دانشکده مهندسی کامپیوتر
دانشگاه صنعتی شریف

ژنومیک محاسباتی

مطهری و فروغمند
پاییز ۱۴۰۰

امتحان پایان ترم

پاسخ سوال های امتحان پایان ترم

نگارنده: الهه بدلی - ۹۸۲۰۹۰۷۲

۱ پاسخ سوال ۱

شرایط گفته شده سوال، شرایط تعادل هاردی_ وینبرگ است که جمعیت زیاد است. اگر دیگر شرایط این تعادل از جمله تصادفی بودن ازدواج و نبودن نیروهای انتخاب طبیعی و اینکه جهش نداشته باشیم و نسل‌ها قاطی نباشند یعنی ازدواج در هر نسل بین افراد همان نسل صورت گیرد. در این صورت طبق تعادل هاردی_ وینبرگ می‌توان گفت بعد از یک نسل به تعادل می‌رسد و در همان تعادل باقی می‌ماند.

$$P[AA] = (f_۱ + \frac{f_۲}{۲}) = f_A^۲ = p^۲$$

$$P[aa] = (f_۳ + \frac{f_۲}{۲}) = f_a^۲ = q^۲$$

$$P[aA] = ۱ - f_A^۲ - f_a^۲ = ۲(f_۱ + \frac{f_۲}{۲})(f_۳ + \frac{f_۲}{۲}) = ۲pq$$

۲ پاسخ سوال ۲

با در نظر V_i به عنوان نود i ام، V_i توزیع دو جمله‌ای دارد.

$$P[V_i = k] = \binom{2N}{k}(\frac{1}{2N})^k(1 - \frac{1}{2N})^{2N-k}$$

$$E[V_i|X_0 = 1] = 2Np = 2N * \frac{1}{2N} = 1$$

یعنی امید داریم در نسل بعد ۱ فرزند ببینیم.

$$Var[V_i] = 2Np(1-p) = 2N(\frac{1}{2N})(1 - \frac{1}{2N}) = 1 - \frac{1}{2N}$$

$$Corr(V_i, V_j) = \frac{Cov(V_i, V_j)}{\sqrt{Var[V_i]Var[V_j]}}$$

$$Cov(V_i, V_j) = E[V_i, V_j] - E[V_i]E[V_j]$$

$$= E[\sum_{i=1}^{2N} V_i \sum_{i=1}^{2N} V_j] - 1 * 1$$

$$= \sum_{i \neq j} E[V_i]E[V_j] + \sum_{i=j} E[V_j, V_j] - 1$$

i و j را دو فرد متفاوت در نظر می‌گیریم. بنابراین جمع دوم صفر است.

$$= \frac{(2N)^2 - 2N}{(2N)^2} - 1 = -\frac{1}{2N}$$

بنابراین برای ضریب هم‌بستگی داریم:

$$Corr(V_i, V_j) = \frac{-\frac{1}{2N}}{\sqrt{(1 - \frac{1}{2N})^2}} = -\frac{1}{2N - 1}$$

قسمت دوم:

در این صورت والدها از بین N نفر انتخاب می‌شوند. بنابراین احتمال والد مشترک برابر است با : $\frac{1}{N}$. بنابراین داریم:

$$P[V_i = k] = \binom{N}{k}(\frac{1}{N})^2(1 - \frac{1}{N})$$

$$E[V_i] = N * p = 1 \; Var[V_i] = Np(1-p) = \frac{N}{N}(1 - \frac{1}{N}) = 1 - \frac{1}{2N}$$

بنابراین برای هم‌بستگی داریم:

$$Corr(V_i, V_j) = \frac{Cov(V_i, V_j)}{\sqrt{(1 - \frac{1}{N})^2}} = \frac{-\frac{1}{N}}{1 - \frac{1}{N}} = -\frac{1}{N - 1}$$

۳ پاسخ سوال ۳

قسمت اول: k برگ داریم: بدون داشتن هیچ شرطی محاسبات را انجام می‌دهیم. هر یال درخت $coalsence$ یک توزیع نمایی با پارامتر متناسب با تعداد افراد آن مرحله است.بنابراین:

$$E[\,] = \frac{1}{\binom{k}{2}} + \frac{1}{\binom{k-1}{2}} + ... + \frac{1}{\binom{2}{2}} = \frac{2}{k(k-1)} + \frac{2}{(k-1)(k-2)} + ... + 1 = \frac{2}{k-1} - \frac{2}{k} + \frac{2}{k-2} - \frac{2}{k-1} + ... + \frac{2}{2} = 2 - \frac{2}{k} = \frac{2k-2}{k}$$

که به صورت حدی برابر ۲ است. یعنی ۲ برابر جمعیت . این عدد برای $k = ۲$ این عدد برابر ۱ است.
قسمت دوم:
در صورتی که ۲ فرد از نصف والدها حق انتخاب برای جد مشترک داشته باشند، آن گاه: احتمال جد مشترک : $\frac{۱}{N}$ می‌شود.
بنابراین:

$$p = \frac{\binom{k}{۲}}{N}, \ = a = \frac{p}{M}a = \frac{\binom{k}{۲}}{NM}$$

در صورتی که فرض کنیم $M = \frac{۱}{N}$ محاسبات مشابه قبل خواهد بود.

