پانزدهمین سمینار احتمال
و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان

15th Seminar On Probability
and Stochastic Processes
August 30-31, 2025
University of Kurdistan

انجمن آمار ایران

برخی نتایج مجانبی در مورد اختلاف میانگین دو جامعه در داده‌های تابعی جزئی مشاهده‌شده

پویا سلیمی فر [۱] ، سید محمد ابراهیم حسینی نسب[۲] ، امید خادم نوع [۳]

[۱] دانشگاه شهید بهشتی

[۲] دانشگاه شهید بهشتی [۳] دانشگاه زنجان

چکیده: امروزه ثبت اطلاعات از مشاهداتی مانند $X_1, X_2, \ldots, X_n$ که به‌صورت تابعی پیوسته از یک پدیده باشند، مرسوم است. مشاهدات تابعی اندازه‌گیری‌هایی از پدیده‌ای است که نسبت به یک متغیر مانند زمان، بسامد و . . . پیوسته است. ماهیت پیوسته این داده‌ها منبع غنی از اطلاعات است و امکان مدلسازی با دقت بالا را فراهم می‌کند، در برخی شرایط و مطالعات تجربی عواملی مانند محدودیت زمانی در مطالعه، ابزارهای اندازه‌گیری ناکارامد یا عوامل ناشناخته سبب می‌شوند ثبت داده‌ها در سراسر دامنه‌ای معین در جریان مطالعه به‌صورت کامل انجام نگیرد، در این وضعیت داده‌های تابعی به‌صورت ناکامل یا جزئی مشاهده‌شده [۱] در دسترس هستند و روش‌های برآوردیابی معمول که برای داده‌های تابعی کامل ارائه شده است، قابل استفاده نیست. یک رویکرد این است که با در نظر گرفتن مقدار تابع در زیر مجموعه‌های مشاهده‌شده از کل دامنه به تحلیل داده‌ها پرداخته می‌شود. در این مقاله، برخی خواص مجانبی از اختلاف میانگین دو جامعه در داده‌های تابعی جزئی مشاهده‌شده ارائه می‌شود.

واژه‌های کلیدی: داده‌های تابعی جزئی مشاهده‌شده، همگرایی در احتمال، فرآیند گاوسی، عملگر قطری .

۱ مقدمه

داده‌های تابعی جزئی مشاهده شده شامل تحقق‌هایی از توابع تصادفی هستند که در کل دامنه مشاهده نشده و هر تابع در نمونه ممکن است در زیر مجموعه‌های متفاوتی از دامنه مشاهده شود و هیچ اطلاعی درباره مقادیر تابع در سایر نقاط دامنه در دسترس نباشد. با در نظر گرفتن $\tau = [0, 1]$ برای i امین متغیر تابعی $X_i \in L^2[0, 1]$ زیر مجموعه‌ای مانند $O_i \subseteq [0, 1]$ وجود دارد به طوری که $X_i(t)$ برای $t \in O_i$ مشاهده شده است و برای $t \in \tau \setminus O_i$ مشاهده نشده است، به طور کلی الگوی گمشدگی در مطالعات تابعی نیز می‌تواند

[۱] سخنران، p_salimifar@sbu.ac.ir

[1] Partially observed

در چهار حالت نمایان شود. اولین حالت که به آن اشاره می‌شود، الگوی گمشدگی کاملا تصادفی [۲] است که در آن گمشدگی بخش‌هایی از تابع، مستقل از تابع (به این معنا که در آن ساختار گمشدگی نه به بخش‌های مشاهده شده و نیز به بخش‌های گمشده از تابع وابسته نیست) است. دومین حالت، الگوی گمشدگی تصادفی نام دارد که در آن گمشدگی تنها با بخش مشاهده شده از تابع مرتبط است و سومین الگوی گمشدگی، گمشدگی غیر تصادفی نام دارد که به هر دو بخش مشاهده شده و گمشده از تابع وابسته خواهد بود. الگوی چهارم که الگوی گمشدگی سیستماتیک [۳] است از سه وضعیت موجود متفاوت است و در آن اغلب یک عامل شناخته شده در جریان مطالعه سبب گمشدگی بخشی از تابع در دامنه می‌شود، همچنین یادآور می‌شویم که تاکنون الگوی گمشدگی تصادفی و غیر تصادفی به دلیل فرض‌های بسیار محدود کننده بر الگوی داده‌ها در پژوهش‌های مربوط به داده‌های تابعی مورد مطالعه قرار نگرفته‌اند. مسائل کاربردی مرتبط با این حالت باعث ایجاد یک مجموعه کارهای تحقیقاتی با جنبه‌های مختلف در این مسئله شده است، باگنی (۲۰۱۲)، به آزمون مشخصه‌سازی برای داده‌های تابعی ناکامل (که به ارزیابی اعتبار فرضیات احتمالی تشکیل دهنده یک مدل آماری اشاره دارد) پرداخت، کراووس (۲۰۱۵)، داده‌های پایش فشار خون سرپایی و نمایه ضربان قلب شامل الگو تغییرات ضربان قلب در حالات مختلف فعالیت، استراحت و یا شرایط خاص پزشکی را مورد مطالعه قرار داد و با استفاده از قسمت مشاهده شده از منحنی ضربان قلب برای قسمت گمشده از آن یک پیشگو ارائه کرد، سپس برای قسمت گمشده از منحنی یک باند پیشگویی ساخت. دلایگل و هال (۲۰۱۶)، داده‌های تابعی ناکامل را با استفاده از بخش‌هایی از یک مدل زنجیر مارکوف تقریب زدند. گرومنکو و همکاران (۲۰۱۷)، در یک مطالعه به بررسی روند سرمایش در لایه یونوسفر با استفاده از رگرسیون تابعی با منحنی‌های ناکامل پرداختند. شیبانی و همکاران (۲۰۱۸)، به طبقه‌بندی با متغیرهای کمکی تابعی ناقص پرداختند. اِستِفنسکی و همکاران (۲۰۱۸)، به تفکیک داده‌های تابعی جزئی مشاهده شده بر اساس تحلیل مولفه‌های اصلی پرداختند. دیسکری و همکاران (۲۰۱۹)، به بازسازی کوواریانس از مشاهدات تابعی ناکامل پرداختند. لیبل و رامسدر (۲۰۱۹)، در یک مطالعه که هدف آن، برآورد منحنی میانگین قیمت مصرف انرژی در بازار ذخیره کنترلی برق در آلمان بود، تمرکز کردند و با توجه به این که در این بازار مزایده‌ای، منحنی‌های قیمت تنها به صورت جزئی در انتها دامنه مشاهده می‌شدند مکانیسم گمشدگی به‌صورت سیتماتیک، وابسته به استراتژی های معاملاتی (عامل شناخته شده) بود و این عامل سبب نقض فرض گمشدگی کاملاً تصادفی می‌شد و نتایج ناسازگاری را در برآوردگر میانگین ایجاد می‌کرد از این رو آنها از طریق قضیه اساسی حساب دیفرانسیل و انتگرال و مشتق پذیری داده‌های تابعی برآوردهای سازگاری برای تابع میانگین و عملگر کوواریانس در این وضعیت ارائه دادند. یان کیو و همکاران (۲۰۲۱)، به مطالعه‌ی آزمون فرض دو نمونه‌ای برای میانگین پرداختند. نایپ و لیبل (۲۰۲۰)، یک روش بازسازی برای قسمت گمشده از منحنی‌های جزئی مشاهده شده ارائه دادند. از دیگر مطالعات می‌توان به کتب منتشر شده توسط روبین (۱۹۷۶)، هی و همکاران (۲۰۲۲)، اشاره کرد.

