FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Introdución ao análise multivariante en bioestatística

Andrea García Pérez

2020/2021

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Introdución ao análise multivariante en bioestatística

Andrea García Pérez

07/2021

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estadística e Investigación Operativa
Título: Introdución ao análise multivariante en bioestatística
Breve descrición do contido
O obxectivo deste TFG é o de revisar algunhas das técnicas que se aplican no análise estatístico de datos no campo das ciencias da saúde. Para elo, revisaranse diferentes métodos estatísticos que permiten reducir a complexidade dos datos. Tamén se realizará unha revisión das técnicas que permiten coñecer que variables aleatorias teñen relevancia á hora de obter conclusións. Finalmente, partindo dun conxunto de datos biosanitarios, aplicarémoslle os métodos revisados coa finalidade de obter conclusións relevantes sobre ditos datos.
Recomendacións
O tratamento dos datos realizarase utilizando o software estatítico de uso libre R.
Outras observacións

Índice xeral

Resumo

Unha das posibles aplicacións das matemáticas é no ámbito sanitario, por exemplo para controlar a efectividade de medicamentos. Para levar isto a cabo, debemos controlar os posibles erros e asegurar fiabilidade dos nosos estudos. Unha das situacións nas que é importante este control é cando se levan a cabo centos ou miles de probas o que nos leva a ter que facer certos axustes para non superar o marxe de erro inicial fixado. Ao longo deste traballo comezamos cos casos sinxelos dun único contraste, nos que se pretende facer inferencia sobre a media ou a comparación de medias entre dúas poboacións, para sentar as bases antes de introducirnos en múltiples contrastes simultáneos. Desenvolvemos o significado dos termos máis importantes á hora de controlar os tipos de erros que se poden cometer, chegando á explicación do funcionamento dos distintos métodos de corrección que podemos aplicar para manter controlada a probabilidade de cometer un certo tipo de erro. Para concluír o traballo decidimos expoñer todos os conceptos sobre un exemplo no que a partir dun conxunto de información sobre a expresión de 2000 xenes en persoas sans e en persoas enfermas, quereremos saber se a expresión en media dos xenes variará en función do estado da saúde da persoa.

Abstract

One of the possible applications of mathematics is in the healthcare field, for example to monitor the effectiveness of medicines. We must minimize possible errors and ensure reliability of our studies to accomplish this. One of the situations in which this control is important is when hundreds or thousands of tests are performed which leads us to make certain adjustments so as not to exceed the initial set margin of error. Throughout this work we begin with the simple cases of a single contrast, in which it is intended to make an inference about the mean or comparison of means hypothesis testing between two populations, to lay the groundwork before introducing ourselves to multiple simultaneous

contrasts. We developed the meaning of the most important terms when controlling the types of errors that can be made, coming to the explanation of the operation of the different correction methods that we can apply to maintain a certain controlled probability of a particular type of error. To conclude the work we decided to set out all the concepts on a specific example in which from a set of information on the expression of 2000 genes in healthy people and in sick people, we want to know if the average expression of genes will vary depending on the state of the person's health.

Introdución

Partindo da base da inferencia estatística, é dicir, pretendemos extraer conclusións para unha poboación a partir dun resultado obtido tras observar unha mostra. Segundo se coñezamos ou non a forma de distribución de probabilidade levaremos a cabo unha inferencia paramétrica ou non paramétrica, respectivamente. Tanto en paramétrica coma en non paramétrica podemos centrarnos en parámetros ou características (coma distribucións) buscando estimacións dos mesmos, o que nos levará a intervalos de confianza, é dicir, a un rango de valores posibles entre os que se atopará o valor real do parámetro que queremos estimar cunha certa probabilidade (nivel de confianza). A amplitude de ditos intervalos será en relación a un marxe de erro, máis concretamente será o erro típico multiplicado por un pequeno valor, un cuantil da distribución na mostraxe que teñamos. Tamén cabe mencionar que poden aparecer intervalos que non sexan simétricos respecto do parámetro que buscamos estimar, pero nós non imos centrarnos nisto.

Chegaremos así tamén aos contrastes de hipóteses, problemas baseados en responder preguntas concretas sobre a poboación, ser capaces de decidir sobre a veracidade de certas hipóteses. Coa finalidade de ilustrar algúns conceptos vexamos o exemplo 1.1.

Exemplo 1.1. Partimos dunha mostra de dúas poboacións, unha de mulleres e outra de homes (entre $25-45$ anos que é cando a cantidade da hormona se mantén máis estable), ambas de tamaño 1000. En cada unha dela analizamos a cantidade de hormona luteinizante (LH) en sangue (encargada do desenvolvemento sexual que se produce pola glándula pituitaria) e obtemos que, en media, a cantidade nas mulleres é de 17,23 e nos homes 21,57 en unidades por litros de sangue. Queremos entón saber se podemos afirmar a un nivel do 5 % que a cantidade media de LH en sangue depende do sexo, é dicir, se é distinta entre os homes e as mulleres. As cantidades medias obtidas son datos aleatorios debido a que a cantidade de dita hormona vese afectada por moitos factores. Para saber máis sobre ela véxase [16].

Comecemos falando dun estimador, que é o valor concreto que tomou a nosa estatística de contraste na realización mostral. Xeralmente un bo estimador para un parámetro da

		Decisión	
		Aceptar H_0	Rexeitar H_0
Realidade	H_0 certa	**Correcto**	**Erro tipo I**
	H_0 falsa	**Erro tipo II**	**Correcto**

Cadro 1: Posibles resultados tras aplicar un contraste de hipóteses.

poboación é aquel que se aproxima a dito parámetro (inesgado, cuxa esperanza matemática coincida co valor do parámetro a estimar) e presenta a menor varianza posible. O parámetro que queiramos estimar adoita ter unha distribución na mostraxe, polo que o estimador do parámetro vai presentar unha distribución asintótica coñecida e, nalgúns casos incluso ata tabulada. Isto vainos servir para calcular os cuantís das distribucións e obter os erros típicos que acaban dando lugar á amplitude dos intervalos de confianza.

Seguindo un pouco a idea de [12] imos empezar a falar dos contrastes de hipóteses, centrándonos sobre todo nos contrastes para a media que resultan de maior interese para o análise de datos que levaremos a cabo. Por un lado falaremos da hipótese nula, H_0, que é a que se da por certa antes de obter unha mostra. Por exemplo H_0: "En media, a cantidade de hormona LH en homes e mulleres é a mesma". Por outro lado a hipótese alternativa, H_a, que é o que sucede cando a nula non é certa. H_a: "En media, a cantidade de LH en homes e mulleres é distinta". Isto é, no noso exemplo, estaremos interesados en contrastar:

$$H_0 : \mu_{\text{muller}} = \mu_{\text{home}};$$

$$H_a : \mu_{\text{muller}} \neq \mu_{\text{home}}.$$

Normalmente facemos o contraste para saber se obtemos probas significativas que nos permitan rexeitar a hipótese nula.

Como ben vemos estamos ante un problema de decisión, o que nos leva aos resultados que se amosan no Cadro 1, que poden ocorrer tras aplicar o contraste.

Vemos que, cando non se toma a decisión correcta, aparecen dous tipos de erros, dos que falaremos a continuación. Entón o noso obxectivo básico será minimizar a probabilidade de ambos, aínda que teremos que ter en conta que o aumento dun provoca a diminución do outro. Ante isto, o método que se adoita empregar para solucionar estes problemas será fixar unha probabilidade para o erro de tipo I e tratar de minimizar o de tipo II (ou, como imos ver agora, maximizar a potencia).

Por un lado temos a probabilidade do erro tipo I, tamén denominada nivel de significación, α, e por outro falaremos da potencia, β, que é a probabilidade de detectar que

unha hipótese é falsa, isto é, o complementario da probabilidade do erro tipo II. Os valores habituais que adoita tomar α son 0,01, 0,05, 0,1, que quere dicir que buscaremos rexeitar a hipótese nula a niveis do 1, 5 e 10 %. Polo tanto, temos que:

$$\mathbb{P}\{\text{Erro tipo I}\} = \alpha,$$

$$\mathbb{P}\{\text{Erro tipo II}\} = 1 - \beta.$$

Noutras palabras, o nivel de significación é a probabilidade de rexeitar H_0 cando esta é certa. Chamaremos a estes casos como falsos positivos. Dito nivel vai merecer unha mención especial neste traballo, xa que teremos que ter en conta certos factores, sobre todo, en contrastes múltiples e levar a cabo unhas correccións que nos permitirán controlar as probabilidades do erro tipo I.

Os criterios de contraste, para saber se vou ter probas significativas para rexeitar a hipótese nula, poden ser varios:

Baseados na estatística de contraste e na distribución que este segue na mostraxe: se a estadística, obtida a partir da mostra, é menor ou maior (dependendo do tipo de contraste no que esteamos) que o cuantil da distribución correspondente, é dicir, se este cae na rexión de rexeitamentom teremos probas significativas para rexeitar a hipótese nula; estando este procedemento ligado aos intervalos de confianza.

Tendo en conta o $p-valor$: un $p-valor$ é a probabilidade de que un valor estatístico calculado sexa posible dada unha H_0 certa. Doutro xeito, axudaranos a diferenciar resultados que son produto do azar na mostraxe de resultados que son estatisticamente significativos, é dicir, dos que permiten rexeitar H_0. Así un $p-valor$ é a probabilidade de obter resultados polo menos tan extremos como os resultados que son realmente observados, baixo a suposición de que H_0 é certa. **Criterio:** se obtemos un $p-valor$ máis pequeno que o nivel de significación podemos obter probas para rexeitar H_0 ao nivel α que fixamos.

Concluíndo: rexeitaremos H_0 cando a estatística caia na rexión de rexeitamento ou cando o $p-valor$ sexa menor que o nivel de significación. Noutro caso non teremos probas significativas para rexeitala e "suporémola" como certa.

Aínda que na maioría dos casos empregamos os criterios de contrastes para tomar unha decisión, como ben diciamos ao comezo, para decidir entre un ou outro temos que facer fincapé sobre a decisión que queremos tomar. Con isto queremos dicir que podemos querer contrastar se dúas poboacións teñen a mesma varianza ou a mesma media, incluso poderiamos facer contrastes sobre proporcións ou moitos outros parámetros. Para nós van

resultar de interese os contrastes relacionados coa media (comparación da media cun valor fixo ou coa media doutros grupos).

Debido a que nos centraremos en contrastar medias comezaremos falando un pouco dunha media poboacional, $\mu = \mathbb{E}[X] = \int_\Omega X(\omega)dP(\omega)$. Esta é o valor esperado dunha variable aleatoria. Noutras palabras é a esperanza matemática de dita variable, co que se formaliza a idea do valor medio dun fenómeno aleatorio. Dada unha mostra, o seu estimador será a media mostral, definida como a estatística que se obtén a partir da media aritmética dos valores da mostra $(x_1, \cdots, x_n)$ dunha variable aleatoria X, $\bar{x} = (x_1 + \cdots + x_n)/n$.

Capítulo 1

Contrastes simples para a media

Como vimos na introdución, na inferencia estatística, podemos estar interesados en extraer conclusións sobre un parámetro ou parámetros. Ao longo deste capítulo centrarémonos na media, xa que esta representa unha medida de centralidade dos datos. Comezamos introducindo os contrastes para a media nunha poboación. Despois, posto que este é o problema que queremos analizar, verase como realizar contrastes para comparar a media de dúas ou máis poboacións.

1.1. Contrastes simples para a media nunha poboación

Nesta primeira sección comezamos co caso máis sinxelo, onde quereremos contrastar se a media dunha poboación toma un valor concreto que nós elixamos. Por exemplo, podemos querer saber se un xene concreto se expresa, en media, nunha cantidade maior ou igual a un número dado, xa que se isto ocorrese este sería un xene dominante que lograría expresarse no individuo. Adiantándonos un pouco ao que se desenvolve nas liñas seguintes, para poder levar a cabo o contraste teremos unha mostra dun tamaño n, onde para cada un deses individuos da mostra analizaremos a expresión do xene correspondente.

Temos que comezar por distinguir se estamos ante un contraste de igualdade (bilateral) ou de desigualdade (unilateral) na hipótese nula. Segundo nos atopemos nun ou noutro teremos unha pequena variación cando rexeitemos empregando a estatística de contraste. Vexamos máis a fondo isto considerando inicialmente un contraste unilateral para a media. Nesta sección suporemos que a varianza é descoñecida, xa que se trata do escenario máis habitual na práctica. No caso de que a varianza fose coñecida os procedementos serán similares (ver Observación 1.1).

No caso do contraste simple unilateral para a media formulamos o contraste:

$$H_0 : \mu = \mu_0,$$
$$H_a : \mu \neq \mu_0. \tag{1.1}$$

Consideramos ademais fixado o nivel de significación α. Como a varianza é descoñecida empregaremos, neste contraste, como estimador desta a cuasivarianza mostral, $S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ onde n é o número de observacións que temos e $\bar{x}$ é a media mostral da mostra $(x_1, \cdots, x_n)$.

Agora escribimos a estatística correspondente incluíndo a distribución que segue na mostraxe baixo H_0 en (1.1). Hai que ter en conta que esta distribución sae como consecuencia de que as nosas observacións proveñen dunha distribución normal de media μ e varianza descoñecida.

$$\frac{\bar{x} - \mu_0}{\frac{S_x}{\sqrt{n}}} \sim T_{n-1},$$

onde T_{n-1} é a distribución T−Student con $(n-1)$ graos de liberdade e S_x é a cuasidesviación típica mostral (a raíz da cuasivarianza mostral). O cociente $S_x/\sqrt{n}$ é o que se coñece como erro típico.

Rexeitaremos H_0 no contraste definido en (1.1) se:

$$\frac{|\bar{x} - \mu_0|}{\frac{S_x}{\sqrt{n}}} > t_{n-1,\frac{\alpha}{2}},$$

onde $t_{n-1,\frac{\alpha}{2}}$ é o cuantil dunha T−Student con $(n-1)$ graos de liberdade que deixa unha probabilidade de $\alpha/2$ á dereita e $(1-\alpha/2)$ á esquerda.