۴ سوال ۴

قسمت ۱:
در مدل WF اگر در نسل اول فرکانس A برابر $\frac{x}{۲N}$ باشد، وقتی در زمان پیش می‌رویم $fixation$ اتفاق می‌افتد.
اثبات: با فرض $X_۰ = \frac{x}{۲N}$ آن‌گاه فرکانس الیل A در نسل بعد از توزیع دو جمله‌ای است:

$$P[X_۱ = k|X_۰ = \frac{x}{۲N}] = \binom{۲N}{k}(\frac{x}{۲N})^۲(۱ - \frac{x}{۲N})^{۲N-k}$$

$$E[X_۱|X_۰ = \frac{x}{۲N}] = \frac{x}{۲N}$$

یعنی انتظار داریم همین نسبت در مراحل بعدی حفظ شود. همچنین این تعریف، یک مارتینگل است. $E[X_n|X_۰ = \frac{x}{۲N}] = \frac{x}{۲N}$
حال $fixation$ یا قبل از لحظه‌ی n ام اتفاق افتاده است یا بعد از آن لحظه. بنابراین امید را تفکیک می‌کنیم.

$$E[X_n|X_۰ = \frac{x}{۲N}, fixation\, before\, n]P[fixation\, before\, n] + E[X_n|X_۰ = \frac{x}{۲N}, fixation\, after\, n]P[fixation\, after\, n]$$

که اگر n زیاد باشد، $fixation$ قبل از n رخ داده است. بنابراین:

$$\frac{x}{۲N} = E[X_n|X_۰ = \frac{x}{۲N}, fixation\ before\ n]$$

حال $fixation$ یا برای a اتفاق افتاده است که امید آن صفر است و یا برای A اتفاق افتاده است که امید آن برابر ۱ است. بنابراین :

$$P[fixation\ is\ A] = \frac{x}{۲N}$$

بنابراین احتمال $fixation$ برای هر الیل برابر فرکانس آن در نسل اول است.
قسمت ۲:
فرض کنیم U و V دو متغیر تصادفی با توزیع نمایی به ترتیب با پارامترهای $\lambda_۱$ و $\lambda_۲$ است. بنابراین:

$$P[U > t] = e^{-\lambda_۱ t}$$

و

$$P[U > t] = e^{-\lambda_۱ t}$$

حال :

$$Z = min(U, V)$$

$$P[min(U, V) > t] = P[U > t, V > t]$$

U و V مستقل‌اند. بنابراین:

$$= P[U > t]P[V > t] = e^{(\lambda_1+\lambda_2)t}$$

قسمت ۳:

از منظر ۲: با فرض اینکه جهش با نرخ θ رخ بدهد، آن‌گاه تعداد جهش نمایی با پارامتر $k\theta$ است.
در این صورت داریم:

$$P[S_n = s] = P[S_n = s|mutation\ first]P[mutation\ first] + P[S_n = s|coalsence\ first]P[coalsence\ first] =$$

$$P[S_n = s-1].\frac{k\theta}{\binom{k}{2}+k\theta} + P[S_{n-1} = s].\frac{\binom{k}{2}}{\binom{k}{2}+k\theta}$$

البته در منابع کلاس نرخ جهش، $\frac{\theta}{2}$ در نظر گرفته شده است.

قسمت ۴:

معمولا مدل‌هایی مثل مدل WF فرض‌هایی دارند که واقعی نیستند. مثلا تعداد جمعیت همیشه ثابت است یا ازدواج تصادفی است یا مثلا تعداد مرد و زن یکسان است. اما در واقعیت اینگونه نیست. در واقعیت جمعیت موثر مطرح می‌شود که از جمعیت اصلی کوچک‌تر است و با تقریب خوبی مدل‌هایی که قبلا داشتیم را می‌توان استفاده کرد. برای مثال می‌توان از روش یافتن جمعیت موثر بر پایه واریانس استفاده کرد. در مدل WF ، با فرض X تعداد A ها در این نسل، داریم:

$$P[X = k] = \binom{2N}{k} p^k(1-p)^{2N-k}$$

$$E[X] = 2Np$$

$$Var[X] = 2Np(1-p)$$

$$p' = \frac{X}{2N}$$

$$E[p'] = \frac{E[X]}{2N} = \frac{2Np}{2N} = p$$

$$Var[p'] = \frac{Var[X]}{(2N)^2} = \frac{2Np(1-p)}{(2N)^2} = \frac{p(1-p)}{2N}$$

