آمار و احتمال مهندسی

نیم‌سال اول ۹۸ـ۹۹

دانشکده‌ی مهندسی کامپیوتر

کوییز شماره ۳

مسئله‌ی ۱. توزیع Pareto

توزیع Pareto با دو پارامتر α و x_m مشخص می‌شود. می‌دانیم که x_1، x_2، ... و x_n داده‌هایی رندوم از توزیع Pareto با $\alpha > 2$ هستند. خصوصیات توزیع Pareto داده‌هایمان در ادامه آورده شده است:

$$PDF: \ \frac{\alpha {x_m}^{\alpha}}{x^{\alpha+1}}, \ \alpha > 2, \ x_m > 0, \ x \geqslant x_m$$

$$CDF: \ 1 - (\frac{x_m}{x})^{\alpha}$$

$$Mean: \ \frac{\alpha x_m}{\alpha - 1}$$

$$Variance: \ \frac{\alpha {x_m}^2}{(\alpha-1)^2(\alpha-2)}$$

الف

با استفاده از تخمین‌گر MLE پارامترهای α و x_m را تخمین بزنید.

ب

وضعیت unbiased بودن و consistent بودن تخمین‌گر MLE برای x_m را مشخص کنید.
(راهنمایی: اگر n متغیر تصادفی iid از توزیعی با $CDF: \ F$ باشند، مینیموم این متغیرها متغیر تصادفی‌ای با $CDF: \ 1-(1-F)^n$ است)

حل.

الف

ابتدا x_m را تخمین می‌زنیم. باید عبارت زیر را بیشینه کنیم:

$$\sum_{i=1}^{n} ln(P(x_i|x_m)) = \sum_{i=1}^{n} ln(\frac{\alpha {x_m}^{\alpha}}{{x_i}^{\alpha+1}}) = nln(\alpha) + n\alpha ln(x_m) - (\alpha+1)(\sum_{i=1}^{n} ln(x_i))$$

برای بیشینه کردن این عبارت بر حسب x_m کافی است x_m را بیشینه کنیم. برای هر i می‌دانیم $x_i \geqslant x_m$. پس $min(x_۱, ..., x_n) \geqslant x_m$. پس تخمین‌مان برای x_m برابر $min(x_۱, ..., x_n)$ است.

حال α را تخمین می‌زنیم. باید عبارت زیر را بیشینه کنیم:

$$\sum_{i=۱}^{n} ln(P(x_i|\alpha)) = nln(\alpha) + n\alpha ln(x_m) - (\alpha + ۱)(\sum_{i=۱}^{n} ln(x_i))$$

مشتق این عبارت بر حسب α را برابر ۰ قرار می‌دهیم. در این صورت داریم:

$$\alpha = \frac{n}{\sum_{i=۱}^{n} ln(x_i) - nln(x_m)}$$

ب

CDF تخمین‌گر $min(x_۱, ..., x_n)$ طبق راهنمایی برابر است با $۱ - (\frac{x_m}{x})^{\alpha n}$. همان‌طور که می‌بینید این عبارت CDF یک توزیع Pareto با پارامترهای $(\alpha n, x_m)$ است. پس MSE این تخمین‌گر برابر است با:

$$(\frac{\alpha n x_m}{\alpha n - ۱} - x_m)^۲ + \frac{\alpha n {x_m}^۲}{(\alpha n - ۱)^۲(\alpha n - ۲)} =$$

$$(\frac{x_m}{\alpha n - ۱})^۲ + \frac{\alpha n {x_m}^۲}{(\alpha n - ۱)^۲(\alpha n - ۲)}$$

حال اگر n را به بی‌نهایت میل دهیم، MSE به ۰ میل می‌کند. پس این تخمین‌گر consistent است. اما چون داریم $bias = \frac{x_m}{\alpha n - ۱}$ پس unbiased است. $\triangleright$

مسئله‌ی ۲.

جدول t در ادامه آمده است.

