Fundamentos para al Análisis de Datos y la Investigación

Tema 6 - Experimentos con datos y pruebas significativas

José María Sarabia

Máster Universitario en Ciencia de Datos
CUNEF Universidad

Contents

1 Introducción

Objetivos:

- Conocer y aplicar las pruebas A/B
- Conocer y aplicar las pruebas de tablas de contingencia 2×2 y las pruebas t
- Conocer y aplicar las pruebas múltiples A/B/C... (ANOVA) y el test de Kruskal-Wallis
- Conocer la teoría de los contrastes de hipótesis. Saber establecer las hipótesis y usar el p-valor, para su aplicación a problemas de estimación.
- Uso de comandos y paquetes de R para estimación y contraste.

Contents

2 Pruebas A/B

Prueba A/B

- Una prueba A/B es un experimento con dos grupos de sujetos para establecer cual de los dos tratamientos, productos o procedimientos o similar es mejor.
- A menudo uno de los tratamientos es el tratamiento estándard, denominado *control*. La hipótesis habitual es que el nuevo tratamiento es mejor que el control.
- Las pruebas A/B son habituales tanto en el diseño web como en marketing, ya que los resultados se evalúan fácilmente. En ciencia de datos las pruebas A/B se usan normalmente en un contexto web.
- Es importante poner atención en el *estadístico de prueba*, o *métrica* que se utiliza para comparar A con B. Quizás la métrica más frecuente en ciencia de datos es una *variable binaria* (hacer clic o no), comprar o no, hay fraude o no etc. Esto supone trabajar con una tabla de contingencia 2×2
- Otra posibilidad es que la *métrica* sea una variable continua (importe de compra, tiempo en la web, etc)

3 Pruebas t y de proporciones

- Existen diversas pruebas A/B significativas, *dependiendo* si los datos son de conteo o 1/0, según el número de muestras y que se está midiendo.
- Si los datos con de conteo (continuos) se usa la prueba t y si los datos son de proporciones (tabla de contingencia) el test de propociones.
- Para el caso de datos de conteo, una de las pruebas más usadas es la t o t-test (t de Student), que permite aproximar la distribución de una media muestral.
- Esta prueba requiere que los datos sean numéricos (mediciones). En R el comando es `t.test()`
- Si los datos son de proporciones se usa en R el comando `prop`

Contents

4 Pruebas múltiples

- En esta situación supongamos que en lugar de una prueba A/B, tenemos que comparar varios grupos, por ejemplo A/B/C/D, siempre con datos numéricos.
- El procedimiento estadístico adecuado que permite probar que existen diferencias significativas entre los grupos, se denomina *análisis de la varianza* (analysis of variance) o ANOVA para el caso de normalidad y de Kruskal-Wallis sin suponer normalidad en los datos.
- Los comandos en R son `aov()` y `kruskal.test()`
- Veamos el modelo de pruebas múltiples o ANOVA a través del siguiente ejemplo.

4 Pruebas múltiples

Prueba A/B/C/D: La siguiente tabla muestra la adherencia (número de segundos que un visitante ha estado en la página) de cuatro páginas web. Las cuatro páginas se cambian para que cada visitante de la web reciba una al azar. Hay un total de cinco visitantes por cada página y cada columna de la tabla es un conjunto de datos independientes.

	Página 1	Página 2	Página 3	Página 4
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Medias	172	185	176	162
Media general				173.75

4 Pruebas múltiples

En relación con el modelos anterior nos hacemos las siguientes preguntas:

1. ¿Existen diferencias importantes entre las cuatro páginas?
2. Si efectivamente hay diferencias: ¿Es A diferentes de B?, ¿Es A diferentes de C? y ¿Es A diferentes de C?, etc.

Las hipótesis del modelos ANOVA son las siguientes:

1. Los datos de las columnas tienen que ser independientes entre si
2. Las distribuciones de las columnas deben ssguir distribuciones normales
3. Se tiene que cumplir la hipótesis de homcedasticidad, es decir, las varianzas de las variables deben ser iguales.

4 Pruebas múltiples: test de Kruskal-Wallis

- El contraste de Kruskal-Wallis es la versión *no paramétrica* del modelo de pruebas múltiples o ANOVA
- El contraste estudia si un conjunto de k muestras independientes proceden de la misma población
- Los datos están formados por muestras

$$(X_{i1}, \ldots, X_{in_i}), \quad i = 1, 2, \ldots, k$$

  donde $\sum_{i=1}^{k} n_i = n$

- La hipótesis nula establece que las muestras proceden de distribuciones idénticas. El test se basa en la suma de *rangos*. Para ello se ordenan el total de observaciones de menor a mayor y se asignan rangos r_{ij}, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n_i$, de modo que a la observación más pequeña se asigna un 1 y a la mayor n

4 Pruebas múltiples: test de Kruskal-Wallis

- A continuación se suman los rangos de cada una de las muestras

$$R_i = \sum_{j=1}^{n_i} r_{ij}, \quad i = 1, 2, \ldots, k$$

- Finalmente se considera el contraste definido por medio del estadístico H:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

que se distribuye aproximadamente como una χ^2 con $k-1$ grados de libertad, siempre que $n_i > 5$.