A modo de resumo e de forma gráfica, podemos visualizar as rexións de aceptación e de rexeitamento como se ilustra na Figura 1.2. Para o contraste (1.1), rexeitaremos H_0 cando a estatística de contraste caia na rexión de rexeitamento. Se o valor da estatística de contraste cae na rexión de aceptación non existen probas significativas que nos permitan rexeitar a hipótese nula.

Aquí os intervalos de confianza para a media μ a un nivel de confianza $(1-\alpha)$ obtéñense como:

$$\left(\bar{x} - t_{n-1,\frac{\alpha}{2}}\frac{S_x}{\sqrt{n}}, \bar{x} + t_{n-1,\frac{\alpha}{2}}\frac{S_x}{\sqrt{n}}\right).$$

Outro criterio para rexeitar H_0 e que nós empregaremos na maior parte deste traballo, é se o $p-valor$ que obtemos é moi pequeno, máis pequeno que o nivel de significación α que temos fixado. O $p-valor$, para o contraste (1.1), podería ser calculado do seguinte

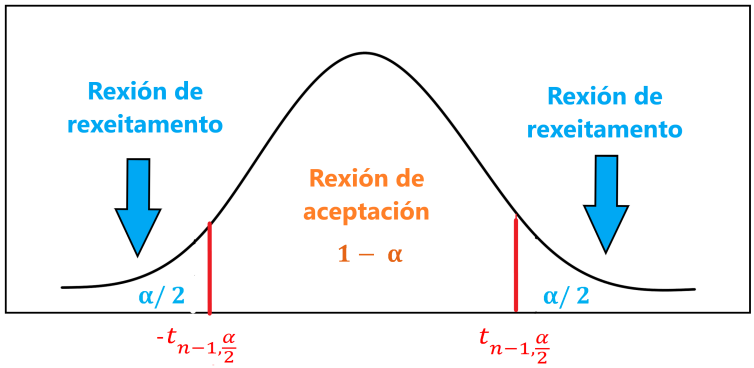

Figura 1.2: Rexións de aceptación e de rexeitamento para o contraste bilateral (1.1).

xeito:

$$\mathbb{P}\left\{|t_{n-1}| > \left|\frac{\bar{x}-\mu_0}{\frac{S_x}{\sqrt{n}}}\right|\right\} = 2\mathbb{P}\left\{t_{n-1} > \left|\frac{\bar{x}-\mu_0}{\frac{S_x}{\sqrt{n}}}\right|\right\} = 2\left(1-\mathbb{P}\left\{t_{n-1} \leq \left|\frac{\bar{x}-\mu_0}{\frac{S_x}{\sqrt{n}}}\right|\right\}\right) = p-valor. \quad (1.2)$$

Se nos pasamos agora ao caso dos contrastes unilaterais, podemos estar interesados en rexeitar a hipótese nula cando o valor da media sexa "moi grande", é dicir,

$$\begin{aligned} H_0 &: \mu \leq \mu_0; \\ H_a &: \mu > \mu_0, \end{aligned} \quad (1.3)$$

ou rexeitar a hipótese nula cando o seu valor sexa "moi pequeno",

$$\begin{aligned} H_0 &: \mu \geq \mu_0; \\ H_a &: \mu < \mu_0. \end{aligned} \quad (1.4)$$

Como seguimos supoñendo que a varianza é descoñecida, continuaremos empregando a cuasivarianza mostral. O procedemento a levar a cabo é moi similar ao que se fai no contraste de igualdade. As pequenas diferenzas que aparecen neste tipo de contrastes, tanto en (1.3) como (1.4), estarán relacionadas co criterio de decisión, o cálculo do $p-valor$ e cos cuantís da distribución T−Student.

Empezando polo contraste en (1.3), o criterio adoptado é rexeitar H_0 se

$$\frac{\bar{x}-\mu_0}{\frac{S_x}{\sqrt{n}}} > t_{n-1,\alpha},$$

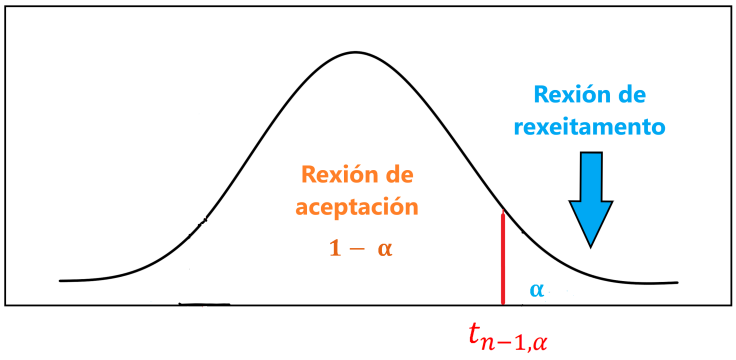

Figura 1.3: Rexións de aceptación e de rexeitamento para o contraste unilateral (1.3).

onde $t_{n-1,\alpha}$ é o cuantil da distribución T−Student de $(n-1)$ graos de liberdade que deixa unha probabilidade α á dereita.

Neste caso, podemos ver unha representación gráfica das rexións de aceptación e de rexeitamento para o contraste (1.3) na Figura 1.3. Rexeitarase a hipótese nula naqueles casos nos que a estatística de contraste caia na rexión de rexeitamento.

Unha maneira de obter os $p-valores$ para o contraste (1.3) pode ser:

$$\mathbb{P}\left\{t_{n-1} < \frac{\bar{x}-\mu_0}{\frac{S_x}{\sqrt{n}}}\right\} = p-valor.$$

Se analizamos agora o contraste (1.4), o criterio de decisión será rexeitar H_0 se

$$\frac{\bar{x}-\mu_0}{\frac{S_x}{\sqrt{n}}} < -t_{n-1,\alpha},$$

onde se emprega o cuantil que mencionabamos para o contraste (1.3) cambiado de signo.

Agora, malia tratarse do mesmo cuantil, se nos fixamos na Figura 1.4, veremos que a rexión de rexeitamento para o contraste (1.4) é aquela que deixa unha probabilidade α á esquerda. Isto é debido á consideración do signo negativo para o cuantil.

Para o contraste (1.4) poderiamos obter os $p-valores$:

$$\mathbb{P}\left\{t_{n-1} > \frac{\bar{x}-\mu_0}{\frac{S_x}{\sqrt{n}}}\right\} = 1-\mathbb{P}\left\{t_{n-1} \leq \frac{\bar{x}-\mu_0}{\frac{S_x}{\sqrt{n}}}\right\} = p-valor.$$

Observación 1.1. No caso en que a varianza σ^2 fose coñecida non necesitariamos estimala, é dicir, deixamos de traballar coa súa cuasivarianza e cuasidesviación típica, xa que a desviación típica tamén se coñece, σ. Como consecuencia disto a distribución na mostraxe que seguirá a estatística de contraste cambiará:

$$\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1),$$

onde $N(0,1)$ é a distribución normal estándar. Os criterios para rexeitar son análogos aos anteriores, intercambiando os cuantís da distribución T de Student de $(n-1)$ graos de liberdade $t_{n-1,\frac{\alpha}{2}}$, $t_{n-1,\alpha}$ e $-t_{n-1,\alpha}$ polos cuantís respectivos da distribución normal estándar $z_{\frac{\alpha}{2}}$, z_α e $-z_\alpha$.

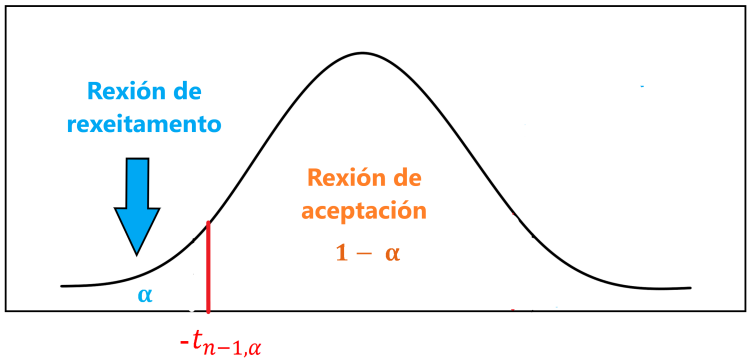

Figura 1.4: Rexións de aceptación e de rexeitamento para o contraste unilateral (1.4).

1.2. Comparación de medias en dúas poboacións

Supoñamos agora que estamos interesados en comparar as medias de dúas poboacións, X e Y. Novamente poderemos atoparnos ante un contraste unilateral ou bilateral. Temos dúas mostras $(x_1, \dots, x_n)$ e $(y_1, \dots, y_m)$ de dúas poboacións normais e independentes, $X \sim N(\mu_x, \sigma_x^2)$ e $Y \sim N(\mu_y, \sigma_y^2)$, onde ambas varianzas son consideradas descoñecidas. Desenvolvemos o caso sen necesidade de asumir que $m = n$, é dicir, o seguinte será válido tanto para cando os tamaños mostrais son iguais coma cando non o son. Presentemos o contraste bilateral:

$$H_0 : \mu_x = \mu_y;$$

$$H_a : \mu_x \neq \mu_y.$$

Como temos mostras de dúas poboacións distintas o que teremos que facer é calcular os estimadores tanto para a media e a varianza da poboación das X como das Y, é dicir, calcularemos as correspondentes medias e cuasivarianzas mostrais.

Así, en función á mostra correspondente a X temos:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Analogamente, para a mostra correspondente a Y, teriamos:

$$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i; S_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2.$$

En virtude das propiedades da varianza tense $\mathrm{Var}(\bar{x} - \bar{y}) = \sigma_x^2/n + \sigma_y^2/m$. Polas propiedades da media, da suposición de independencia (que implica incorrelación) e aplicando o teorema central do límite (ver [5]) chegamos a que, definindo $\bar{z} = \bar{x} - \bar{y}$, a distribución asintótica sería, $\bar{z} \sim N(0, \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}})$ baixo a hipótese nula de igualdade de medias. Con todo isto o que se nos pode ocorrer é contrastar se a diferenza de medias $\mu_z = \mu_x - \mu_y$ é significativamente distinta de cero ou non, é dicir, se as medias de dúas poboacións son significativamente diferentes. Entón o contraste a levar a cabo será:

$$\begin{aligned} H_0 &: \mu_z = 0; \\ H_a &: \mu_z \neq 0. \end{aligned} \tag{1.5}$$

Para poder falar da estatística de contraste, agora temos que distinguir dous casos en función de se as varianzas son iguais ou distintas para as poboacións, considerando, analogamente á Sección 1.1, o suposto de que en ambos casos as varianzas son descoñecidas.

Varianzas descoñecidas pero iguais: $\sigma_x^2 = \sigma_y^2 = \sigma^2$. Neste caso temos entón que a distribución sería $\bar{z} \sim N(0, \sigma\sqrt{\frac{1}{n} + \frac{1}{m}})$. Se estimamos a varianza como unha media ponderada das cuasivarianzas mostrais, obteriamos que a cuasedesviación típica común é $S_z = \sqrt{\frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}$. A partires do anterior, pódese chegar a que a estatística de contraste é:

$$\frac{\bar{z}}{S_z\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim T_{n+m-2}$$

É dicir, a estatística de contraste segue unha distribución T-Student de $(n + m - 2)$ graos de liberdade baixo a hipótese nula.

Rexeitaremos H_0 en (1.5) se:

$$\frac{|\bar{z}|}{S_z\sqrt{\frac{1}{n} + \frac{1}{m}}} > t_{n+m-2,\frac{\alpha}{2}}$$

onde $t_{n+m-2,\frac{\alpha}{2}}$ é o cuantil dunha T−Student con $(n+m-2)$ graos de liberdade que deixa unha probabilidade de $\alpha/2$ á dereita e $(1-\alpha/2)$ á esquerda.

Aquí os intervalos de confianza para a diferenza de medias, a nivel de confianza $(1-\alpha)$ obtéñense como:

$$\left(\bar{z} - t_{n+m-2,\frac{\alpha}{2}} S_z \sqrt{\frac{1}{n}+\frac{1}{m}}, \bar{z} + t_{n+m-2,\frac{\alpha}{2}} S_z \sqrt{\frac{1}{n}+\frac{1}{m}}\right).$$

Como podemos observar, a diferenza do anterior só temos que estimar adecuadamente a varianza e os graos de liberdade para que realmente esteamos ante unha distribución T-Student, polo que, para rexeitar os cuantís da distribución só cambian os graos de liberdade. Vexamos como última cuestión, para este caso de varianzas descoñecidas pero iguais, que unha posible maneira de obter os $p-valores$ sería calculando a seguinte probabilidade:

$$\mathbb{P}\left\{|t_{n+m-2}| > \left|\frac{\bar{z}}{S_z\sqrt{\frac{1}{n}+\frac{1}{m}}}\right|\right\} = 2\left(1 - \mathbb{P}\left\{t_{n+m-2} \leq \left|\frac{\bar{z}}{S_z\sqrt{\frac{1}{n}+\frac{1}{m}}}\right|\right\}\right).$$

Os contrastes unilaterais faríanse de maneira análoga.

Varianzas descoñecidas e distintas: $\sigma_x^2 \neq \sigma_y^2$.

Neste caso a distribución da nosa nova variable sería $\bar{z} \sim N\left(0, \sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}\right)$ baixo H_0. Se temos que estimar as varianzas, a cuasidesviación típica será $S_z = \sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}$, polo que a estatística de contraste para (1.5) agora vai ser (tendo en conta novamente que $\mu_z = 0$ baixo H_0):

$$\frac{\bar{z}}{S_z} \sim T_d,$$

onde T_d é a distribución T-Student con d graos de liberdade, sendo

$$d = \frac{\left(\frac{S_x^2}{n}+\frac{S_y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_x^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_y^2}{m}\right)^2}. \tag{1.6}$$

Rexeitaremos H_0 se:

$$\frac{|\bar{z}|}{S_z} > t_{d,\frac{\alpha}{2}},$$

onde $t_{d,\frac{\alpha}{2}}$ é o cuantil dunha T−Student con d graos de liberdade, véxase (1.6), que deixa unha probabilidade de $\alpha/2$ á dereita e $(1-\alpha/2)$ á esquerda.

