multiple testing

luis manuel

18/3/2021

En esta lección, analizaremos las pruebas múltiples. Podrías preguntar, "¿Qué es eso?"

Dado que los datos son valiosos y nos gustaría aprovecharlos al máximo, podríamos usarlos para probar varias hipótesis. Si tenemos un nivel alfa de .05 y probamos 20 hipótesis, entonces, en promedio, esperamos un error, solo por casualidad.

Otro problema potencial es que después de ejecutar varias pruebas, solo se puede informar el valor p más bajo O todos los valores p por debajo de algún umbral pueden considerarse significativos. Sin duda, algunos de estos serían falsos.

Afortunadamente, tenemos formas inteligentes de minimizar los errores en esta situación. Eso es lo que abordaremos. Definiremos medidas de error específicas y luego formas estadísticas de corregirlas o limitarlas.

Las pruebas múltiples son particularmente relevantes ahora en esta era de datos GRANDES. A los estadísticos se les encargan preguntas como "¿Qué variables importan entre las miles medidas?" y "¿Cómo se relaciona la información no relacionada?"

El valor p es "la probabilidad bajo la hipótesis nula de obtener evidencia tan o más extrema que su estadística de prueba (obtenida de sus datos observados) en la dirección de la hipótesis alternativa". Por supuesto, los valores p están relacionados con la significancia o los niveles alfa, que se establecen antes de que se realice la prueba (a menudo en 0,05).

Si se encuentra que un valor p es menor que alfa (digamos 0.05), entonces el resultado de la prueba se considera estadísticamente significativo, es decir, sorprendente e inusual, y la hipótesis nula (el status quo) es rechazada

Ahora considere este cuadro copiado de http://en.wikipedia.org/wiki/Familywise_error_rate. Supongamos que hemos probado m hipótesis nulas, m_0 de las cuales son realmente verdaderas y m-m_0 son en realidad falsas. De las pruebas m, R se ha declarado significativo, es decir, los valores p asociados fueron menores que alfa, y m-R fueron resultados no significativos o aburridos.

Otro nombre para un error de Tipo II es Falso Negativo, ya que afirma falsamente un resultado no significativo (negativo).

Una rosa con cualquier otro nombre, ¿verdad? Considere la fracción V/R. La R observada representa el número de resultados de prueba declarados significativos. Estos son "descubrimientos", algo diferente del status quo. V es el número de aquellos falsamente declarados significativos, por lo que V/R es la proporción de descubrimientos FALSOS. Dado que V es una variable aleatoria (es decir, desconocida hasta que hacemos un experimento), llamamos al valor esperado de la relación, E[V/R], la Tasa de Descubrimiento Falso (FDR).

Una rosa con cualquier otro nombre, ¿verdad? ¿Qué tal la fracción V/m_0? En el gráfico, m_0 representa el número de H_0 verdaderos y se desconoce m_0. V es el número de los falsamente declarados significativos, por lo que V/m_0 es la proporción de FALSOS positivos. Dado que V es una variable aleatoria (es decir, desconocida hasta que hacemos un experimento), llamamos al valor esperado de la razón, E (V/m_0), la tasa FALSO POSITIVO.

A la probabilidad de al menos un falso positivo, Pr (V> = 1), la llamamos Tasa de error familiar (FWER).

	Null hypothesis is True	Alternative hypothesis is True	Total
Declared significant	V	S	R
Declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

- m_0 is the number of true null hypotheses, an unknown parameter
- R is an observable random variable, while S, T, U, and V are unobservable random variables.

Figure 1: A caption

Entonces, ¿cómo controlamos la tasa de falsos positivos?

Supongamos que somos realmente inteligentes, calculamos nuestros valores p correctamente y declaramos todas las pruebas con p <alfa como significativas. Esto significa que nuestra tasa de falsos positivos es como máximo alfa, en promedio.

Supongamos que realizamos 10,000 pruebas y alfa = .05. Esperamos 500 falsos positivos, 500 falsos positivos parece mucho. ¿Cómo evitamos tantos?