- La función en R es

```
kruskal.test(variable  grupos, data = datos)
```

- En análisis post-hoc, es posible contratar en qué grupos existen diferencias significativas

Contents

5 Prueba chi cuadrado y bondad de ajuste

- Se trata de estudiar si un conjunto de datos *procede de una distribución de probabilidad conocida*. Ya estudiamos el contraste de Shapiro-Wilk para el caso de normalidad
- Dicha distribución puede estar totalmente especificada o depender de un conjunto de parámetros que se estiman a partir de la muestra
- Estudiaremos los siguientes contastes:
  - Contraste de la chi-cuadrado de Pearson (variables discretas)
  - Contraste de Kolmogorov-Smirnov (variables continuas)

5 Contrastes de la chi-cuadrado de Pearson

- Partimos de un conjunto de sucesos $E_1, \ldots, E_k$ tales que

$$P(E_i) = p_i,\ i = 1, 2, \ldots, k;\ \sum_{i=1}^{k} p_i = 1$$

- Se realiza un experimento aleatorio y se obtienen frecuencias $n_1, \ldots, n_k$ de $E_1, \ldots, E_k$ donde $\sum_{i=1}^{k} n_i = n$
- Se establecen las hipótesis:
  H_0: *los datos proceden de la distribución* $\{p_i\}$, $i = 1, 2, \ldots, k$
  H_1: *los datos no proceden de la distribución* $\{p_i\}$
- La región de rechazo se basa en el estadistico χ^2 de Pearson,

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i},$$

cuya distribución es una chi-cuadrado de Pearson con $k - 1$ grados de libertad. Si el valor de χ^2 es grande, se rechaza la hipótesis nula

5 Contraste de Kolmogorov-Smirnov

- Tenemos que contrastar si un conjunto de datos procede de una distribucion continua F_0
- Las hipótesis son: H_0 : los datos proceden de $F = F_0$ frente a $H_1 : F \neq F_0$
- La región crítica es:

$$D = \max_x |F_n(x) - F_0(x)|$$

  donde $F_n(x)$ representa la función de distribución empírica de los datos
- Es importante saber que la distribución del estadístico D bajo H_0 no depende de la distribución de F_0
- El test se realiza mediante el función `ks.test(datos,pdist)`

Contents

6 Contrastes en tablas de contingencia

- La idea de este contraste es estudiar si existen o no diferencias significativas entre k poblaciones de las cuáles se han extraido muestras aleatorias simples
- Suponemos que disponemos de m clases en las que se han dividido las k poblaciones, de modo que disponemos de una tabla de contingencia como la que se muestra a continuación
- La hipótesis nula es H_0 : las k poblaciones son homogéneas frente a la alternativa H_1 de que las k poblaciones no son homogéneas.

Muestras / Clases	c_1	c_2	$\cdots$	c_m	
M_1	n_{11}	n_{12}	$\cdots$	n_{1m}	n_{1+}
M_2	n_{21}	n_{22}	$\cdots$	n_{2m}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
M_k	n_{k1}	n_{k2}	$\cdots$	n_{km}	n_{k+}
	n_{+1}	n_{+2}	$\cdots$	n_{+m}	n

6 Contrastes en tablas de contingencia

- Para realizar el contraste, comparamos la tabla original con la tabla construida suponiendo homogeneidad mediante el estadístico chi-cuadrado, para obtener

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - n_{i+} n_{+j}/n)^2}{n_{i+} n_{+j}/n}$$

- El estadístico se distribuye (en el límite) según una distribución chi-cuadrado de Pearson con $(k-1) \times (m-1)$ grados de libertad

6 Contrastes en tablas de contingencia

- En el caso del contraste de independencia suponemos dos caracteres A y B con k y m modalidades, respectivamente, de modo que los n individuos de la muestra se clasifican en una tabla de doble entrada o de contingencia similar a la antes vista.
- Para realizar el contraste, comparamos la tabla original con la tabla construida suponiendo independencia:

$$P(A_iB_j) = P(A_i)P(B_j) \Rightarrow \frac{n_{ij}}{n} = \frac{n_{i+}}{n} \cdot \frac{n_{+j}}{n}$$

- Se establece la hipótesis nula H_0: las variables A y B son independientes
- Análogamente al caso anterior, el estadístico chi-cuadrado χ^2 sigue aproximadamente una distribución chi-cuadrado de Pearson con $(k-1) \times (m-1)$ grados de libertad