Aquí os intervalos de confianza obtéñense como:

$$\left(\bar{z} - t_{d,\frac{\alpha}{2}} S_z, \bar{z} + t_{d,\frac{\alpha}{2}} S_z\right).$$

Por outra parte, unha posibilidade para a obtención dos $p - valores$ neste caso sería:

$$\mathbb{P}\left\{|t_d| > \left|\frac{\bar{z}}{S_z}\right|\right\} = 2\left(1 - \mathbb{P}\left\{t_d \leq \left|\frac{\bar{z}}{S_z}\right|\right\}\right) = p - valor.$$

Observación 1.2. Este último test explicado na Sección 1.2, que permite comparar a media de dúas poboacións, é coñecido como o test t de Welch.

Observación 1.3. Se volvemos ao Exemplo 1.1, esta sería unha das situacións ás que lle poderiamos aplicar o test t de Welch, tendo en conta que os tamaños das dúas poboacións coinciden (en relación á notación anterior sería $m = n = 1000$). Para este caso, $17{,}23 = \bar{x}$ e $21{,}57 = \bar{y}$, sendo X a poboación das mulleres e Y a dos homes.

1.3. Comparación de medias en varias poboacións. ANOVA

O modelo ANOVA é un modelo de análise da varianza que veremos que dará lugar a unha táboa de descomposición da variabilidade. Por outra parte, neste tipo de modelos, hai unha única variable explicativa que ademais será discreta. Esta dará lugar a varios grupos, digamos que da lugar a I grupos. Cada grupo contará con n_i observacións con $i = 1, \ldots, I$ (podemos ter para cada poboación distinto número de observacións). Denotaremos por $Y_{i,j}$ a observación $j-$ésima do grupo i. Ademais, contamos coa suposición de que todas as (sub)poboacións seguirán unha distribución normal $N(\mu_i, \sigma^2)$ onde a varianza é a mesma para todas elas, σ^2, e a media do grupo $i-$ésimo, μ_i, pode cambiar en función do grupo. Como consecuencia disto último teremos que os erros (a diferenza entre o valor real e o valor observado/estimado), ε_i, serán independentes e seguen unha distribución normal $N(0, \sigma^2)$.

En resumo, as hipóteses para poder realizar un test ANOVA son as seguintes: independencia das observacións e dos grupos, homoscedasticidade (quere dicir que as varianzas de todas as subpoboacións son iguais) e normalidade (os datos de cada grupo seguen unha distribución normal estándar). Estas hipóteses pódense comprobar empregando tests estatísticos, que aparecen nas librarías de R, o software estatístico que podemos ver en [15], como poden ser o test de Shapiro Wilks para contrastar normalidade ou o test de Levene para a homoscedasticidade. Se algunha destas hipóteses non se satisfixera, podería procederse a unha transformación dos datos ou a un método non paramétrico como pode ser o denominado Kruskal−Wallis, que poderiamos consultar e velo comparado cun ANOVA no traballo [14].

As estimacións das medias poboacionais para cada grupo serán as medias mostrais locais en cada grupo.

Vexamos agora a descomposición de variabilidade á que da lugar tendo en conta o test de contraste de que tódalas medias son iguais. O contraste será:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_I; \\ H_a &: \exists 1 \leq l, k \leq I : \mu_l \neq \mu_k. \end{aligned} \tag{1.7}$$

Se a hipótese nula fose certa todos os individuos terían como axuste a media global, μ, mentres que baixo a alternativa sería a media local, $\mu_i, i \in \{1, \ldots, I\}$ con algunha poboación que cumpra $\mu_l \neq \mu_k$.

Se denotamos por ε o vector que contén os erros de todas as poboacións e por $\hat{\varepsilon}$ o vector que contén os residuos (cada residuo é o valor observado menos a estimación), vemos que baixo as distintas hipóteses teremos distintas sumas residuais de cadrados, RSS, que permiten medir a cantidade de varianza nun conxunto de datos.

Baixo a hipótese nula e baixo a alternativa teremos, respectivamente:

$$RSS_0 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{i,j} - \bar{Y})^2,$$

$$RSS = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{i,j} - \bar{Y}_i)^2.$$

Teremos a seguinte descomposición da variabilidade:

Variabilidade total (VT) =Variabilidade explicada(VE)+ Variabilidade non explicada (VNE).

$RSS_0 = \sum_{i,j=1}^{I,n_i} (\bar{Y}_i - \bar{Y})^2 + RSS.$

A estatística de contraste para (1.7) seguirá a seguinte distribución na mostraxe baixo a hipótese nula:

$$\frac{\frac{RSS_0 - RSS}{I-1}}{\frac{RSS}{n-I}} \sim F_{I-1,n-I},$$

onde $F_{I-1,n-I}$ é unha distribución F de Snedécor con $(I-1, n-I)$ graos de liberdade, xa que estamos facendo o cociente de varianzas. Valores altos desta estatística indicarán que a hipótese nula non sería certa.

Rexeitaremos a hipótese nula cando a estatística sexa grande ou, novamente, cando teñamos un $p-valor$ suficientemente pequeno. Se o quixésemos, este $p-valor$ poderíase

obter do seguinte xeito:

$$\mathbb{P}\left\{|F_{I-1,n-I}| > \left|\frac{\frac{RSS_0-RSS}{I-1}}{\frac{RSS}{n-I}}\right|\right\} = 2\left(1-\mathbb{P}\left\{F_{I-1,n-I} \leq \left|\frac{\frac{RSS_0-RSS}{I-1}}{\frac{RSS}{n-I}}\right|\right\}\right) = p-valor.$$

$$\mathbb{P}\left\{F_{I-1,n-I} < \frac{\frac{RSS_0-RSS}{I-1}}{\frac{RSS}{n-I}}\right\} = p-valor.$$

No caso en que puidésemos rexeitar a igualdade de tódalas medias gustaríanos poder estudar que variables aleatorias presentan unha media diferente.

O problema co que nos atopamos é que se facemos estes contrastes simultaneamente teremos que non se respectará o nivel de significación, polo que, necesitaremos facer algunha corrección, que será no que nos centraremos ao longo dos seguintes capítulos.

Capítulo 2

Contrastes múltiples

Como ben rematabamos introducindo, unha vez que rexeitamos a igualdade de todas as medias, podemos querer contrastar, para cada par de variables aleatorias, cales das medias se poden asumir iguais simultaneamente. Por dicilo dalgunha maneira, sería como facer contrastes simples de igualdade de medias para todas as posibles combinacións de grupos distintos.

Un problema co que nos poderiamos atopar é o seguinte: supoñamos que temos $k = 10$ comparacións e un nivel de significación fixado $\alpha = 0{,}05$. Nun contraste simple este nivel indica que se aplicásemos o test en 100 mostras distintas sabendo que a hipótese nula é certa, deberiamos rexeitar 5 veces esta hipótese, é dicir, habería 5 falsos positivos. Se agora asumimos que os k contrastes son independentes, a probabilidade do erro de tipo I pasaría a ser, tendo en conta que a independencia proporciona que a intersección de probabilidades sexa igual ao produto, a seguinte probabilidade de cometer un erro tipo I:

$\mathbb{P}\{\text{Erro tipo I}\} = \mathbb{P}\{\text{Rexeitar algún}|H_0 \text{ certa}\} = 1 - \mathbb{P}\{\text{Non rexeitar ningún}|H_0 \text{ certa}\} = 1 - (1 - \mathbb{P}\{\text{Rexeitar } H_0|H_0 \text{ certa}\})^k = 1 - (1-\alpha)^k$.

Volvendo ao exemplo anterior, $\mathbb{P}\{\text{Erro tipo I}\} = (1-(1-\alpha)^k) = (1-(1-0{,}05)^{10}) = 0{,}4$, polo que no caso particular dun contraste con $k = 10$, $\alpha = 0{,}05$ estariamos rexeitando o 40 % das veces a hipótese nula, supoñendo que esta é certa, cando esperariamos rexeitar polo menos unha un 5 % das veces.

Chegamos así a que, normalmente, teremos que corrixir este efecto dalgunha forma. Unha excepción ao caso anterior, onde non hai que considerar ningunha corrección é se podemos asumir que os k contrastes son completamente dependentes.

Para as posibles correccións debemos recordar o que era o nivel de significación, α (probabilidade de rexeitar H_0 cando esta é certa, isto é, probabilidade dun falso positivo) e a potencia, β (probabilidade de detectar cando unha hipótese sexa falsa). Ademais disto, deberemos tamén introducir os termos do False Discovery Rate (FDR) e o Family

Wise Error Rate (FWER) que serán chave á hora de introducir os distintos métodos que empreguemos para solucionar o problema que se mencionou anteriormente.

O **FDR** é a proporción esperada de falsos positivos de entre todos os tests considerados como significativos, noutras palabras, é a proporción esperada de erros tipo I. Traballando con isto, teremos fixado o α e estableceremos un límite de significación para que, de entre todos os tests significativos, a proporcións de H_0 certas non supere dito valor.

O FDR tamén se define como a proporción esperada de rexeitamentos falsos, é dicir de rexeitamentos de hipóteses nulas cando estas serían certas, entre todas as hipóteses nulas rexeitadas. Podemos escribilo como,

$$FDR = \mathbb{E}\left(\frac{v}{R}\right)\mathbb{P}(R > 0),$$

onde $\mathbb{E}$ é a esperanza matemática, R é o número de hipóteses rexeitadas e v é o número de hipóteses rexeitadas cando estas eran certas. Se $R = 0$, entón $FDR = 0$.

Por outro lado, o **FWER** é a probabilidade de obter ao menos un falso positivo, é dicir, que ao menos rexeitamos unha das hipóteses nulas cando esta é certa en tódalas comparacións. Pensándoo, podemos ver que queremos asegurar que o risco dun falso positivo sexa menor ou igual que α. Isto vai ser significativamente máis forte que o FDR, polo que se podemos facer un control do FWER poderemos facer un do FDR, pero á inversa non sempre.

Unha definición máis formal, seguindo o Capítulo 1 en [3], sería que o FWER é a máxima probabilidade de rexeitar H_0 baixo todas as posibles combinacións de hipóteses nulas. O supremo destas probabilidades vai ser o que controlemos e lle asignemos

$$FWER = \sup_{H_0} \mathbb{P}\{\text{rexeitar } H_0\} \leq \alpha.$$

Os procedementos de control do FDR poden concluír en máis erros de tipo I e menos de tipo II que os procedementos de control do $FWER$. En xeral, $FWER \leq FDR$, de aí que o control do FDR sexa significativamente máis potente. No caso en que todas as hipóteses nulas sexan certas teremos que $FWER = FDR$.

Proposición 2.1. *$FWER \leq FDR$.*

Demostración. ▪ Se todas as hipóteses nulas son certas.

$$FDR = \mathbb{E}\left(\frac{v}{R}\right)\mathbb{P}(R > 0).$$

Baixo o suposto de que todas as hipóteses nulas son certas, temos que $\frac{v}{R} = 0$ se $v = 0$ e $\frac{v}{R} = 1$ se $v \geq 1$.

Entón

$$FDR = 0 \cdot \mathbb{P}(v = 0) + 1 \cdot \mathbb{P}(v \geq 1) = \mathbb{P}(v \geq 1) = FWER.$$

- Se non todas as hipóteses nulas son certas:

  Neste caso temos que $v < R \leftrightarrow \frac{v}{R} < 1$, entón

  $FDR = \frac{v}{R} \cdot \mathbb{P}(v \geq 1) + 0 \cdot \mathbb{P}(v = 0) = \frac{v}{R} \cdot \mathbb{P}(v \geq 1) < \mathbb{P}(v \geq 1) = FWER.$

□

Capítulo 3

Métodos de correccións

Neste capítulo veremos distintos métodos que permiten o control do FWER, como son a corrección de Bonferroni na Sección 3.1, a de Holm na 3.2, a de Hochberg na 3.3 ou a de Hommel na 3.4, e outros para o control do FDR, como son as correccións de Benjamini−Hochberg na Sección 3.5, Benjamini−Yekutieli na 3.6 ou a de Benjamini−Liu na 3.7. Todos estes métodos, agás o de Benjamini−Liu, atópanse implementados nas librarías `stats`, do software estatístico R. Quero mencionar que todos os métodos que aparecen, agás Bonferroni e Hommel, están baseados na ordenación dos $p-valores$ obtidos tras un contraste simple. Distinguiremos entre os chamados métodos paso a paso cara arriba ("step−up") coma o de Hochberg na Sección 3.3, no que os $p-valores$ se ordenan dende o menos ao máis significativo (equivalentemente de maior a menor valor numérico), e os métodos paso a paso cara abaixo ("step−down methods"), as restantes Seccións 3.2, 3.4, 3.5, 3.6 e 3.7, e que toman os $p-valores$ ordenados dende o máis significativo ao menos; é dicir, de menor a maior. Trala ordenación dos $p-valores$ os criterios de decisión varían en función do método adoptado.

3.1. Método de Bonferroni

O método de Bonferroni, que se pode atopar no Capítulo 1 do libro [3], introduce unha corrección no nivel de significación que consiste en dividir o α orixinal entre o número de comparacións que teñamos, dando lugar a un nivel de significación axustado. Isto escríbese como $\alpha_k = \alpha/k$ sendo k o número de comparacións que teñamos.

Este α axustado lévanos ao seguinte criterio: rexeitaremos H_0 cando $p_i < \alpha/k$, ou equivalentemente $k \cdot p_i < \alpha$ con $i = 1, \ldots, k$, tendo en conta que p_i é o $p-valor$ para o contraste $i-$ésimo.

Un dos problemas que ten o método proposto por Bonferroni é que produce un au-

mento nos erros de tipo II. Algunhas das alternativas que se basean na idea orixinal de Bonferroni e tratan de mellorar a potencia dos tests presentan o problema de que deixan de respectar o nivel de significación (véxase [2]). Outra limitación que tamén se lle atribúe a este método é que non podemos traballar cun número demasiado grande de contrastes porque se considerásemos un k suficientemente grande o α_k sería moi pequeno, chegando así a que ningún dos tests sería significativo.

Imos ver agora que o método de Bonferroni controla o FWER a través da desigualdade de Boole empregando álxebra de sucesos.

Proposición 3.1. *A desigualdade de Boole establece que para toda familia finita de sucesos, a probabilidade de que ao menos un deses sucesos ocorra é menor ou igual que a suma dos sucesos individuais:*

Temos a seguinte familia de sucesos (finita) $A_1, \dots, A_n$ entón

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) \leq \sum_n \mathbb{P}(A_n).$$

Demostración. Este resultado pode obterse por indución en n:

Para $n = 1$ é certo que $\mathbb{P}(A_1) \leq \mathbb{P}(A_1)$.