بنابراین:

$$2N_{eff} = \frac{p(1-p)}{V\hat{[p']}}$$

بنابراین با داشتن تخمینی از واریانس می‌توان N_{eff} را محاسبه نمود.

۵ سوال ۵:

سوال۱:

تخمین واترسون: برای تخمین زدن نرخ جهش استفاده می‌شود. با حل از منظر ۱:
تعداد واقعی جهش‌ها روی درخت به ازای n نفر را S_n در نظر می‌گیریم.

$$P[S_n = s] = \frac{n-1}{\theta}\sum_{i=1}^{n-1}(-1)^{(i-1)}\binom{n-2}{i-1}\left(\frac{\theta}{i+\theta}\right)^{s+1}$$

با به دست آوردن $E[S_n]$ می‌توان تخمین واترسون را به دست آورد.
در صورتی که تعداد جهش‌ها در شاخه‌ی n ام را X_n در نظر بگیریم.

$$S_n = X_n + X_{n-1} + ... + X_2$$

تعداد جهش‌ها در هر شاخه از توزیع پواسون هستند.

$$E[S_n] = E[X_n] + E[X_{n-1} + ... + E[X_2]$$

و داریم:

$$E[S_k] = E_t[E[S_k|t]]$$

با توجه به اینکه kt تا فاصله روی درخت داریم و تعداد جهش‌ها دارای توزیع پواسون هستند، بنابراین :

$$E[S_k|t]] = 2Nkt$$

بنابراین:

$$E[S_k] = 2NkE[t] = \frac{2Nk}{\binom{k}{2}} = \frac{4N}{k-1}$$

زیرا با فرض $2Nx = \frac{\theta}{2}$ و با فرض $x = 1$ داریم: $\theta = 4N$. بنابراین:

$$\frac{4N}{k-1} = \frac{\theta}{k-1}$$

حال با جمع زدن برای n های مختلف داریم:

$$\frac{\theta}{n-1} + \frac{\theta}{n-2} + ... + \frac{\theta}{1}$$

که برابر است با :

$$E[S_k] = \theta \sum_{i=1}^{n-1} \frac{1}{i} = \theta h_n$$

که در آن: $h_n = logn$

سوال ۲:

در روش $TajimaD$ ابتدا فاصله جفت جفت رشته‌ها را به دست می‌آوریم. فرض می‌کنیم d_{ij} فاصله ۲ رشته است. این فاصله با تعداد جهش‌های روی ۲ شاخه برابر است.

حال با فرض وجود k نفر، داریم:

احتمال اینکه اول جهش داشته باشیم:

$$\frac{\frac{\theta k}{2}}{\frac{\theta k}{2} + \binom{k}{2}}$$

برای دو نفر داریم: احتمال اینکه اول جهش داشته باشیم: $\frac{\theta}{\theta+1}$ و احتمال اینکه اول $coalsence$ برابر $\frac{1}{\theta+1}$ است.

حال فاصله این دو نفر را این گونه در نظر می‌گیریم که چند جهش تا قبل از $coalsence$ دارند.

$$P[d_{ij} = k] = (\frac{\theta}{\theta+1})^k \frac{1}{1+\theta}$$

یعنی k بار جهش رخ داده است و در نهایت $coalsence$ انجام شده است.

$$E[d_{ij}] = \frac{1-p}{p} \quad , k = 0, 1, 2, ...$$

$$E[d_{ij}] = \frac{۱ - \frac{۱}{۱+\theta}}{\frac{۱}{۱+\theta}} = \frac{\frac{\theta}{\theta+۱}}{\frac{۱}{۱+\theta}} = \theta$$

و برای تخمین $TajimaD$ نیز داریم:

$$E[\hat{\theta_T}] = \frac{\sum_{i,j} E[d_{ij}]}{\binom{n}{۲}} = \theta$$

سوال ۳:

استفاده از $TajimaD$ به این صورت است که در حالتی که جهش ها خنثی باشند باید نرخ جهش تخمین زده شده توسط $Waterson$ و $TajimaD$ یکی باشند. بنابراین باید اختلافشان صفر باشد و D در این نام هم به همین معنا است.