6 Contrastes en tablas de contingencia

- Los dos contrastes se calculan de la misma forma. Sin embargo, en el contraste de homogeneidad los totales de las marginales las fija el investigador, que decide cuantos individuos se han de elegir en cada población
- Sin embargo, en el contraste de independencia el investigador fija el valor de n, de modo que las marginales no vienen determinadas de antemano
- Los comandos en R para dichos contraste, así como sus principales elementos son:
  - `chisq.test(tabla)`
  - `chisq.test(tabla)$expected`
  - `chisq.test(tabla)$statistic`
- En los dos contrastes, los datos pueden proceder de dos fuentes: de una fuente primaria (encuesta), y para la tabulación hacemos uso de `table()`; de una fuente secundaria, de modo que la tabla de resultados viene dada

Contents

7 Contraste de aleatoriedad de Wald-Wolfowitz

- Se trata de estudiar si un conjunto de datos son independientes e igualmente distribuidos
- Las hipótesis son: H_0 : los datos son i.i.d.; H_1 : los datos no son i.i.d.
- La región crítica del test se basa en conjuntos de rachas por encima y debajo de la mediana o bien de un punto de corte:

$$R = \frac{U - \left(\frac{2n_1 n_2}{n1+n_2} + 1\right)}{\sqrt{\frac{2n1n_2(2n_1 n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}}$$

- El contraste se realiza mediante el paquete `randtests` y el comando `runs.test`

Contents

8 Contraste de hipótesis

- Muchos de los problemas en estadística y en CD consisten en decidir una de dos hipótesis a partir de la evidencia proporcionada por los datos.
- En un *contraste de hipótesis* se establecen dos hipótesis: *la hipótesis nula* H_0 y la *hipótesis alternativa* H_1
- La hipótesis nula es la que se establece de forma neutra, y no se rechaza a no ser que los datos digan los contrario. La hipótesis nula es la hipótesis de partida.
- La teoría de los contrastes consiste en aportar evidencias en contra de la hipótesis nula. Esto se realiza mediante la *región crítica o region de rechazo*

8 Contraste de hipótesis

- A la hora e realizar un contraste se pueden comenter dos tipos de errores:
  - *Error tipo I*: Es el error que se comete cuando se rechaza H_0 cuando es cierta. Su probabilidad se denota por α
  - *Error tipo II*: Es el error que se comete cuando se acepta H_0 siendo falsa. Su probabilidad se denota por β

8 Contraste de hipótesis

- En la práctica, para rechazar o no la hipótesis nula se trabaja con la *probabilidad crítica o p-valor*
- El p-valor es el mínimo valor a partir del cual se rechaza la hipótesis nula. El p-valor es la probabilidad bajo H_0 de obtener un resultado menos compatible con H_0 que el obtenido por la muestra.
- En la práctica, se elige un umbral de probabilidad (habitualmente el 5 o el 1 por ciento), de modo que si el p-valor es menor que el 5 por ciento, se rechaza la hipótesis nula, mientras que si es mayor del 5 por ciento no se rechaza

8 Contraste de hipótesis

- Un aspecto importante es establecer correctamente las hipótesis nula y alternativa
- Un primer aspecto es que la hipótesis nula no se demuestra, únicamente se rechaza o no se rechaza
- Si se desea estudiar un determinado efecto, en la hipótesis nula se establece el hecho de que no hay efecto, de modo que los datos probarán o no si se rechaza dicha hipótesis. Hay que tener en cuenta que la hipótesis nula es la hipótesis neutra
- Existen diversos tipos de hipótesis. Una *hipótesis simple* es del tipo $H_0 : \theta = \theta_0$. Si $H_1 : \theta \neq \theta_0$, se trata de un *contraste bilateral* o de dos colas. Si $H_1 : \theta < \theta_0$ o bien $H_1 : \theta > \theta_0$ se trata de un *contraste unilateral*.

8 Estimación y contraste en R

- **En el módulo base** existen tres tipos básicos de intervalos de confianza y contrastes de hipótesis de medias, varianzas y proporciones. En los tres casos se pueden especificar varias hipótesis alternativas, unilaterales o bilaterales.
- **Estimación de media o diferencia de medias**: se usa la función `t.test(x, y, paired=FALSE, var.equal=False, conf.level=0.95)`
- **Estimación del cociente de varianzas**: `var.test(x, y, conf.level=0.95)`
- **Estimación de una proporción o diferencia de proporciones**: `prop.test(x, n)`

Fundamentos para al Análisis de Datos y la Investigación

Tema 6 - Experimentos con datos y pruebas significativas

José María Sarabia

Máster Universitario en Ciencia de Datos
CUNEF Universidad