Consideremos agora que o resultado se cumpre para $n-1$ sucesos, vexamos que é certo para n tamén. Denotemos por $E = \bigcup_{i=1}^{n-1} A_i$. Por hipótese de indución sabemos $\mathbb{P}(E) \leq \sum_{i=1}^{n-1} A_i$.

Entón, $\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \mathbb{P}(E \cup A_n) = \mathbb{P}(E) + \mathbb{P}(A_n) - \mathbb{P}(E \cap A_n)$.

Tendo en conta que a probabilidade toma valores entre 0 e 1 e a hipótese de indución que diciamos antes

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n-1} A_i + A_n = \sum_{i=1}^{n} A_i.$$

□

Con isto, se volvemos ao método de Bonferroni con k contrastes e un nivel de significación axustado α/k, podemos ver que a probabilidade de que os falsos positivos, que denotaremos por η, sexan cando menos un, $\mathbb{P}(\eta \geq 1)$, sería:

$$\mathbb{P}(\eta \geq 1) \leq \sum_{i=1}^{k} \mathbb{P}_{H_{0,i}}(\text{Rexeitar } H_{0,i}) = \sum_{i=1}^{k} \frac{\alpha}{k} = \alpha.$$

Concluímos así que o método de Bonferroni respecta a probabilidade do erro tipo I, α, que tiñamos, é dicir, controla o FWER sen supoñer ningunha hipótese a maiores.

3.2. Método de Holm

Como ben diciamos comezamos a falar agora dos métodos paso a paso cara abaixo. Consideramos k contrastes de hipóteses e o nivel de significación fixado, α.

Empezamos ordenando os $p-valores$, tendo en conta que comezabamos co máis significativo, é dicir, $p_{(1)} \leq \cdots \leq p_{(k)}$, onde $p_{(i)}$ non fai referencia ao contraste $i-$ésimo senón ao que ocupa o lugar i unha vez ordenados.

Como criterio, que se pode atopar en [1], o método de Holm rexeita todas as hipóteses nulas $H_{0,(j)}$ con $j = 1, \ldots, i$, sendo i o maior índice que cumpre $p_{(i)} < \alpha/(k-i+1)$ con $i = 1, \ldots, k$.

Observación 3.2. Rexeitamos todas as $H_{0,(j)}$, é dicir, as hipóteses que están asociadas a $p_{(j)}$ que non ten porque corresponderse co contraste $j-$ésimo, $H_{0,j}$. Por exemplo: Teñamos $k = 10$, $\alpha = 0{,}05$ e $p_7 = 0{,}002$ cumprindo que $p_7 < p_m$ con $m \neq 7$. Neste caso $p_7 = p_{(1)} \leq \frac{0{,}05}{10-1+1}$ cumprindo o criterio de Holm. Polo que poderiamos rexeitar o contraste $H_{0,(1)}$ que se correspondería co contraste $H_{0,7}$.

Outro criterio para rexeitar será obter os $p-valores$ axustados, $p_{(i)} \cdot (k-i+1)$, que se obteñen da desigualdade anterior e comparalo co α de partida, é dicir, rexeitaranse todas as hipóteses nulas $H_{0,(j)}$, con $j = 1, \cdots, i$; sendo i o maior índice verificando que $p_{(i)} \cdot (k-i+1) \leq \alpha$.

O método de Holm basease na desigualdade de Bonferroni e é válido independentemente da distribución conxunta das estatísticas que se empregan.

3.3. Método de Hochberg

Só para facernos unha pequena idea visual de como funcionarían os métodos step−up introducimos aquí o esquema seguido para realizar o contraste múltiple segundo as ideas de Hochberg. Hochberg é un método máis potente que Holm, véxase, por exemplo, [9]. Unha suposición adicional que debe cumprirse é a independencia ou a dependencia positiva dos $p-valores$, como consecuencia de que o método de Hochberg se basea na proba de Simes, que se verá na Proposición 3.4. Comezamos ordenando os $p-valores$: $p_{\{1\}} \geq \cdots \geq p_{\{k\}}$, sendo k o número de contrastes e α o nivel de significación fixado. $p_{\{i\}}$ fai referencia á hipótese nula $H_{0,\{i\}}$.

O criterio tomado de [9] será rexeitar todas as hipóteses nulas a partir dun certo índice. Decidiremos rexeitar todas as $H_{0,\{j\}}$ con $j \geq i$, sendo i o menor índice que verifica $p_{\{i\}} < \alpha/i$ con $i = 1, \ldots, k$.

A outra opción, novamente, será rexeitar aquelas hipóteses nulas que cumpran $ip_{\{i\}} < \alpha$.

Observación 3.3. Introducimos a notación $p_{\{i\}}$ no lugar de $p_{(i)}$ para diferenciar a ordenación descendente dos $p-valores$ da ascendente.

3.4. Método de Hommel

Un dos últimos métodos de correccións para o control do FWER que decidimos incluír é o de Hommel. Para a súa descrición decidimos seguir [13].

O método de Hommel, ao igual que o de Hochberg, funciona para contrastes de hipóteses independentes ou cando os $p-valores$ presentan unha dependencia positiva. As dúas diferenzas que podemos atopar entre eles é, unha á hora de computalos, debido a que os $p-valores$ de Hochberg son máis rápidos de computar que os de Hommel, e outra é que o método de Hommel é máis potente que o de Hochberg.

Recordando que k é o número de contrastes que estamos a levar a cabo e que os $p-valores$ os ordenaremos de forma ascendente, o criterio do método de Hommel é buscar o enteiro j que cumpra: $j = \text{máx}\{i \in \{1, \cdots, k\} : p_{(k-i+l)} > l\alpha/i, l = 1, \cdots, i\}$. Se non existe o máximo entón rexeitaremos tódalas hipóteses nulas. En caso contrario, rexeitaremos aquelas hipóteses nulas tales que $p_{(i)} \leq \alpha/j$. Novamente, outro criterio de decisión é calcular os $p-valores$ axustados, é dicir $jp_{(k-j+i)}/i$, e rexeitar aquelas hipóteses nulas cuxos $p-valores$ axustados non superen α.

Novamente observamos que os resultados obtidos dependerán da ordenación das hipóteses nulas, $H_{0,i}$.

3.5. Método de Benjamini−Hochberg

Pasemos aos métodos que controlarán o FDR, empezando polo de Benjamini−Hochberg que se pode localizar no Capítulo 1 do libro [3]. Tendo novamente un nivel de significación fixado, α, e k contrastes, o método de Benjamini−Hochberg funcionaría como:

Ordenamos os $p-valores$ de menor a maior, por exemplo $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k)}$. O criterio será rexeitar aquelas hipóteses nulas $H_{0,(j)}$ con $j = 1, \cdots, i$ onde $i \in \{1, \cdots, k\}$ é o maior índice cuxo $p-valor$ asociado verifica $p_{(i)} \leq i\alpha/k$. Hai que ter en conta que, ao igual que antes, $p_{(i)}$ é o $p-valor$ asociado ao contraste cuxo $p-valor$ ocupa a posición $i-$ésima unha vez ordenados (ver Observación 3.2).

Unha das hipóteses que necesitaremos verificar para poder aplicar este método será que todos os $p-valores$ sexan independentes ou positivamente dependentes, condición que se

establece en [7].

Para comprobar que o método funciona no caso de independencia emprégase a desigualdade de Simes, cuxa demostración podemos atopar en [17].

Proposición 3.4. *A desigualdade de Simes enúnciase como: Sexan* $p_{(1)}, \cdots, p_{(k)}$ *os* $p-$*valores ordenados de maior a menor significación. Se asumimos que* $p_{(1)}, \cdots, p_{(k)}$ *son variables aleatorias independentes que seguen unha distribución* $N(0,1)$ *e sexa* $A_n(\alpha) = \mathbb{P}\left\{p_{(j)} > \frac{j\alpha}{n}; j = 1, \cdots, n\right\}$ *e* $0 \leq \alpha \leq 1$. *Entón* $A_n(\alpha) = 1 - \alpha$.

3.6. Método de Benjamini−Yekutieli

Cando nos atopamos que entre os $p-$*valores* existe dependencia positiva podemos, tamén, recorrer ao método de Benjamini−Yekutieli, véxase [4].

Seguimos no caso en que os $p-$*valores* se atopan ordenados como antes $p_{(1)} \leq \cdots \leq p_{(k)}$. Agora o criterio de decisións diranos que rexeitaremos todas as hipóteses nulas $H_{0,(j)}$ con $j = 1, \cdots, i$, onde $i \in \{1, \cdots, k\}$ é o maior índice cuxo $p-$*valor* verifique:

$$p_{(i)} \leq \frac{i}{k \cdot c(k)}\alpha,$$

onde, na práctica, a proposta de Benjamini−Yekutieli é empregar $c(k) = \sum_{j=1}^{k} \frac{1}{j}$, coñecido como o número harmónico.

Dito número harmónico pode ser aproximado para valores k suficientemente grandes, tendo en conta o desenvolvemento dunha serie de Taylor e empregando a constante de Euler, $\gamma \approx 0{,}57721$,

$$c(k) = \sum_{j=1}^{k} \frac{1}{j} \approx \ln(k) + \gamma + \frac{1}{2k}.$$

O método de Benjamini−Hochberg pode escribirse como un caso particular do de Benjamini−Yekutieli, tomando $c(k) = 1$.

3.7. Método de Benjamini−Liu

Introducimos este novo método que tamén se pode consultar en [6], para situarnos nun entorno no que se controla o FDR e non necesitaremos saber de antemán nin contrastar as hipóteses de dependencia positiva ou independencia dos $p-$*valores*. Isto resulta importante xa que contrastar e verificar ditas hipóteses é complexo, escapándosenos a súa análise do obxectivo deste TFG.

O funcionamento deste método é unha pequena modificación do de Benjamini-Hochberg. Comezamos ordenando os $p-valores$ dende o máis significativo ao menos, é dicir, partimos de que $p_{(1)} \leq \cdots \leq p_{(k)}$ onde k representa o número de tests estatísticos que estamos levando a cabo. Comezando dende o $p-valor$ máis pequeno, rexeitaremos todas as $H_{0,(j)}$ $j = 1, \cdots, i$ onde i é o maior índice que verifica $p_{(i)} \leq h_{(i)}$, sendo $h_{(i)} = \text{mín}\left(0{,}05, \frac{0{,}05 \cdot k}{(k+1-i)^2}\right)$.

3.8. Aclaracións mediante un exemplo

Incluímos unha última sección neste capítulo na que consideraremos un conxunto de $p-valores$ concretos, para os cales poderemos observar as diferenzas entre os métodos step−up e step−down, vendo que non rexeitan sempre as mesmas hipóteses.

Realizamos agora $k = 10$ contrastes de hipóteses e consideramos o nivel de significación habitual $\alpha = 0{,}05$. É dicir, temos 10 hipóteses nulas ás que lle corresponden os seguintes $p-valores$: $p_1 = 0{,}037$, $p_2 = 0{,}024$, $p_3 = 0{,}048$, $p_4 = 0{,}0021$, $p_5 = 0{,}0067$, $p_6 = 0{,}058$, $p_7 = 0{,}00221$, $p_8 = 0{,}0093$, $p_9 = 0{,}074$ e $p_{10} = 0{,}04$.

Inicialmente, sen ningún tipo de corrección, rexeitaríamos aquelas hipóteses nulas cuxo $p-valor \leq 0{,}05$, é dicir, para esta serie de valores concretos rexeitamos todas as hipóteses nulas agás as correspondentes a p_6 e p_9, $H_{0,6}$ e $H_{0,9}$.

Se agora decidimos empregar a corrección dada polo método de Bonferroni, rexeitaríamos aquelas hipóteses nulas cuxos $p-valores$ non superasen o cociente $0{,}05/10 = 0{,}005$. Usando os $p-valores$ anteriores, temos que o método de Bonferroni rexeita unicamente as hipóteses nulas $H_{0,4}$ e $H_{0,7}$.

Agora para os métodos step−down necesitamos que os $p-valores$ se mostren ordenados de menor a maior, polo que teriamos: $p_{(1)} = 0{,}0021$, $p_{(2)} = 0{,}00221$, $p_{(3)} = 0{,}0067$, $p_{(4)} = 0{,}0093$, $p_{(5)} = 0{,}024$, $p_{(6)} = 0{,}037$, $p_{(7)} = 0{,}04$, $p_{(8)} = 0{,}048$, $p_{(9)} = 0{,}058$ e $p_{(10)} = 0{,}074$. Tendo en conta a Observación 3.2 debemos recordar que $p_{(i)}$ fai referencia a hipótese nula $H_{0,(i)}$ e non $H_{0,i}$ con $i = 1, \cdots, 10$. Tendo en conta isto, as hipóteses nulas correspondentes á ordenación dos $p-valores$ anteriores serían, respectivamente: $H_{0,(1)} = H_{0,4}$, $H_{0,(2)} = H_{0,7}$, $H_{0,(3)} = H_{0,5}$, $H_{0,(4)} = H_{0,8}$, $H_{0,(5)} = H_{0,2}$, $H_{0,(6)} = H_{0,1}$, $H_{0,(7)} = H_{0,10}$, $H_{0,(8)} = H_{0,3}$, $H_{0,(9)} = H_{0,6}$ e $H_{0,(10)} = H_{0,9}$.

Para aplicar a corrección dada polo método de Holm sabemos que, en virtude da Sección 3.2, rexeitaremos todas as hipóteses nulas $H_{0,(j)}$ con $j = 1, \cdots, i$; sendo i é o maior índice tal que $p_{(i)} < 0{,}05/(10 - i + 1)$, onde $1 \leq i \leq 10$. Estes últimos aparecen calculados no Cadro 3.2, comparando estes valores cos $p-valores$ obtidos neste exemplo, obtemos os resultados que se amosan no Cadro 3.1. Isto é, rexeitamos as hipóteses nulas $H_{0,(1)} = H_{0,4}$ e $H_{0,(2)} = H_{0,7}$, xa que $p_{(3)} = 0{,}0067 > 0{,}05/8 = 0{,}0063$ e $p_{(1)} = 0{,}0021 < 0{,}05/10 = 0{,}005$,

$p_{(2)} = 0{,}00221 < 0{,}05/8 = 0{,}0093$.