اداره هواشناسی یک شهر، ۴ دستگاه سنجش آلودگی هوا را در یک منطقه قرار داده است. فرض کنید شاخص آلودگی هوا در این منطقه ثابت است اما این دستگاه‌ها دقیق نیستند و شاخص را با کمی نویز گزارش می‌دهند. در یک روز نسبتا آلوده، مقادیر گزارش شده توسط این ۴ دستگاه به شرح زیر است.

$$۱۵۲, ۱۴۸, ۱۵۳, ۱۵۳$$

الف

با کمک داده‌های جمع آوری شده، یک بازه اطمینان ۹۵ درصد برای شاخص آلودگی هوا در آن ایستگاه ارائه دهید.

ب

در صورتی که شاخص آلودگی هوا از ۱۵۰ بیشتر باشد، هوا در شرایط ناسالم برای تمامی گروه‌ها قرار می‌گیرد. عده‌ای از دانشمندان معتقدند که میانگین شاخص آلودگی ۱۵۰ بوده بنابراین هوای این منطقه ناسالم نیست، در حالی که عده‌ی دیگری معتقد میانگین شاخص آلودگی به طور معنی‌داری از ۱۵۰ بیشتر بوده و هوا ناسالم است. برای بررسی این افراد یک آزمون فرض طراحی کنید. فرض صفر و فرض دیگر این آزمون را بیان کرده و سپس مشخص کنید آیا با سطح اهمیت ۰/۰۵ می‌توان فرض صفر را رد کرد یا خیر.

پ

برای کاهش خطای نوع اول، باید سطح اهمیت را افزایش دهیم یا کاهش؟ برای کاهش خطای نوع دوم چه طور؟

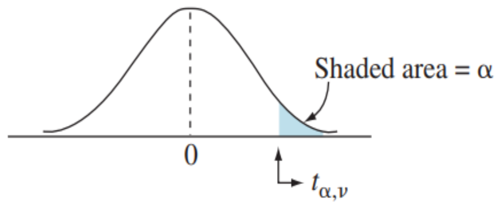

TABLE 2
Percentage points of Student's *t* distribution

df/α =	.40	.25	.10	.05	.025	.01	.005	.001	.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073

حل.

الف

با توجه به اینکه توزیع عدد گزارش شده توسط هر دستگاه* توزیع نرمالی است که واریانس آن را نداریم، می‌توانیم توزیع میانگین را Student's t در نظر بگیریم. در این صورت برای بازه اطمینان داریم

$$-t_{/۰۲۵} \leqslant \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \leqslant t_{/۰۲۵}$$

از روی داده‌ها داریم $s = ۲/۳۸$ $\bar{x} = ۱۵۱/۵$ و با توجه به اینکه درجه آزادی در اینجا ۳ است از روی جدول t داریم $t_{/۰۲۵} = ۳/۱۸$. با جایگذاری مقادیر بازه اطمینان [۱۴۷/۷, ۱۵۵/۳] به دست می‌آید.

ب

از آزمون تک نمونه‌ای t-test با فرض صفر $\mu = ۱۵۰$ و فرض دیگر $\mu > ۱۵۰$ استفاده می‌کنیم. t-value داده‌ها برابر است با $\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = ۱/۲۶$ که از $t_{۰/۰۵} = ۲/۳۵$ کمتر است بنابراین نمی‌توانیم فرض صفر را رد کنیم.

پ

خطای نوع اول برابر با سطح اهمیت آزمون است بنابراین برای کاهش خطای نوع اول باید مقدار عددی سطح اهمیت را کاهش دهیم. خطای نوع دوم هنگامی اتفاق می‌افتد که فرض صفر غلط باشد و ما به اشتباه نتوانیم آن را رد کنیم. افزایش مقدار عددی سطح اهمیت رد کردن فرض صفر را ساده کرده و احتمال آن را بیشتر می‌کند (در حالی که تاثیری روی درست یا غلط بودن فرض صفر در عالم واقعیت ندارد!) بنابراین افزایش مقدار عددی سطح اهمیت باعث کاهش احتمال خطای نوع دوم می‌شود. $\triangleright$

موفق باشید