Podemos intentar controlar la tasa de error familiar (FWER), la probabilidad de al menos un falso positivo, con la corrección de Bonferroni, la corrección de prueba múltiple más antigua.

Es muy sencillo. Hacemos m pruebas y queremos controlar el FWER en el nivel alfa para que Pr (V> = 1) <alfa. Simplemente reducimos alfa dramáticamente. Establezca alpha_fwer en alpha / m. Solo llamaremos significativo a un resultado de prueba si su valor p <alpha_fwer.

Suena bien, ¿verdad? Fácil de calcular. ¿Cuál sería el inconveniente de este método?, el problema seria que demasiados resultados fallarán

Otra forma de limitar la tasa de falsos positivos es controlar la tasa de falsos descubrimientos (FDR). Recuerde que esto es E [V/R]. Esta es la corrección más popular cuando se realizan muchas pruebas. Se utiliza en muchas áreas, como genómica, imágenes, astronomía y otras disciplinas de procesamiento de señales.

Nuevamente, haremos m pruebas pero ahora configuraremos el FDR o E [V/ R] en el nivel alfa. Calcularemos los valores p como de costumbre y los ordenaremos de menor a mayor, p_1, p_2, ... p_m. Llamaremos significativo a cualquier resultado con p_i <= (alpha * i) / m. Este es el método Benjamini-Hochberg (BH). Un valor p se compara con un valor que depende de su clasificación.

Esto es equivalente a encontrar el k más grande tal que p_k <= (k * alfa) / m, (para un alfa dado) y luego rechazar todas las hipótesis nulas para i = 1, ..., k.

Al igual que la corrección de Bonferroni, es fácil de calcular y mucho menos conservadora. Podría dejar pasar más falsos positivos y puede comportarse de manera extraña si las pruebas no son independientes.

Ahora considere este cuadro copiado de las diapositivas. Muestra los valores p para 10 pruebas realizadas en el nivel alfa = .2 y tres líneas de corte. Los valores p se muestran en orden de izquierda a derecha a lo largo del eje x. La línea roja es el umbral para Sin correcciones (los valores p se comparan con alfa = .2), la línea azul es el umbral de Bonferroni, alfa = .2 / 10 = .02 y la línea gris muestra la corrección BH. Tenga en cuenta que no es horizontal pero tiene una pendiente positiva como esperamos.

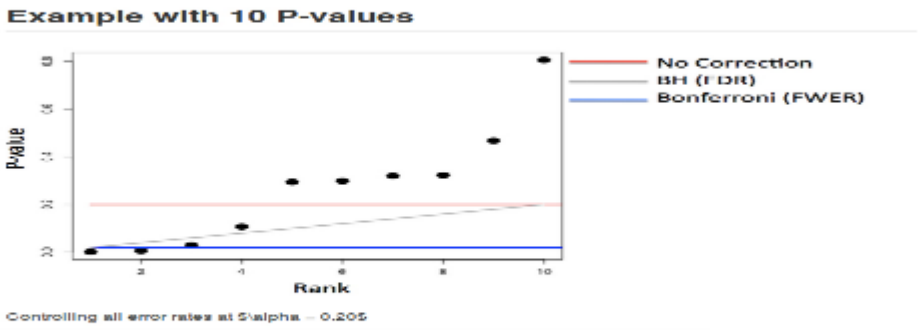

Figure 2: A caption

Así que Bonferroni pasó solo la mitad de los resultados que pasó el método Sin corrección (comparando valores p con alfa).

Entonces, la corrección de BH que limita el FWER está entre la No Corrección y el Bonferroni. Es más conservador (menos resultados significativos) que el Sin corrección, pero menos conservador (resultados más significativos) que el Bonferroni. Tenga en cuenta que con este método el umbral es proporcional a la clasificación de los valores, por lo que tiene una pendiente positiva mientras que los otros dos umbrales son planos.

Observe cómo los métodos de Bonferroni y BH ajustaron el nivel de umbral (alfa) para rechazar las hipótesis nulas. Otro enfoque correctivo equivalente es ajustar los valores p, por lo que ya no son valores p clásicos, pero se pueden comparar directamente con el alfa original.