Se continuamos aplicando o método de Hommel explicado na Sección 3.4, buscaremos o maior enteiro $j = \text{máx}\{i \in \{1, \cdots, 10\} : p_{(10-i+l)} > 0{,}05l/i, n = 1, \cdots, i\}$ e rexeitaremos as hipóteses nulas $H_{0,(i)}$ se $p_{(i)} \leq 0{,}05/j$ se existe o máximo, xa que noutro caso serán rexeitadas todas as $H_{0,(i)}$ con $i = 1, \cdots, 10$. Observando o Cadro 3.3, onde incluímos os cálculos $\{i \in \{1, \cdots, 10\} : 0{,}05l/i, l = 1, \cdots, i\}$, e comparándoos cos $p_{(10-i+l)}$ correspondente chegamos a que $j = 7$. Entón rexeitamos as hipóteses nulas $H_{0,(1)} = H_{0,4}$, $H_{0,(2)} = H_{0,7}$ e $H_{0,(3)} = H_{0,5}$ xa que $p_{(1)}$, $p_{(2)}$ e $p_{(3)}$ son menores que $0{,}05/7 = 0{,}0071$.

Continuando por orde pasamos agora ao método de Benjamini−Hochberg. O criterio de decisión que atopamos na Sección 3.5 permite rexeitar as hipóteses nulas $H_{0,(j)}$ con $j = 1, \ldots, i$; onde $i \in \{1, \cdots, 10\}$ é o maior índice para o cal $p_{(i)} \leq 0{,}05i/10$. Comparando cada $p_{(i)}$ para os métodos step−down, que podemos atopar no Cadro 3.1, co resultado correspondente da operación $0{,}05i/10$ que atopamos no Cadro 3.2, concluímos que o método de Benjamini−Hochberg rexeita as hipóteses nulas $H_{0,4}$, $H_{0,7}$, $H_{0,5}$, $H_{0,8}$ e $H_{0,2}$.

Pasemos agora a aplicar a corrección dada polo método de Benjamini−Yekutieli na Sección 3.6. Aquí, similar ao anterior, buscamos o maior índice $i \in \{1, \cdots, 10\}$ para o cal $p_{(i)} \leq 0{,}05i/(10 \cdot c(10))$, onde $c(10)$ era o número harmónico que se obtiña como

$$c(10) = \sum_{j=1}^{10} \frac{1}{j} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{10} \approx 2{,}93,$$

e rexeitaremos as hipóteses nulas $H_{0,(j)}$ con $j = 1, \cdots, i$. Tendo en conta agora a ordenación dos $p - valores$ para os métodos step−down que podemos atopar no Cadro 3.1 e comparándoos cos resultados de $0{,}05i/(10 \cdot 2{,}93)$ que podemos observar no Cadro 3.2, chegamos á conclusión de que só se poden rexeitar as hipóteses nulas $H_{0,4}$ e $H_{0,7}$.

Como último método step−down tiñamos o de Benjamini−Liu, para o cal, seguindo o criterio explicado na Sección 3.6, rexeitaremos aquelas hipóteses nulas $H_{0,(j)}$ para $j = 1, \cdots, i$ onde $i \in \{1, \cdots, 10\}$ é o maior índice para o que se verifica $p_{(i)} \leq h_{(i)}$, onde $h_{(i)} = \text{mín}\left(0{,}05, \frac{0{,}05 \cdot 10}{(10+1-i)^2}\right)$. Novamente, comparando os diferentes valores de $h_{(i)}$ que podemos atopar no Cadro 3.2 cos correspondentes $p - valores$ para un método step−down, que se observan no Cadro 3.1, chegamos finalmente a que se rexeitan as hipóteses nulas $H_{0,4}$, $H_{0,7}$, $H_{0,5}$, $H_{0,8}$ e $H_{0,3}$.

Vendo por último o método step−up de Hochberg, necesitamos que os $p - valores$ se mostren ordenados de maior a menor, polo que teriamos, recordando a notación que empregamos na Sección 3.5 para diferenciar a ordenación dos $p - valores$ deste método dos step−down: $p_{\{1\}} = 0{,}074$, $p_{\{2\}} = 0{,}058$, $p_{\{3\}} = 0{,}048$, $p_{\{4\}} = 0{,}04$, $p_{\{5\}} = 0{,}037$, $p_{\{6\}} = 0{,}024$, $p_{\{7\}} = 0{,}0093$, $p_{\{8\}} = 0{,}0067$, $p_{\{9\}} = 0{,}00221$ e $p_{\{10\}} = 0{,}0021$. Novamente,

tendo en conta a Observación 3.2, as hipóteses nulas correspondentes á ordenación dos $p-valores$ anteriores serían, respectivamente: $H_{0,\{1\}} = H_{0,9}$, $H_{0,\{2\}} = H_{0,6}$, $H_{0,\{3\}} = H_{0,3}$, $H_{0,\{4\}} = H_{0,10}$, $H_{0,\{5\}} = H_{0,1}$, $H_{0,\{6\}} = H_{0,2}$, $H_{0,\{7\}} = H_{0,8}$, $H_{0,\{8\}} = H_{0,5}$, $H_{0,\{9\}} = H_{0,7}$ e $H_{0,\{10\}} = H_{0,4}$.

Polo criterio estudado na Sección 3.3, rexeitaremos aquelas hipóteses nulas $H_{0,\{j\}}$ con $j \geq i$, onde $i \in \{1, \cdots, 10\}$ é o menor índice que verifica $p_{\{i\}} < 0{,}05/i$. Observando a fila de Hochberg no Cadro 3.2 e comparando a fila i co $p_{\{i\}}$ correspondente aos métodos step−up, chegamos a que o menor índice i para o que se verifica a desigualdade é $i = 9$. Polo tanto rexeitamos as hipóteses nulas $H_{0,\{9\}} = H_{0,7}$ e $H_{0,\{10\}} = H_{0,4}$.

Finalmente, a modo de esquema e resumo incluímos o Cadro 3.1 no que se destacan en grosa as hipóteses que se rexeitan.

i	1	2	3	4	5	6	7	8	9	10	Criterio
Sen ordenar											
$H_{0,i}$	$H_{0,1}$	$H_{0,2}$	$H_{0,3}$	$H_{0,4}$	$H_{0,5}$	$H_{0,6}$	$H_{0,7}$	$H_{0,8}$	$H_{0,9}$	$H_{0,10}$	
p_i	0,037	0,024	0,048	0,0021	0,0067	0,058	0,00221	0,0093	0,074	0,04	
Sen corrección	**0,037**	**0,024**	**0,048**	**0,0021**	**0,0067**	0,058	**0,00221**	**0,0093**	0,074	**0,04**	Rexeito $H_{0,i}$ se $p_i \leq 0{,}05$
Bonferroni	0,037	0,024	0,048	**0,0021**	0,0067	0,058	**0,00221**	0,0093	0,074	0,04	Rexeito $H_{0,i}$ se $p_i \leq \frac{0{,}05}{10} = 0{,}005$
Step-down											
$H_{0,(i)}$	$H_{0,4}$	$H_{0,7}$	$H_{0,5}$	$H_{0,8}$	$H_{0,2}$	$H_{0,1}$	$H_{0,10}$	$H_{0,3}$	$H_{0,6}$	$H_{0,9}$	
$p_{(i)}$	0,0021	0,00221	0,0067	0,0093	0,024	0,037	0,04	0,048	0,058	0,074	
Holm	**0,0021**	**0,00221**	0,0067	0,0093	0,024	0,037	0,04	0,048	0,058	0,074	Rexeito $H_{0,(j)}$, $j = 1, \cdots, i$, sendo i o maior índice: $p_{(i)} < \frac{0{,}05}{11-i}$
Hommel	**0,0021**	**0,00221**	**0,0067**	0,0093	0,024	0,037	0,04	0,048	0,058	0,074	Rexeito $H_{0,(i)}$ se $p_{(i)} \leq \frac{0{,}05}{j}$, $j = \max\{i \in \{1, \cdots, 10\} : p_{(10-i+l)} > \frac{0{,}05l}{i}, l = 1, \cdots, i\}$
Benjamini-Hochberg	**0,0021**	**0,00221**	**0,0067**	**0,0093**	**0,024**	0,037	0,04	0,048	0,058	0,074	Rexeito $H_{0,(j)}$, $j = 1, \cdots, i$, sendo i o maior índice: $p_{(i)} \leq \frac{0{,}05i}{10}$
Benjamini-Yekutieli	**0,0021**	**0,00221**	0,0067	0,0093	0,024	0,037	0,04	0,048	0,058	0,074	Rexeito $H_{0,(j)}$, $j = 1, \cdots, i$, sendo i o maior índice: $p_{(i)} \leq \frac{0{,}05i}{10 \cdot 2{,}93}$ $c(k) \approx 2{,}93$
Benjamini-Liu	**0,0021**	**0,00221**	**0,0067**	**0,0093**	0,024	0,037	0,04	**0,048**	0,058	0,074	Rexeito $H_{0,(j)}$, $j = 1, \cdots, i$, sendo i o maior índice: $p_{(i)} \leq h_i$, $h_i = \min\{0{,}05, \frac{0{,}05 \cdot 10}{(11-i)^2}\}$
Step-up											
$H_{0,\{i\}}$	$H_{0,9}$	$H_{0,6}$	$H_{0,3}$	$H_{0,10}$	$H_{0,1}$	$H_{0,2}$	$H_{0,8}$	$H_{0,5}$	$H_{0,7}$	$H_{0,4}$	
$p_{\{i\}}$	0,074	0,058	0,048	0,04	0,037	0,024	0,0093	0,0067	0,00221	0,0021	
Hochberg	0,074	0,058	0,048	0,04	0,037	0,024	0,0093	0,0067	**0,00221**	**0,0021**	Rexeito $H_{0,\{j\}}$, $j \geq i$, sendo i o menor índice: $p_{\{i\}} < \frac{0{,}05}{i}$

Cadro 3.1: Exemplo de como rexeitan os distintos métodos para un conxunto concreto de 10 $p - valores$.

Para finalizar esta sección, a modo de resumo podemos distinguir que hipóteses nulas se rexeitan en función do escenario que escollamos:

1. Decidimos levar a cabo un control do FWER: como ben estudamos no Capítulo 3, poderemos empregar as correccións de Bonferroni, Holm, Hommel ou Hochberg. En

función de se asumamos algún tipo de dependencia ou non entre os $p-valores$ poderemos rexeitar unhas ou outras hipóteses nulas.

a) Se **non facemos ningunha suposición adicional** sobre a dependencia dos $p-valores$ o máis acertado sería empregar o método de Bonferroni ou o de Holm, rexeitando con ambos métodos as hipóteses nulas $H_{0,4}$ e $H_{0,7}$.

b) Se **asumimos dependencia positiva ou independencia** dos $p-valores$ calquera dos métodos para o control do FWER podería ser utilizado. En función dos rexeitamentos para os distintos métodos indicados, que se poden observar no Cadro 3.1, escollemos empregar o método de Hommel xa que é o que permite rexeitar un maior número de hipóteses nulas. Así, para este caso rexeitariamos as hipóteses nulas $H_{0,4}$, $H_{0,7}$ e $H_{0,5}$.

2. Se decidimos levar a cabo un control do FDR: para este caso temos os métodos de Benjamini–Hochberg, Benjamini–Yekutieli e Benjamini–Liu. Novamente, distinguindo en función do tipo de dependencia que asumamos:

a) **Sen asunción ningunha sobre o tipo de dependencia**, o método máis axeitado é empregar o de Benjamini–Liu, rexeitando así as hipóteses nulas $H_{0,4}$, $H_{0,7}$, $H_{0,5}$ e $H_{0,8}$.

b) **Se asumimos independencia ou dependencia positiva entre os** $p-valores$ poderiamos empregar calquera dos distintos métodos para o control do FDR. Polo tanto, escollemos aquel que permite rexeitar un maior número de hipóteses nulas, neste caso vale tanto Benjamini–Hochberg como Benjamini–Liu, rexeitando as hipóteses nulas $H_{0,4}$, $H_{0,7}$, $H_{0,5}$ e $H_{0,8}$.

Método	i	1	2	3	4	5	6	7	8	9	10
Holm	$\frac{0,05}{11-i}$	0,005	0,0056	0,0063	0,0071	0,0083	0,01	0,013	0,017	0,025	0,05
Benjamini–Hochberg	$\frac{0,05i}{10}$	0,005	0,01	0,015	0,02	0,025	0,03	0,035	0,04	0,045	0,05
Benjamini–Yekutieli	$\frac{0,05i}{10 \cdot 2,93}$	0,0017	0,0034	0,0051	0,0068	0,0085	0,01	0,012	0,014	0,015	0,017
Benjamini–Liu	$\frac{0,05 \cdot 10}{(11-i)^2}$	0,005	0,0062	0,0078	0,01	0,014	0,02	0,031	0,056	0,125	0,5
Hochberg	$\frac{0,05}{i}$	0,05	0,025	0,017	0,013	0,01	0,0083	0,0071	0,0063	0,0056	0,005

Cadro 3.2: Valores cos que se comparan os $p-valores$ para cada un dos distintos métodos.

$\frac{0,05l}{i}$ $l=1,\cdots,i$	l	1	2	3	4	5	6	7	8	9	10
i											
1		0,05									
2		0,025	0,05								
3		0,017	0,033	0,05							
4		0,0125	0,025	0,0375	0,05						
5		0,01	0,02	0,03	0,04	0,05					
6		0,0083	0,017	0,025	0,033	0,042	0,05				
7		0,0071	0,014	0,021	0,028	0,035	0,043	0,05			
8		0,0063	0,013	0,019	0,025	0,031	0,038	0,044	0,05		
9		0,0056	0,011	0,017	0,022	0,028	0,033	0,039	0,044	0,05	
10		0,005	0,01	0,015	0,02	0,025	0,03	0,035	0,4	0,45	0,05

Cadro 3.3: Operacións correspondentes para localizar o máximo j do método de Hommel.

Capítulo 4

Aplicacións

Trala visión global da parte teórica das técnicas e métodos estatísticos que queremos levar a cabo, decidimos adentrarnos agora nun caso práctico e expoñer ditos coñecementos exemplificándoos. Tentaremos responder que axuste sería mellor en función do que desexemos responder e facer fincapé sobre as vantaxes e os inconvenientes en cada método.

