

به نام خدا

شماره دانشجویی : ۴۰۲۱۰۶۶۰۴

محمد رضا منعمیان



Mohammad Reza Monemian

bioinformatic



سوال اول :

برای حل این سوال از آزمون فیشر یا همان آزمون دقیق جایگشت استفاده می‌کنیم. ابتدا به بیان نحوه خود آزمون فیشر می‌پردازیم. این آزمون برای بررسی استقلال یا عدم استقلال و وابستگی دو موضوع به همدیگر است. طبق آزمایش دو متغیر تصادفی A , B را که برنولی هستند حالت های مختلفشان را به صورت جدول زیر در می‌آوریم

متغیر تصادفی A خیر	متغیر تصادفی A آری	
b	a	متغیر تصادفی B آری
d	c	متغیر تصادفی B خیر

حال مقدار فرمول زیر را برای جدول بالا و همچنین جدول های اکستریم تر (یعنی جدول هایی که در آن a , d را زیاد کنیم) به دست می‌آوریم :

$$P = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}$$

p-value برابر با جمع مقادیر به دست آمده است. در ادامه برای رد کردن فرض صفر باید به مقدار سطح معناداری توجه کنیم اگر p-value به دست آمده از سطح معناداری کمتر بود این امکان را داریم که فرض صفر را رد کنیم. حال برای سوال گفته شده آزمون فیشر را به روشی که بیان کردیم انجام می‌دهیم. دو متغیر تصادفی ما در اینجا بیان افزایشی و فراوانی وقوع موتیف ۶ تایی ACGTGC است.

در اینجا فرض صفر و جایگزین ما به صورت زیر است :

H0 : ارتباط خاصی بین فراوانی وقوع موتیف ۶ تایی ACGTGC با بیان افزایشی وجود ندارد.

HA : خاصی بین فراوانی وقوع موتیف ۶ تایی ACGTGC با بیان افزایشی ارتباط معناداری است.

اگر کمی دقت کنیم در ژن های A,C,E موتیف ۶ تایی گفته شده وجود دارد که در بین آنها A ژن با بیان افزایشی نیست در صورتی که دوتای دیگر هستند. بنابراین جدول گفته شده به صورت زیر به دست می‌آید:

عدم داشتن موتیف	موتیف دار بودن	
b = 1	a = 2	بیان افزایشی
d = 2	c = 1	عدم بیان افزایشی

همچنین جدول اکستریم تر یعنی a و d سه شوند. در ادامه به محاسبه p-value می‌پردازیم :

$$P - value = \frac{\binom{2+1}{2} \binom{1+2}{1}}{\binom{1+2+2+1}{2+1}} + \frac{\binom{3+0}{3} \binom{0+3}{0}}{\binom{3+0+0+3}{3+0}} = 0.5$$

مقدار p-value به دست آمده برابر با ۰.۵ شد که فاصله بسیار زیادی با سطح معناداری ۰.۰۵ دارد و این یعنی با داشتن فرض درست بودن فرض H0 احتمال ۵۰ درصد ما به صورت شانس می‌توانیم نتایج به دست آمده را



مشاهده کنیم پس امکان رد کردن فرض صفر وجود ندارد و این نتیجه می‌دهد که فراوانی وقوع موتیف ۶ تایی گفته شده در سوال ارتباط خاصی با بیان افزایشی ژن‌ها ندارد و نسبتاً مستقل هستند.

سوال دوم)

قسمت آ: p-value مقداری از خطای نوع اول ما است یعنی با شرط درست بودن فرض صفر چقدر احتمال دارد ما فرض صفر را رد کنیم اگر این مقدار از سطح اطمینانی که ما می‌گذاریم بیشتر باشد به این معناست که احتمال اینکه اتفاق به وجود آمده به خاطر شانس باشد و ما دچار اشتباه مثبت کاذب شده باشیم زیاد است. حال اگر به جای یک آزمایش چندین آزمایش را با همدیگر انجام بدهیم اگر فرض کنیم سطح اطمینان ما ۰.۰۵ هم باشد یعنی ما ریسک خطای ۵ درصد را پذیرفتیم حال اگر ۶ آزمایش انجام دهیم احتمال اینکه یکی از آزمایش‌ها دچار خطای مثبت کاذب باشد برابر با :

$$1 - (0.95)^6 = 0.26$$

می‌بینیم که خطا به نسبت بسیار بزرگی افزایش یافت این به خاطر این است که تعداد آزمایش‌های ما زیاد است. بنابراین اگر سطح اطمینان را همان مقدار قبلی بگیریم به احتمال زیاد دچار خطا خواهیم شد به خاطر تجمع آزمایش‌ها روی هم. پس استفاده از همان p-value های خام نامناسب است و باید به روشی این مقدار را adjust کنیم.