Suponga que los valores p son p_1, . . . , p_m. Con el método de Bonferroni, los ajustaría estableciendo p'_i = max (m * p_i, 1) para cada valor p. Entonces, si llama a todo p'_i <alpha significativo, controlará el FWER.

Para demostrar algunos de estos conceptos, hemos creado una serie de valores p para usted. Tiene una longitud de 1000 y es el resultado de una regresión lineal realizada en pares x, y normales aleatorios, por lo que no existe una verdadera relación significativa entre las x y las y.

```
pValues<-read.csv("C:/Users/luism/Documents/pv1.csv")$x
pValues2<-read.csv("C:/Users/luism/Documents/pv2.csv")$x
head(pValues)
```

```
## [1] 0.5334915 0.2765785 0.8380943 0.6721730 0.8122037 0.4078675
```

```
sum(pValues < 0.05)
```

```
## [1] 51
```

Así que obtuvimos alrededor de 50 falsos positivos, tal como esperábamos (.05 * 1000 = 50). La belleza de R es que proporciona una gran cantidad de funciones estadísticas integradas. La función p.adjust es un ejemplo. El primer argumento es la matriz de pValues. Otro argumento es el método de ajuste. Una vez más, use la función R suma y una expresión booleana usando p.adjust con method = "bonferroni" para controlar el FWER.

```
sum(p.adjust(pValues,method="bonferroni") < 0.05)
```

```
## [1] 0
```

Entonces, la corrección eliminó todos los falsos positivos que habían pasado la prueba alfa sin corregir. Repite el mismo experimento, esta vez usando el método "BH" para controlar el FDR.

Entonces, el método BH también eliminó todos los falsos positivos. Ahora hemos generado otra matriz de valores p de 1000 de largo, esta llamada pValues2. En estos datos, la primera mitad (500 pares x / y) contiene valores xey que son aleatorios y la segunda mitad contiene pares xey que están relacionados, por lo que ejecutar un modelo de regresión lineal en los 1000 pares debería encontrar algunos (relación no aleatoria).

También creamos una matriz de cadenas de caracteres de 1000 de largo, trueStatus. Las primeras 500 entradas son "cero" y las últimas son "distintas de cero". Utilice la cola de la función R para ver el final de trueStatus.

```
trueStatus<-read.csv("C:/Users/luism/Documents/tru.csv")$x
tail(trueStatus)
```

```
## [1] "not zero" "not zero" "not zero" "not zero" "not zero" "not zero"
```

Una vez más, podemos usar la grandeza de R para contar y tabular por nosotros. Podemos llamar a la tabla de funciones R con dos argumentos, un booleano como pValues2 <.05 y el arreglo trueStatus. El booleano obviamente tiene dos resultados y cada entrada de trueStatus tiene uno de dos valores posibles. La tabla de funciones alinea los dos argumentos y cuenta cuántos de cada combinación (VERDADERO, "cero"), (VERDADERO, "no cero"), (FALSO, "cero") y (FALSO, "no cero") aparecen. Pruebalo ahora.

```
table(pValues2 < 0.05, trueStatus)
```

```
##        trueStatus
##         not zero zero
##   FALSE        0  476
##   TRUE       500   24
```

Vemos que sin ninguna corrección, las 500 pruebas verdaderamente significativas (no aleatorias) se identificaron correctamente en la columna "no cero". En la columna cero (las pruebas verdaderamente aleatorias), sin embargo, 24 resultados se marcaron como significativos.por lo tanto el porcentaje de falsos positivos Tal como esperábamos, alrededor del 5% o .05 * 100.

Ahora ejecute la misma función de tabla, sin embargo, esta vez use la llamada a p.adjust con el método "bonferroni" en la expresión booleana. Esto controlará el FWER.

```r
table(p.adjust(pValues2,method="bonferroni") < 0.05, trueStatus)
```

```
##        trueStatus
##         not zero zero
##   FALSE       23  500
##   TRUE       477    0
```

Dado que el método de corrección de Bonferroni es más conservador que simplemente comparar valores p con alfa, todas las pruebas verdaderamente aleatorias se identifican correctamente en la columna cero. En otras palabras, no tenemos falsos positivos. Sin embargo, el umbral se ha ajustado tanto que 23 de los resultados verdaderamente significativos se han identificado erróneamente en la columna distinta de cero.