Para levar a cabo isto apoiarémonos nos resultados obtidos no Anexo, que está composto por unha análise completa sobre os datos colon que se poden atopar na libraría [8] de R, que atopamos mencionados e estudados dunha maneira diferente en artigos como [11].

Os datos inclúen un análise xenético de 2000 xenes en 62 individuos diferentes. Destes individuos sabemos que 40 presentan un tumor de cancro de colon e os 22 restantes teñen un tecido normal. Queremos levar a cabo un contraste múltiple de igualdade de medias, é dicir, queremos responder para cada xene, se os individuos con tumor e os sans se van expresar distinto en media, tomando un certo nivel de significación $\alpha = 0{,}05$.

Partindo dunha idea sinxela podemos decidir levar a cabo o contraste en cada xene sen efectuar ningún tipo de corrección, co que vemos na análise levada a cabo no Anexo que se están obtendo 600 xenes que me permiten rexeitar a hipótese nula de igualdade de medias. O erro que se nos presenta, supoñendo que os 2000 tests fosen independentes, é o aumento da probabilidade do erro tipo I, que viría a ser $1 - (1 - 0{,}05)^{2000} \approx 1$, é dicir estaría a rexeitar ao menos unha hipótese nula case o 100 % das veces, supoñendo esta hipótese é certa para os 2000 xenes, cando só debería rexeitar o 5 %. Isto motiva a necesidade de realizar algún tipo de corrección.

Seguindo a literatura, véxase por exemplo [11], levaremos a cabo un test t de Welch, que se introduciu no Capítulo 1 para a comparación de medias, asumindo que se verifican as hipóteses necesarias para poder aplicalo. Unha das suposicións máis habituais para poder aplicar o test de comparación de medias é asumir que a distribución dos individuos das dúas poboacións segue unha normal. Co fin de comprobar dita suposición, nos 2000

xenes de sans e os 2000 de enfermos, no Anexo contrastaremos a hipótese de normalidade co test de Shapiro−Wilks. Para controlar o FWER e ver se ao sumo rexeitamos unha das hipóteses nulas contrastadas empregaremos a corrección de Bonferroni. Cos resultados obtidos, chegamos á conclusión de que non se pode asumir normalidade para a expresión xénica, xa que, como vemos no Anexo, podemos afirmar que ao menos unha das hipóteses a vai rexeitar. Na maioría dos estudos publicados que se poden atopar, incluso na análise de [11], estas comprobacións omítense e lévanse a cabo igualmente as correccións que lle interesen. Así, no que segue, asumiremos que se pode aplicar o test t de Welch. Noutro caso, unha alternativa sería empregar métodos non paramétricos coma o test de Mann−Whitney−Wilcoxon (coñecido tamén como Wilcoxon rank−sum ou u−test, que podemos consultar en [10]).

4.1. Control do FWER

4.1.1. Significado

Tiñamos que o FWER era a probabilidade de atopar ao menos un falso positivo, entre os 2000 contrastes que tiñamos. É dicir, se tomamos a decisión de levar a cabo un control do FWER, isto permitiranos saber se hai ao menos unha hipótese nula que se poida rexeitar, ao nivel de significación imposto, o cal nos di que hai ao menos un xene que nos permite distinguir entre sans e enfermos.

4.1.2. Resultados obtidos

Para controlar o FWER podemos empregar algún dos métodos vistos no Capítulo 3. Tanto se levamos a cabo unha corrección de Bonferroni, como de Holm ou de Hochberg, obtemos 47 xenes dos cales a súa expresión en media é diferente entre sans e enfermos. Se o que decidimos é considerar os resultados dados pola corrección de Hommel, detectaremos un xene a maiores, 48. Debemos ter en conta que os resultados que se obteñen dependen do tipo de dependencia que se asuma entre os $p - valores$, é dicir, se non asumimos dependencia positiva nin independencia entre os $p - valores$ empregar os métodos de corrección de Hochberg ou o de Hommel non sería correcto, porque non se verificarían as hipóteses necesarias.

4.1.3. Conclusións

Por un lado, todos os tests permiten afirmar que hai ao menos un xene que nos permite diferenciar sans de enfermos. E, por outro, todos detectan 47 xenes para os cales a expresión

en media entre sans e enfermos é distinta. Hommel detecta un xene a maiores, pero este caso só é válido se asumimos dependencia positiva.

Unha cuestión que podemos destacar é que os 47 atopados son os mesmos e incluso se atopan incluídos dentro dos detectados a través da corrección de Hommel. Outro detalle é que, en virtude da Proposición 2.1, teremos que estes xenes se atoparán tamén en estudos nos que decidamos empregar un control do FDR. Estas comprobacións poden apreciarse no documento anexado.

4.2. Control do FDR

4.2.1. Significado

Controlando o FDR establecíamos un límite de significación, neste caso para o conxunto dos 2000 test, de forma que, entre todos os que son considerados como significativos, a proporción esperada de falsos positivos non supere un valor concreto. É dicir, para o caso particular tratado no Anexo, de entre os 600 que se detectaron sen ningún tipo de corrección (referentes aos tests significativos), para un nivel de significación do 5 % fixado, tralo control do FDR teremos asegurado que o número de falsos positivos esperados non será superior a 30 xenes. Así, controlando o FDR podemos detectar que xenes presentan unha diferenza significativa en media, controlando sempre a probabilidade do erro tipo I.

4.2.2. Resultados obtidos

Novamente, practicando o control do FDR coas correspondentes correccións analizadas no Capítulo 3. Explorando as saídas, obtemos que Benjamini−Hochberg nos permite afirmar que 359 xenes se van expresar de xeito distinto en persoas sans e en persoas enfermas; para Benjamini−Yekutieli obtemos 138 e para Benjamini−Liu 47.

4.2.3. Conclusións

Independentemente da corrección empregada, todas elas detectan como significativas ao menos 47 xenes. Os 47 xenes que rexeitan para Benjamini−Liu están incluídos nos 138 que rexeitamos para Benjamini−Yekutieli e ambos aparecen nos que se rexeitan empregando Benjamini−Hochberg, como ben se comproba ao final do análise completo que incluímos no Anexo. Cada un deles permite identificar que xenes en concreto se expresan, en media, distinto entre persoas sans e enfermas.

Unha vez asumidas as hipóteses necesarias para aplicar o test t de Welch, podemos propor por outra banda o tipo de dependencia existente entre os 2000 $p-valores$. Baixo

a suposición de independencia ou dependencia positiva podemos decidir levar a cabo a corrección de Benjamini−Hochberg, entre outros, que nos permite identificar un maior número de xenes que calquera dos outros métodos. Porén, se non puidésemos asumir este tipo de dependencia, teremos que acudir, por exemplo, á corrección de Benjamini−Liu.

4.3. Consideracións finais

Tendo en conta que para un gran número de contrastes pode resultar moi complexo verificar que tipo de dependencia presentan os $p-valores$, deberiamos proceder empregando os métodos que menos hipóteses precise comprobar. Para o noso caso en concreto tomariamos os resultados dados aplicando os métodos de Bonferroni, Holm ou Benjamini−Liu.

Analicemos agora cando se decide empregar uns ou outros métodos. Na maioría dos casos, se realizamos unha corrección para controlar o FDR, os xenes que esteamos a obter mediante este tipo de métodos tamén incluirán os que se obteñen se fixésemos un control sobre o FWER. A diferenza principal radica no tipo de cuestións que queiramos responder; por unha banda se queremos saber se unha persoa está enferma fariamos un estudo sobre o FDR porque este devolve que xenes en concreto nos permitirían distinguir persoas sans de persoas enfermas e sería suficiente con observar que ocorre nestes xenes para esa persoa. Porén, se o que quixésemos facer é un estudo sobre a expresión xénica e concluír se é igual, en media, tanto para sans coma para enfermos, resultaría suficiente controlar o FWER.

Na últimas liñas do Anexo podemos observar que se inclúe unha alternativa para descargar un ficheiro que contén o nome dos respectivos xenes empregados nesta análise, de onde sería posible seleccionar os xenes concretos que estariamos atopando tras aplicar correccións que controlan o FDR.

Anexos

Anexo A: Análise completo de datos

Contraste simple de dúas poboacións

Simplifiquemos a situación considerando un único xene, o 25. Queremos saber se, en media, hai diferenzas significativas na expresión do xene 25 entre as persoas sans e as enfermas.

```
san<-x[y==0,25]
enferma<-x[y==1,25]
length(san);
```

```
## [1] 22
```

```
length(enferma)
```

```
## [1] 40
```

É dicir, estamos no contraste de medias para dúas poboacións de tamaño distinto. Comprobemos se estamos ante poboacións normais (onde, en tal caso, traballamos con varianzas descoñecidas e distintas).

Nótese que neste caso poderiamos contrastar a hipótese de se as varianzas son iguais comprobando a hipótese de homoscedasticidade co test de Levene (é dicir, contrastar se a hipótese nula de que a homoxeneidade das varianzas é certa). En caso de poder asumir homoscedasticidade, poderiamos empregar o test correspondente. Neste caso, en concordancia co resto do documento suporemos que non se pode asumir que as varianzas son iguais.

Para comprobar normalidade empregamos o test de Shapiro−Wilks que está xa implementado en R. A grandes trazos diremos que o que fai é un contraste de hipóteses onde a hipótese nula é que a variable aleatoria asociada aos datos segue unha distribución normal e a alternativa que non a segue. É dicir, poderemos rexeitar a hipótese de normalidade se obtivésemos $p - valores$ máis pequenos que o nivel de significación imposto.

```
shapiro.test(san);shapiro.test(enferma)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  san
## W = 0.97197, p-value = 0.7563
```

```
##
##  Shapiro-Wilk normality test
##
## data:  enferma
## W = 0.98471, p-value = 0.8552
```

Obtemos para ambos casos un $p-valor$ moi alto, 0,7563 e 0,8552 respectivamente, polo que non temos probas significativas para rexeitar a hipótese de normalidade para calquera dos niveis de significación habituais 0,01, 0,05 ou 0,10 e podemos asumir que as nosas poboacións proveñen dunha distribución normal.

```
t.test(san,enferma,alternative="two.sided", paired=FALSE, var.equal = FALSE)
#As nosas mostras non son apareadas, polo que empregaremos paired=FALSE
```

```
##
##  Welch Two Sample t-test
##
## data:  san and enferma
## t = 0.46854, df = 34.106, p-value = 0.6424
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1911591  0.3057315
## sample estimates:
## mean of x mean of y
##  2.227369  2.170083
```

Obtemos un $p-valor$ 0,6424 que é demasiado grande, polo que non podemos rexeitar que a hipótese nula de igualdade de medias para $\alpha = 0{,}05$. É dicir, non hai diferenzas significativas ao nivel do 5 % da expresión en media entre sans e enfermos para o xene 25.

Nótese que se se quixese cambiar o nivel de confianza α nos intervalos de confianza, poderíase cambiar a función anterior incluíndo o argumento `conf.level` $= 1-\alpha$. Neste caso o intervalo de confianza ao nivel $(1-0{,}05)$ sería o que se mostra baixo o rótulo `confidence interval`, $(-0{,}19, 0{,}31)$.

Por outra banda, quero aproveitar é incluír que se o que quixésemos é levar a cabo un contraste unilateral o único que habería que cambiar é o tipo de alternativa que queremos empregar. Por exemplo, para contrastar para o xene 25 que en media a súa expresión é menor en sans que enfermos sería:

```
t.test(san,enferma,alternative="less", paired=FALSE, var.equal = FALSE)
#se o contraste fose menor sería alternative="greater"
```

```
##
##  Welch Two Sample t-test
##
## data:  san and enferma
## t = 0.46854, df = 34.106, p-value = 0.6788
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.2640102
## sample estimates:
## mean of x mean of y
##  2.227369  2.170083
```

Obtemos un $p-valor$ que non é significativo polo que para o xene 25, en media, a expresión en sans non se pode asumir menor que a de enfermos.

Contrastes múltiples

Queremos agora facer 2000 contrastes simultáneos, un para cada xene, para contrastar a hipótese nula de se en media a expresión entre os sans e enfermos é igual. Para iso necesitamos comprobar se os datos proceden de poboacións normais para estar nas hipóteses axeitadas para empregar o test t de Welch. Para levalo a cabo teremos por un lado 2000 variables aleatorias (xenes) de persoas sans e 2000 de enfermas, polo que estamos a realizar un total de 4000 contrastes de normalidade.

```
#creo matrices de datos das cales unha ten a información dos 2000 xenes para
#os sans e a outra os 2000 xenes nas persoas con tumor.
sans<-x[y==0,] #poboacións de persoas sans
enfermos<-x[y==1,] #poboacións de persoas enfermas
```

Agora o que vou facer é crear un vector no que vou almacenar os 2000 $p-valores$ que obteño tras aplicar o contraste de normalidade para sans e outro para os de enfermos. Os $p-valores$ serán os correspondentes de realizar un contraste de normalidade mediante o test de Shapiro-Wilks para cada xene.

```
pvs<-c()
pve<-c()
for (i in 1:2000) {
  pvs[i]<-shapiro.test(sans[,i])$p.value
  pve[i]<-shapiro.test(enfermos[,i])$p.value
}
```

Tendo en conta que realizamos 4000 contrastes, para que estes sexan significativos, empregando o criterio de Bonferroni, buscaremos aqueles $p-valores$ que sexan menores que 0,05/4000.

```
pvs[pvs<0.05/4000]
```

```
## [1] 4.270262e-06
```

```
pve[pve<0.05/4000]
```

```
##  [1] 1.910592e-06 2.811979e-06 6.308665e-06 4.427120e-06 4.113882e-07
##  [6] 2.521306e-06 9.375622e-06 1.731609e-06 1.628694e-08 4.513447e-07
## [11] 1.786977e-08 4.123549e-06 6.201729e-07 7.190603e-06 6.313923e-07
## [16] 2.505073e-07 1.286861e-06 8.020074e-07 4.887897e-06 8.004962e-08
## [21] 6.833367e-07
```

Obtemos así que para ambas poboacións existe polo menos un xene que permite rexeitar a hipótese de normalidade.

