidas, tenemos que distinguir dos posibles escenarios: esas varianzas aún siendo desconocidas pueden asumirse iguales (será útil el contraste de comparación de varianzas), o no hay información al respecto y se suponen distintas.

En el primer caso, el estadístico de contraste es:

$$\frac{\overline{X}-\overline{Y}}{S_{c\,T}\sqrt{\frac{1}{n}+\frac{1}{m}}} \in T_{n+m-2},$$

donde recordemos que $S_{c\,T}$ es la estimación conjunta de la varianza:

$$S^2_{c\,T} = \frac{(n-1)S^2_{c_X} + (m-1)S^2_{c_Y}}{n+m-2}.$$

Si σ^2_X, σ^2_Y son desconocidas y distintas, se utilizará la aproximación de Welch, esto es,

$$\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{S^2_{c_X}}{n}+\frac{S^2_{c_Y}}{m}}} \in T_\delta$$

donde δ es el entero más próximo a:

$$\frac{\left(\frac{S^2_{c_X}}{n}+\frac{S^2_{c_Y}}{m}\right)^2}{\frac{\left(S^2_{c_X}/n\right)^2}{n-1}+\frac{\left(S^2_{c_Y}/m\right)^2}{m-1}}.$$

Ejercicio. Se ha realizado una prueba para evaluar si dos versiones de un algoritmo de búsqueda, A y B, tienen el mismo rendimiento en términos de tiempo de ejecución sobre un conjunto de datos idéntico. Se han ejecutado ambas versiones de los algoritmos de manera independiente en dos grupos de pruebas y se obtuvieron los siguientes resultados en segundos para el tiempo de ejecución de cada prueba:

Algoritmo A: 12.3, 15.4, 21.7, 17.2, 38.8, 42.1, 10.5, 23.3, 35.6, 28.4

Algoritmo B: 21.2, 18.6, 25.1, 14.7, 52.3, 65.2, 40.8, 43.4, 35.6, 42.0

Se supone que los tiempos de ejecución de ambos algoritmos siguen una distribución Normal con la misma varianza. ¿Podemos afirmar que existen diferencias significativas en los tiempos medios de ejecución entre los dos algoritmos a un nivel de significación del 5 %?

```
library(LearningStats)
A=c(12.3, 15.4, 21.7, 17.2, 38.8, 42.1, 10.5, 23.3, 35.6, 28.4)
B=c(21.2, 18.6, 25.1, 14.7, 52.3, 65.2, 40.8, 43.4, 35.6, 42.0)
diffmean.test(x1=A,x2=B,paired=F,var.equal=T,alternative = "two.sided",
+     alpha=0.05)
```

Para comparar las varianzas, σ_X^2 y σ_Y^2, utilizamos el siguiente estadístico de contraste:

$$\frac{S_{c_X}^2}{S_{c_Y}^2} \in F_{n-1,m-1}.$$

Ejercicio. Una empresa de tecnología está comparando la estabilidad de dos versiones de su software de gestión de redes en función de si hay una diferencia significativa en la variabilidad del uso de la CPU. Se han realizado pruebas midiendo el porcentaje de uso de la CPU durante un periodo de 30 minutos obteniendo:

Versión A (uso de CPU en %): 34.5, 36.1, 35.9, 33.4, 34.8, 36.2, 35.6, 34.7, 35.0, 34.3, 35.1, 34.9

Versión B (uso de CPU en %): 27.8, 28.3, 29.0, 28.5, 28.2, 29.3, 28.7, 29.5, 28.8, 28.4, 29.1, 28.6

Suponiendo que en ambos casos la variable sigue una distribución Normal, ¿es posible concluir que las varianzas del uso de la CPU entre las dos versiones de software son significativamente diferentes a un nivel de significación del 5 %.?

```
library(LearningStats)
A=c(34.5, 36.1, 35.9, 33.4, 34.8, 36.2, 35.6, 34.7, 35.0, 34.3, 35.1, 34.9)
B=c(27.8, 28.3, 29.0, 28.5, 28.2, 29.3, 28.7, 29.5, 28.8, 28.4, 29.1, 28.6)
diffvariance.test(x1=A, x2=B, alternative = "two.sided",alpha=0.05)
```

Cuando las muestras **no son independientes**, como por ejemplo cuando recogemos datos antes y después de alguna intervención sobre los mismos individuos, los resultados anteriores no pueden aplicarse. Consideraremos en ese caso la variable diferencia $D = X - Y$. Disponemos de dos m.a.s. $X_1, \ldots, X_n$ y $Y_1, \ldots, Y_n$, del mismo tamaño n (puesto que son muestras pareadas) y por tanto, también de una muestra de D, dada por $(X_1 - Y_1), \ldots, (X_n - Y_n)$. El estadístico de contraste será:

$$\frac{\overline{D}}{S_{c\,D}/\sqrt{n}} \in T_{n-1},$$

donde $\overline{D}$ y $S_{c\,D}$ son la media muestral y la cuasivarianza de la variable diferencia D.