Ahora ejecute la misma función de tabla una última vez. Utilice la llamada a p.adjust con el método "BH" en la expresión booleana. Esto controlará la tasa de descubrimiento falso.

```r
table(p.adjust(pValues2,method="BH") < 0.05, trueStatus)
```

```
##        trueStatus
##         not zero zero
##   FALSE        0  487
##   TRUE       500   13
```

Una vez más, los resultados son un compromiso entre las No Correcciones y Bonferroni. Todos los resultados significativos se identificaron correctamente en la columna "no cero", pero en la columna aleatoria ("cero") 13 resultados se identificaron incorrectamente. Estos son los falsos positivos. Esto es aproximadamente la mitad del número de errores en las otras dos ejecuciones.

Aquí hay una gráfica de los dos conjuntos de valores p ajustados, Bonferroni a la izquierda y BH a la derecha. El eje x indica los valores p originales. Para Bonferroni, (ajustando multiplicando por 1000, el número de pruebas), solo algunos de los valores ajustados están por debajo de 1. Para el BH, los valores ajustados son ligeramente mayores que los valores originales.

Concluiremos diciendo que las pruebas múltiples son un subcampo completo de inferencia estadística. Por lo general, una corrección básica de Bonferroni / BH es lo suficientemente buena para eliminar los falsos positivos, pero si existe una fuerte dependencia entre las pruebas, puede haber problemas. Otro método de corrección a considerar es "BY".

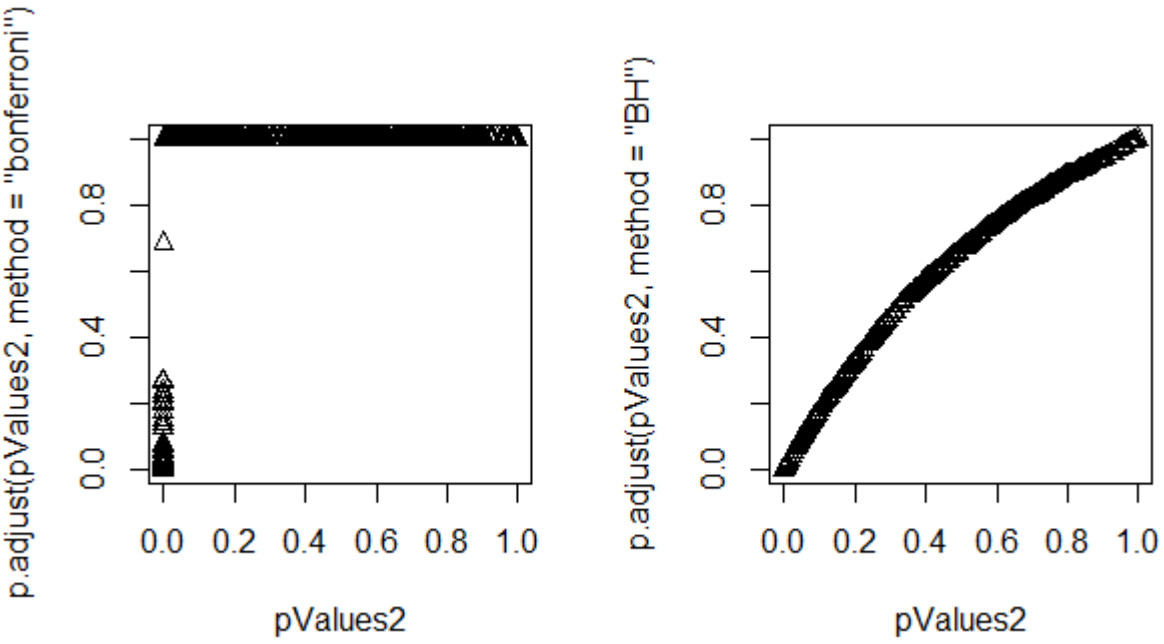

Figure 3: A caption