Observación .1. O test t de Welch funciona ben cando se aplica, por exemplo, a poboacións que proveñen dunha distribución normal. Como estamos nunha situación de estatística multivariante, comprobar que esta condición se verifique non é unha tarefa doada que nos interese para este traballo. Polo tanto, decidimos asumir como certa a hipótese de normalidade para o noso conxunto de datos.

Asumindo que, neste caso, pódese empregar o test t de Welch para comparar as medias de cada un dos xenes, creamos un vector de $p-valores$ onde almacenaremos os 2000 $p-valores$ resultantes de levar a cabo o contraste t de Welch sen aplicar ningún tipo de corrección.

```
p_valores<-c()
for (i in 1:2000){
  p_valores[i]<-t.test(sans[,i],enfermos[,i],var.equal=FALSE,paired=FALSE,
  alternative="two.sided")$p.value
}

length(p_valores[p_valores<0.05])
```

```
## [1] 600
```

Neste caso inicial no que non realizamos ningún tipo de corrección obtemos que podemos rexeitar para 600 xenes a hipótese nula de igualdade de medias entre sans e enfermos.

Vaiamos agora cos distintos métodos de corrección que podemos atopar no Capítulo 3.

Comezamos polos métodos de correccións que permiten controlar o FWER. Teremos dúas liñas para cada un dos distintos métodos de corrección (Bonferroni, Holm, Hochberg e Hommel respectivamente), unha das liñas para aplicar a corrección correspondente e a outra devolvendo o número de xenes que rexeitan a hipótese nula para un nivel de significación $\alpha = 0{,}05$.

```
Bonferroni<-p.adjust(p_valores,method="bonferroni") #Bonferroni
length(Bonferroni[Bonferroni<0.05]) #obtemos 47 xenes
```

```
## [1] 47
```

```
Holm<-p.adjust(p_valores,method="holm") #método de Holm
length(Holm[Holm<0.05]) #47
```

```
## [1] 47
```

```
Hochberg<-p.adjust(p_valores,method="hochberg") #método de Hochberg
length(Hochberg[Hochberg<0.05]) #47
```

```
## [1] 47
```

```
Hommel<-p.adjust(p_valores,method="hommel") #método de Hommel
length(Hommel[Hommel<0.05]) #48
```

```
## [1] 48
```

Observamos que facendo un control sobre o FWER poderemos rexeitar 47 xenes, agás no caso en que decida levar a cabo a corrección de Hommel que atopará un xene a maiores dos anteriores, pero este necesita cumprir a hipótese de dependencia positiva entre os $p-valores$.

Vaiamos agora cos métodos de corrección que controlan o FDR cando asumimos independencia ou dependencia positiva entre os $p-valores$, que serán, respectivamente, Benjamini−Hochberg e Benjamini−Yekutieli.

```
BH<-p.adjust(p_valores,method="BH") #Benjamini-Hochberg
length(BH[BH<0.05]) #359 xenes
```

```
## [1] 359
```

```
BY<-p.adjust(p_valores,method="BY") #Benjamini-Yekutieli
length(BY[BY<0.05]) #138
```

```
## [1] 138
```

Podemos ver que os tests que controlan o FDR obteñen moitos máis xenes para os que podemos distinguir entre sans e enfermos, temos 359 para Benjamini−Hochberg e 138 para Benjamini−Yekutieli fronte aos 47 ou 48 que obtiñamos se controlabamos o FWER.

Por unha banda, se queremos identificar os xenes aos que se refiren os métodos que controlan o FDR podemos proceder do seguinte xeito:

a) Para os 47 atopados por Bonferroni, Holm e Hochberg, respectivamente:

```
hipbon<-c()
hipholm<-c()
hiphoch<-c()
for (i in 1:47){
  hipbon[i]<-which(Bonferroni==Bonferroni[Bonferroni<0.05][i])
  hipholm[i]<-which(Holm==Holm[Holm<0.05][i])
```

```
  hiphoch[i]<-which(Hochberg==Hochberg[Hochberg<0.05][i])
}
hipbon #hipóteses detectadas por Bonferroni
hipholm #Holm
hiphoch #Hochberg
```

b) Para os 48 atopados polo método de Hommel:

```
hiphomm<-c()
for (i in 1:48){
  hiphomm[i]<-which(Hommel==Hommel[Hommel<0.05][i])
}
hiphomm #hipóteses de Hommel
```

Por outra, se queremos identificar os xenes aos que se refiren os métodos que controlan o FDR podemos proceder do seguinte xeito:

a) Para os 359 que atopamos con Benjamini-Hochberg, se temos en conta que este seleccionaba os xenes en funcións dos $p-valores$ ordenados en orde ascendente, as hipóteses correspondentes serán aquelas cuxos $p-valores$ ocupen as 359 primeiras posicións se os dispoñemos ordenados, é dicir,

```
orden<-sort(p_valores)
ordenbh<-orden[1:359] #collemos os 359 $p-valores$ máis pequenos
length(ordenbh)
```

```
## [1] 359
```

```
hipbh<-c() #inicializamos un vector que vai incluír as hipóteses
#correspondentes á corrección de Benjamini-Hochberg.
for (i in 1:359){
  j<-which(p_valores==ordenbh[i])
  hipbh[i]<-j
}
```

```
hipbh #obtemos as 359 hipóteses correspondentes
```

b) Para os 138 que atopamos empregando o método de Benjamini-Yekutieli, procedemos de xeito análogo xa que os $p-valores$ tamén se atopan dende o máis significativo ao menos,

```
ordenby<-orden[1:138] #collemos os 138 $p-valores$ máis pequenos
hipby<-c() #inicializamos un vector que vai incluír as hipóteses
#correspondentes á corrección de Benjamini-Yekutieli.
for (i in 1:138){
  j<-which(p_valores==ordenby[i])
  hipby[i]<-j
}
hipby #obtemos as 138 hipóteses correspondentes
```

Como ben mencionamos no seu momento estamos facendo suposicións que non somos capaces de contrastar, como a dependencia positiva dos $p-valores$, polo que agora imos levar a cabo un contraste de Benjamini-Liu no que non necesitamos dita hipótese. Este método non está entre as funcións programadas de R polo que o introduzo manualmente.

```
#orden era o vector que contiña os $p-valores$ en orde ascendente.
LIU<-c()
for (i in 1:200){
  hi<-min(0.05,(0.05*2000)/(2000+1-i)^2)
  if (orden[i]<=hi) {
    LIU[i]<-orden[i]
  } else {'Rexeito todas as h_{(j)} con j< '}
}
length(LIU)
```

```
## [1] 47
```

```
hipLIU<-c()
for(i in 1:47){
  j<-which(p_valores==LIU[i])
```

```
  hipLIU[i]<-j
}
hipLIU #hipóteses rexeitadas por Benjamini-Liu
```