Ejercicio. Para estudiar el impacto de un programa de capacitación en el rendimiento de la red de una empresa, se ha realizado el siguiente experimento con 11 técnicos de soporte. Antes de la capacitación, se midió el tiempo promedio de respuesta (en milisegundos) para resolver problemas de conectividad en la red de cada técnico. Después de un programa intensivo de formación sobre redes y troubleshooting, se midió nuevamente el tiempo promedio de respuesta de cada técnico. De este modo, se dispone de dos conjuntos de observaciones del tiempo medio de respuesta de los técnicos en milisegundos:

Técnico	1	2	3	4	5	6	7	8	9	10	11
Previo	68	77	94	73	37	131	77	24	99	629	116
Posterior	95	90	86	58	47	121	136	65	131	630	104

Suponiendo que los tiempos de respuesta siguen una distribución Normal, ¿hay pruebas suficientes para afirmar que el programa de capacitación ha reducido el tiempo de respuesta de los técnicos a un nivel de significación del 1 %?

```
library(LearningStats)
previo=c(68, 77, 94, 73, 37, 131, 77, 24, 99, 629, 116)
post=c(95, 90, 86, 58, 47, 121, 136, 65, 131, 630, 104)
diffmean.test(x1=previo,x2=post,paired=T,alternative = "less",alpha=0.01)
```

4. Contrastes de hipótesis para dos proporciones

Supongamos que tenemos $X \in Ber(p_X)$ e $Y \in Ber(p_Y)$, con m.a.s. $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$, de tamaño n y m respectivamente. Para comparar proporciones, utilizaremos el siguiente estadístico de contraste:

$$\frac{\widehat{p}_X - \widehat{p}_Y}{\sqrt{\frac{p_X(1-p_X)}{n} + \frac{p_Y(1-p_Y)}{m}}} \sim N(0,1).$$

Ejercicio. En un estudio realizado sobre la seguridad informática, se examinan dos grupos de usuarios en función de las condiciones de uso de sus sistemas. El primer grupo está compuesto por 307 usuarios que realizan su trabajo principalmente en redes no seguras, mientras que el segundo grupo está formado por 75 usuarios que trabajan en redes protegidas por sistemas de seguridad avanzados. En el grupo de usuarios de redes no seguras, 230 han sido afectados por ataques de *malware* durante el último año. En el grupo de usuarios de redes seguras, 30 fueron víctimas de ataques. A un nivel de significación del 1 %, ¿podemos concluir que la prevalencia de ataques de textitmalware es significativamente menor entre los usuarios de redes seguras?

```
library(LearningStats)
n1=307; exitos1=230
n2=75; exitos2=30
diffproportion.test(x1=exitos2, 2=exitos1,n1=n2,n2=n1,alternative="less",
+     alpha=0.01)
```

Este contraste también se podría plantear de la siguiente manera análoga:

```
library(LearningStats)
n1=307; exitos1=230
n2=75; exitos2=30
diffproportion.test(x1=exitos1, x2=exitos2,n1=n1,n2=n2,alternative="greater",
+     alpha=0.01)
```

5. Contraste χ^2 de independencia

Una de las propiedades estadísticas que nos puede interesar contrastar empleando este procedimiento es la independencia entre dos variables. Supongamos que tenemos una muestra aleatoria simple de tamaño n de una población. Sobre esta muestra, recogemos datos de dos variables X e Y, cualitativas. Estamos interesados en contrastar:

$$H_0: \ X \text{ e } Y \text{ son independientes}, \quad \text{vs.} \quad H_a: \ X \text{ e } Y \text{ están asociadas.}$$

El estadístico de contraste sería:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} \frac{(n_{ij} - \widehat{E}_{ij})^2}{\widehat{E}_{ij}} \in \chi^2_{(k-1)(p-1)},$$

donde $k =$ número de clases de X, $p =$ número clases de Y, $\widehat{E}_{ij}$ son los valores esperados bajo independencia y n_{ij} los valores observados (frecuencias absolutas). Se rechaza la hipótesis nula cuando el valor observado del estadístico de contraste es mayor que el cuantil $\chi^2_{(k-1)\cdot(p-1),1-\alpha}$.

Ejercicio. La siguiente tabla de contingencia muestra la relación entre el uso de diferentes métodos de autenticación en una empresa y la ocurrencia de brechas de seguridad en las cuentas de los empleados.

Autenticación	Brecha de seguridad (Sí)	Brecha de seguridad (No)
Contraseña simple	50	200
Autenticación de dos factores	15	180
Autenticación biométrica	5	195

A partir de esta tabla, ¿se puede afirmar que existe una relación significativa entre el tipo de autenticación y el riesgo de brechas de seguridad en las cuentas de los empleados?

```
x=matrix(c(50,15,5,200,180,195),nr=3,nc=2)
chisq.test(x)
```