قسمت ب :

روش بنفورونی benforoni روشی است که در آن می‌خواهیم از ایجاد یک خطای مثبت کاذب نیز تا حد امکان جلوگیری کنیم. برای این کار مقدارهای p-value ای که در آزمایش به دست آوردیم را در m یعنی تعداد آزمایش‌ها ضرب می‌کنیم (البته با ۱ نیز ماکسیمم می‌گیریم) و حالا مقدار به دست آمده را با سطح اطمینان قبلی یعنی ۰.۰۵ مقایسه می‌کنیم. مقدارهای p-value adjusted به صورت زیر است :

ژن	p-value adjusted	مقایسه با ۰.۰۵	معنی داری
A	0.024	کمتر	بله
B	0.066	بیشتر	خیر
C	0.126	بیشتر	خیر
D	0.234	بیشتر	خیر
E	0.486	بیشتر	خیر
F	1	بیشتر	خیر

همانطور که مشاهده می‌کنیم تنها معنی داری ژن A مورد تایید قرار گرفت.



قسمت ج:

در روش benjamine-hachberg آسانگیرانه تر عمل می‌کنیم. به این معنی که می‌خواهیم از میان همه آزمایش‌ها حداکثر به اندازه سطح معناداری قبلی یعنی ۰.۰۵ خطا داشته باشیم بدین معنی که اگر ۲۰ آزمایش داشته باشیم اینکه ۱ خطا داشته باشیم نیز مورد تایید ما است.

در این روش مقادیر p-value را به صورت صعودی مرتب می‌کنیم و با $\alpha \frac{k}{m}$ مقایسه می‌کنیم و از جایی که مقدار خام ما از این مقدار بیشتر شد از آنجا به بعد را عدم معناداری گزارش می‌کنیم.

ژن	p-value	$\alpha \frac{k}{m}$	مقایسه با ۰.۰۵	معنی داری
A	0.004	0.008	کمتر	بله
B	0.011	0.016	کمتر	بله
C	0.021	0.024	کمتر	بله
D	0.039	0.033	بیشتر	خیر
E	0.081	0.042	بیشتر	خیر
F	0.25	0.050	بیشتر	خیر

مشاهده می‌کنیم در این حالت سه ژن A,B,C معنا دار گزارش می‌شوند.

قسمت د:

چون روش بنفرونی بسیار محافظه‌کارانه است و فرض می‌کند تمام فرضیات مستقل هستند. اما روش BH اجازه می‌دهد درصد کمی از کشفیات، خطا باشند تا قدرت آماری برای پیدا کردن تفاوت‌های واقعی ژن‌های B و C حفظ شود. همانطور که گفتیم روش بنفرونی می‌خواهد جلوی حتی یک خطای مثبت کاذب را نیز بگیرد اما روش BH اجازه می‌دهد به اندازه سطح اطمینانی که در نظر گرفتیم از کل آزمایش‌ها دچار خطا شویم.

بنفرونی رخ خطای خانواده یا **Family-Wise Error Rate (FWER)** را کنترل می‌کند. یعنی تلاش می‌کند احتمال وقوع حتی «یک» مثبت کاذب را به زیر سطح اطمینان برساند.

BH : نرخ کشف اشتباه یا **False Discovery Rate (FDR)** را کنترل می‌کند. یعنی تضمین می‌کند که از بین ژن‌هایی که معنادار اعلام شده‌اند، به طور متوسط حداکثر به اندازه سطح اطمینان آن‌ها اشتباه باشند.

سوال سوم و چهارم به خاطر درخواست تیم تدریس به صورت دست نویس نوشته شده است.