Identificación concreta dos xenes

Como ben sabemos, debido ao control do FDR podemos distinguir os xenes que se expresan diferente, en media, entre sans e enfermos. Como contamos co ficheiro de datos que contén o nome de devanditos xenes podemos coñecelos:

```
xenes<-read.table("datos2.txt",sep="")
xenes[hipbon,1] #xenes identificados por Bonferroni, que coinciden con Holm
#e Hochberg.
```

```
##  [1] "Hsa.467"   "Hsa.1013"  "Hsa.8125"  "Hsa.2361"  "Hsa.36694" "Hsa.957"
##  [7] "Hsa.832"   "Hsa.692"   "Hsa.8147"  "Hsa.692"   "Hsa.821"   "Hsa.36689"
## [13] "Hsa.3152"  "Hsa.1221"  "Hsa.37937" "Hsa.831"   "Hsa.3306"  "Hsa.3305"
## [19] "Hsa.692"   "Hsa.773"   "Hsa.2863"  "Hsa.41280" "Hsa.1131"  "Hsa.41280"
## [25] "Hsa.36952" "Hsa.3331"  "Hsa.9218"  "Hsa.549"   "Hsa.462"   "Hsa.8781"
## [31] "Hsa.3263"  "Hsa.1832"  "Hsa.2344"  "Hsa.2818"  "Hsa.9353"  "Hsa.2928"
## [37] "Hsa.2097"  "Hsa.662"   "Hsa.14069" "Hsa.2821"  "Hsa.2645"  "Hsa.601"
## [43] "Hsa.6814"  "Hsa.2291"  "Hsa.25322" "Hsa.466"   "Hsa.11616"
```

```
xenes[hiphomm,1] #Hommel
```

```
##  [1] "Hsa.467"   "Hsa.1013"  "Hsa.8125"  "Hsa.2361"  "Hsa.36694" "Hsa.957"
##  [7] "Hsa.832"   "Hsa.692"   "Hsa.8147"  "Hsa.692"   "Hsa.821"   "Hsa.36689"
## [13] "Hsa.3152"  "Hsa.1221"  "Hsa.37937" "Hsa.831"   "Hsa.3306"  "Hsa.3305"
## [19] "Hsa.692"   "Hsa.773"   "Hsa.2863"  "Hsa.41280" "Hsa.1131"  "Hsa.41280"
## [25] "Hsa.36952" "Hsa.3331"  "Hsa.9218"  "Hsa.549"   "Hsa.462"   "Hsa.8781"
## [31] "Hsa.3263"  "Hsa.1832"  "Hsa.2344"  "Hsa.2818"  "Hsa.9353"  "Hsa.2928"
## [37] "Hsa.10664" "Hsa.2097"  "Hsa.662"   "Hsa.14069" "Hsa.2821"  "Hsa.2645"
## [43] "Hsa.601"   "Hsa.6814"  "Hsa.2291"  "Hsa.25322" "Hsa.466"   "Hsa.11616"
```

```
xenes[hipbh,1] #xenes identificados por Benjamini-Hochberg
```

```
##   [1] "Hsa.2097"  "Hsa.37937" "Hsa.2291"  "Hsa.36689" "Hsa.549"   "Hsa.1832"
##   [7] "Hsa.831"   "Hsa.36952" "Hsa.8147"  "Hsa.3306"  "Hsa.601"   "Hsa.6814"
##  [13] "Hsa.957"   "Hsa.1131"  "Hsa.2344"  "Hsa.3331"  "Hsa.2645"  "Hsa.3263"
##  [19] "Hsa.462"   "Hsa.1013"  "Hsa.36694" "Hsa.8125"  "Hsa.2928"  "Hsa.467"
##  [25] "Hsa.41280" "Hsa.41280" "Hsa.25322" "Hsa.773"   "Hsa.14069" "Hsa.3305"
##  [31] "Hsa.3152"  "Hsa.821"   "Hsa.2361"  "Hsa.692"   "Hsa.1221"  "Hsa.11616"
##  [37] "Hsa.2821"  "Hsa.662"   "Hsa.692"   "Hsa.832"   "Hsa.9218"  "Hsa.692"
##  [43] "Hsa.8781"  "Hsa.2818"  "Hsa.9353"  "Hsa.466"   "Hsa.2863"  "Hsa.10664"
##  [49] "Hsa.36696" "Hsa.10755" "Hsa.4689"  "Hsa.41323" "Hsa.31933" "Hsa.2705"
##  [55] "Hsa.853"   "Hsa.1047"  "Hsa.8175"  "Hsa.7"     "Hsa.951"   "Hsa.5971"
##  [61] "Hsa.1130"  "Hsa.1387"  "Hsa.2196"  "Hsa.3296"  "Hsa.2553"  "Hsa.627"
##  [67] "Hsa.2250"  "Hsa.1073"  "Hsa.1610"  "Hsa.1454"  "Hsa.41338" "Hsa.2456"
##  [73] "Hsa.3803"  "Hsa.5444"  "Hsa.2715"  "Hsa.41282" "Hsa.1207"  "Hsa.229"
##  [79] "Hsa.27537" "Hsa.329"   "Hsa.27686" "Hsa.1902"  "Hsa.11673" "Hsa.1479"
##  [85] "Hsa.8068"  "Hsa.1165"  "Hsa.2598"  "Hsa.8219"  "Hsa.3093"  "Hsa.3349"
##  [91] "Hsa.3016"  "Hsa.9246"  "Hsa.2747"  "Hsa.1410"  "Hsa.2451"  "Hsa.28914"
##  [97] "Hsa.42186" "Hsa.1205"  "Hsa.1143"  "Hsa.2827"  "Hsa.94"    "Hsa.2646"
## [103] "Hsa.5211"  "Hsa.31630" "Hsa.2800"  "Hsa.43331" "Hsa.8223"  "Hsa.9972"
## [109] "Hsa.2084"  "Hsa.3230"  "Hsa.1013"  "Hsa.6472"  "Hsa.2191"  "Hsa.2316"
## [115] "Hsa.43279" "Hsa.24944" "Hsa.2959"  "Hsa.8040"  "Hsa.1588"  "Hsa.7395"
## [121] "Hsa.879"   "Hsa.168"   "Hsa.24582" "Hsa.286"   "Hsa.1806"  "Hsa.103"
## [127] "Hsa.2710"  "Hsa.2700"  "Hsa.810"   "Hsa.451"   "Hsa.2591"  "Hsa.878"
## [133] "Hsa.732"   "Hsa.3001"  "Hsa.2862"  "Hsa.39753" "Hsa.3007"  "Hsa.2939"
## [139] "Hsa.2795"  "Hsa.3083"  "Hsa.1726"  "Hsa.1198"  "Hsa.812"   "Hsa.2773"
## [145] "Hsa.6030"  "Hsa.17564" "Hsa.36655" "Hsa.33"    "Hsa.316"   "Hsa.1240"
## [151] "Hsa.9994"  "Hsa.1132"  "Hsa.24506" "Hsa.17901" "Hsa.3250"  "Hsa.4252"
## [157] "Hsa.6458"  "Hsa.490"   "Hsa.1985"  "Hsa.678"   "Hsa.929"   "Hsa.18790"
## [163] "Hsa.26528" "Hsa.442"   "Hsa.2513"  "Hsa.1617"  "Hsa.3239"  "Hsa.3252"
## [169] "Hsa.491"   "Hsa.26920" "Hsa.56"    "Hsa.816"   "Hsa.21195" "Hsa.3088"
## [175] "Hsa.3141"  "Hsa.1435"  "Hsa.1648"  "Hsa.5392"  "Hsa.702"   "Hsa.2560"
## [181] "Hsa.330"   "Hsa.2584"  "Hsa.663"   "Hsa.3068"  "Hsa.1272"  "Hsa.1598"
## [187] "Hsa.3566"  "Hsa.2964"  "Hsa.2618"  "Hsa.1955"  "Hsa.60"    "Hsa.2357"
## [193] "Hsa.954"   "Hsa.2275"  "Hsa.7203"  "Hsa.1579"  "Hsa.22762" "Hsa.37541"
```

```
## [199] "Hsa.5398"  "Hsa.6288"  "Hsa.960"   "Hsa.1423"  "Hsa.2568"  "Hsa.127"
## [205] "Hsa.1591"  "Hsa.6317"  "Hsa.1515"  "Hsa.40063" "Hsa.1278"  "Hsa.36685"
## [211] "Hsa.25522" "Hsa.562"   "Hsa.2933"  "Hsa.2126"  "Hsa.2644"  "Hsa.23699"
## [217] "Hsa.1660"  "Hsa.558"   "Hsa.447"   "Hsa.1317"  "Hsa.120"   "Hsa.6782"
## [223] "Hsa.8214"  "Hsa.25748" "Hsa.9255"  "Hsa.28939" "Hsa.2357"  "Hsa.1860"
## [229] "Hsa.996"   "Hsa.57"    "Hsa.3963"  "Hsa.726"   "Hsa.16742" "Hsa.3003"
## [235] "Hsa.41283" "Hsa.2665"  "Hsa.2576"  "Hsa.477"   "Hsa.305"   "Hsa.2837"
## [241] "HSAC07"    "HSAC07"    "HSAC07"    "HSAC07"    "Hsa.13610" "Hsa.865"
## [247] "Hsa.2966"  "Hsa.404"   "Hsa.489"   "Hsa.7652"  "Hsa.16742" "Hsa.341"
## [253] "Hsa.20164" "Hsa.2487"  "Hsa.717"   "Hsa.44244" "Hsa.1737"  "Hsa.1516"
## [259] "Hsa.2950"  "Hsa.1140"  "Hsa.6317"  "Hsa.3015"  "Hsa.2613"  "Hsa.42949"
## [265] "Hsa.43405" "Hsa.594"   "Hsa.1045"  "Hsa.12465" "Hsa.852"   "Hsa.23824"
## [271] "Hsa.1145"  "Hsa.1763"  "Hsa.212"   "Hsa.3115"  "UMGAP"     "UMGAP"
## [277] "UMGAP"     "UMGAP"     "Hsa.25777" "Hsa.2867"  "Hsa.5464"  "Hsa.32463"
## [283] "Hsa.42746" "Hsa.3876"  "Hsa.9744"  "Hsa.36665" "Hsa.2957"  "Hsa.16296"
## [289] "Hsa.902"   "Hsa.2419"  "Hsa.72"    "Hsa.22614" "Hsa.3209"  "Hsa.2856"
## [295] "Hsa.2829"  "Hsa.24373" "Hsa.1171"  "Hsa.934"   "Hsa.579"   "Hsa.3010"
## [301] "Hsa.33965" "Hsa.2565"  "Hsa.45446" "Hsa.2354"  "Hsa.1592"  "Hsa.34431"
## [307] "Hsa.3348"  "Hsa.1694"  "Hsa.2503"  "Hsa.9235"  "Hsa.1732"  "Hsa.422"
## [313] "Hsa.1280"  "Hsa.1732"  "Hsa.7048"  "Hsa.367"   "Hsa.21901" "Hsa.7736"
## [319] "Hsa.1682"  "Hsa.1316"  "Hsa.3194"  "Hsa.2555"  "Hsa.2842"  "Hsa.33572"
## [325] "Hsa.34575" "Hsa.2997"  "Hsa.42625" "Hsa.41187" "Hsa.10358" "Hsa.1920"
## [331] "Hsa.5544"  "Hsa.2051"  "Hsa.2387"  "Hsa.2809"  "Hsa.1209"  "Hsa.2967"
## [337] "Hsa.1994"  "Hsa.5226"  "Hsa.3157"  "Hsa.4996"  "Hsa.5908"  "Hsa.3002"
## [343] "Hsa.5346"  "Hsa.2280"  "Hsa.2777"  "Hsa.421"   "Hsa.35496" "Hsa.3454"
## [349] "Hsa.2221"  "Hsa.3307"  "Hsa.1877"  "Hsa.24877" "Hsa.1288"  "Hsa.1978"
## [355] "Hsa.24121" "Hsa.667"   "Hsa.3952"  "Hsa.2917"  "Hsa.2951"
```

```
xenes[hipby,1] #Benjamini-Yekutieli
```

```
##   [1] "Hsa.2097"  "Hsa.37937" "Hsa.2291"  "Hsa.36689" "Hsa.549"   "Hsa.1832"
##   [7] "Hsa.831"   "Hsa.36952" "Hsa.8147"  "Hsa.3306"  "Hsa.601"   "Hsa.6814"
##  [13] "Hsa.957"   "Hsa.1131"  "Hsa.2344"  "Hsa.3331"  "Hsa.2645"  "Hsa.3263"
##  [19] "Hsa.462"   "Hsa.1013"  "Hsa.36694" "Hsa.8125"  "Hsa.2928"  "Hsa.467"
##  [25] "Hsa.41280" "Hsa.41280" "Hsa.25322" "Hsa.773"   "Hsa.14069" "Hsa.3305"
```

```
##  [31] "Hsa.3152"  "Hsa.821"   "Hsa.2361"  "Hsa.692"   "Hsa.1221"  "Hsa.11616"
##  [37] "Hsa.2821"  "Hsa.662"   "Hsa.692"   "Hsa.832"   "Hsa.9218"  "Hsa.692"
##  [43] "Hsa.8781"  "Hsa.2818"  "Hsa.9353"  "Hsa.466"   "Hsa.2863"  "Hsa.10664"
##  [49] "Hsa.36696" "Hsa.10755" "Hsa.4689"  "Hsa.41323" "Hsa.31933" "Hsa.2705"
##  [55] "Hsa.853"   "Hsa.1047"  "Hsa.8175"  "Hsa.7"     "Hsa.951"   "Hsa.5971"
##  [61] "Hsa.1130"  "Hsa.1387"  "Hsa.2196"  "Hsa.3296"  "Hsa.2553"  "Hsa.627"
##  [67] "Hsa.2250"  "Hsa.1073"  "Hsa.1610"  "Hsa.1454"  "Hsa.41338" "Hsa.2456"
##  [73] "Hsa.3803"  "Hsa.5444"  "Hsa.2715"  "Hsa.41282" "Hsa.1207"  "Hsa.229"
##  [79] "Hsa.27537" "Hsa.329"   "Hsa.27686" "Hsa.1902"  "Hsa.11673" "Hsa.1479"
##  [85] "Hsa.8068"  "Hsa.1165"  "Hsa.2598"  "Hsa.8219"  "Hsa.3093"  "Hsa.3349"
##  [91] "Hsa.3016"  "Hsa.9246"  "Hsa.2747"  "Hsa.1410"  "Hsa.2451"  "Hsa.28914"
##  [97] "Hsa.42186" "Hsa.1205"  "Hsa.1143"  "Hsa.2827"  "Hsa.94"    "Hsa.2646"
## [103] "Hsa.5211"  "Hsa.31630" "Hsa.2800"  "Hsa.43331" "Hsa.8223"  "Hsa.9972"
## [109] "Hsa.2084"  "Hsa.3230"  "Hsa.1013"  "Hsa.6472"  "Hsa.2191"  "Hsa.2316"
## [115] "Hsa.43279" "Hsa.24944" "Hsa.2959"  "Hsa.8040"  "Hsa.1588"  "Hsa.7395"
## [121] "Hsa.879"   "Hsa.168"   "Hsa.24582" "Hsa.286"   "Hsa.1806"  "Hsa.103"
## [127] "Hsa.2710"  "Hsa.2700"  "Hsa.810"   "Hsa.451"   "Hsa.2591"  "Hsa.878"
## [133] "Hsa.732"   "Hsa.3001"  "Hsa.2862"  "Hsa.39753" "Hsa.3007"  "Hsa.2939"
```

```
xenes[hipLIU,1] #Benjamini-Liu
```

```
##  [1] "Hsa.2097"  "Hsa.37937" "Hsa.2291"  "Hsa.36689" "Hsa.549"   "Hsa.1832"
##  [7] "Hsa.831"   "Hsa.36952" "Hsa.8147"  "Hsa.3306"  "Hsa.601"   "Hsa.6814"
## [13] "Hsa.957"   "Hsa.1131"  "Hsa.2344"  "Hsa.3331"  "Hsa.2645"  "Hsa.3263"
## [19] "Hsa.462"   "Hsa.1013"  "Hsa.36694" "Hsa.8125"  "Hsa.2928"  "Hsa.467"
## [25] "Hsa.41280" "Hsa.41280" "Hsa.25322" "Hsa.773"   "Hsa.14069" "Hsa.3305"
## [31] "Hsa.3152"  "Hsa.821"   "Hsa.2361"  "Hsa.692"   "Hsa.1221"  "Hsa.11616"
## [37] "Hsa.2821"  "Hsa.662"   "Hsa.692"   "Hsa.832"   "Hsa.9218"  "Hsa.692"
## [43] "Hsa.8781"  "Hsa.2818"  "Hsa.9353"  "Hsa.466"   "Hsa.2863"
```

Comprobacións

Como último detalle podemos incluír unhas breves liñas nas que comezamos comprobando que se eliximos controlar o FWER, independentemente do método de corrección elixido, rexeitaremos as mesmas hipóteses, agás que poidamos verificar dependencia positiva entre os $p-valores$, que detectamos a maiores o xene 1634:

```
table(hipbon==hiphoch)
```

```
##
## TRUE
##   47
```

```
table(hipbon==hipholm)
```

```
##
## TRUE
##   47
```

Continuamos vendo que os xenes rexeitados por un método que controle o FDR serán tamén rexeitados (e aparecerán ao comezo) se cambiamos a un método que atope máis.

```
table(xenes[hipbh,1][1:138]==xenes[hipby,1]) #os 138 atopados por Benjamini-
#Yekutieli atópanse xa nas 138 primeiras posicións dos atopados por
#Benjamini-Hochberg.
```

```
##
## TRUE
##   138
```

```
table(xenes[hipbh,1][1:47]==xenes[hipLIU,1]) #análogo para os 47 atopados
#por Benjamini-Liu.
```

```
##
## TRUE
##   47
```

```
table(xenes[hipby,1][1:47]==xenes[hipLIU,1]) #do mesmo xeito tamén aparecen
#rexeitados por Benjamini-Yekutieli os 47 que se rexeitaron por Benjamini-Liu.
```

```
##
## TRUE
##   47
```

É máis, podemos ver que as hipóteses rexeitadas tras levar a cabo un control do FWER estarán incluídas se empregamos métodos de corrección para controlar o FDR. Será suficiente con ver que as hipóteses rexeitadas por Bonferroni están incluídas/son as mesmas que as de Benjamini-Liu, polo que levamos visto ata este punto:

```
a<-c() #inicializo un vector que vai conter as hipóteses nas que coincidan
#ambos métodos de correccións.
for (i in 1:47){
  a[i]<-which(hipLIU==hipbon[i])
}
length(a)
```

```
## [1] 47
```

Observamos que o vector a ten lonxitude de 47, é dicir, ambos métodos consideran as mesmas hipóteses.

Documentación completa

Se queremos ver cales son os xenes correspondentes podemos descargar o ficheiro completo de datos que contén o nome dos xenes na mesma orde coa que aparecen na matriz dos datos. Os comandos a empregar poden ser:

```
download.file(
   url = "http://genomics-pubs.princeton.edu/oncology/affydata/names.html",
   dest = "cancer-colon.xml"
)
```

O ficheiro gardarase no cartafol do ordenador que esteamos a empregar.

Bibliografía

[1] Herve Abdi. Holm's sequential Bonferroni procedure. *Encyclopedia of Research Design*, 1(8):1–8, 2010.

[2] Pablo Martínez Camblor. Ajuste del valor-p por contrastes múltiples. *Revista Chilena de Salud Pública*, 16(3):225–232, 2012.

[3] Mark Chang. *Modern Issues and Methods in Biostatistics*. Springer Science & Business Media, Nueva York, 2011.

[4] Yoav Benjamini e Abba M.Krieger e Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

[5] Hugo Alvarado e Carmen Batanero. Significado del teorema central del límite en textos universitarios de probabilidad y estadística. *Estudios Pedagógicos (Valdivia)*, 34(2):7–28, 2008.

[6] Yoav Benajamini e Dan Drai e Greg Elmer e Neri Kafkani e Ilan Golani. Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research*, 125(1-2):279–284, 2001.

[7] Yoav Benjamini e Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

[8] Boxiang Wang e Hui Zou. sdwd: Sparse distance weighted discrimination. Dispoñible en https://CRAN.R-project.org/package=sdwd, 2020. R package version 1.0.5.

[9] Yifan Huang e Jason C Hsu. Hochberg's step-up method: cutting corners off Holm's step-down method. *Biometrika*, 94(4):965–975, 2007.

[10] Morten W Fagerland e Leiv Sandvik. The Wilcoxon–Mann–Whitney test under scrutiny. *Statistics in Medicine*, 28(10):1487–1497, 2009.

[11] James J Chen e Paula K Robenson e Michael J Schell. The false discovery rate: a key concept in large-scale genetic studies. *Cancer Control*, 17(1):58–62, 2010.

[12] Rosa María Crujeiras Casais e Pedro Roca Faraldo. *Manual de Estadística Básica para Ciencias de la Salud.* Universidade de Santiago de Compostela, Departamento de Estadística e Investigación Operativa, 2010.

[13] G. Hommel. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2):383–386, 06 1988.

[14] Hangcheng Liu. Comparing Welch's ANOVA, a Kruskal-Wallis test and traditional ANOVA in case of Heterogeneity of Variance. Trabajo fin de máster, Universidad de Virginia Commonwealth. 2015.

[15] R Core Team. R: A language and environment for statistical computing. Dispoñible en https://www.R-project.org/, 2020.

[16] Sanitas. Información sobre a hormona luteinizante. Dispoñible en https://www.sanitas.es/sanitas/seguros/es/particulares/biblioteca-de-salud/prevencion-salud/hormona-luteinizante.html, 2020. Accedido 28-05-2021.

[17] R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.