با توجه به این که استنباط در مورد میانگین داده‌های تابعی جزئی مشاهده‌شده از اهمیت بالایی برخوردار است و در بسیاری از مطالعات شامل آزمون فرض مورد استفاده قرار می‌گیرد در این پژوهش نیز به مطالعه‌ی برخی خواص مجانبی در مورد اختلاف میانگین یک متغیر تابعی در دو جامعه پرداخته می‌شود. نتایج این پژوهش می‌تواند در آزمون فرض برای برابری میانگین دو جامعه و در حالت تعمیم یافته آن به چند جامعه مورد استفاده قرار گیرد.

[2] Missing-Completely-At-Random

[3] Systematic Missingness

۲ برخی نتایج حدی

در مطالعه‌ی داده‌های تابعی جزئی باید دو عامل مهم در نظر گرفته شود، نخست این که وزن های مربوط به شانس گمشدگی در هر گروه متفاوت است و باید در ساختار اختلاف میانگین‌ها اعمال شود، دوم این که ممکن است در اکثر مطالعات واقعی برای هر گروه تابع کوواریانس متمایزی وجود داشته باشد که هر دو این عوامل باید در ساختار مسئله اعمال شود. اکنون به تعریف برخی نماد مورد استفاده در مقاله می‌پردازیم:

$O_{ij}(t)$ معرف تابع نشانگر در لحظه t برای گروه j است. از این رو $N_j(t) = \sum_{i=1}^{n} O_{ij}(t)$ مجموع مشاهدات در دسترس برای $j = 1, 2$ است که در آن اندیس j نشانگر تعداد گروه‌ها است و برای هر گروه مشاهدات تابعی،می‌توانند متفاوت باشند. اگر $N_j(t) = 0$، آنگاه $J_j(t) = I_{[N_j(t))>0]} = 0$، به این معنا است که مقدار مشاهده شده‌ای از تابع برای گروه j ام در لحظه‌ی t وجود ندارد.

$$\begin{cases} \hat{\rho}(s,t) = \dfrac{I(s,t)}{M(s,t)} \displaystyle\sum_{i=1}^{n} U_i(s,t)\{X_i(s) - \hat{\mu}_{st}(s)\}\{X_i(t) - \hat{\mu}_{st}(t)\}, \\ I(s,t) = I_{[M(s,t)>0]}, \\ \hat{\mu}_{st}(s) = \dfrac{I_{[M(s,t)>0]}}{M(s,t)} \displaystyle\sum_{i=1}^{n} U_i(s,t) X_i(s). \end{cases}$$

$U_i(s,t) \equiv O_i(s)O_i(t)$ معرف تابع نشانگر است و به این معناست که در دو لحظه‌ی معین s و t مقدار تابع به‌صورت توأم در دسترس باشد ، از این رو $M(s,t) \equiv \sum_{i=1}^{n} U_i(s,t)$ معرف مجموع مشاهدات توام در هر گروه است و در صورتی که $I_{[M(s,t)>0]} = 0$ (به این معنا که هیچ مقادیری از تابع در نقاط جفتی مشاهده نشود) ، برآوردگر $\hat{\mu}_{st} = 0$ در نظر گرفته می‌شود، در ادامه برآورد میانگین گروهی و میانگین کل به‌صورت زیر است:

$$\sum_{j=1}^{2} \hat{\omega}_j(t) = 1, \quad N(t) = \sum_{i=1}^{n} O_i(t), \quad \hat{\mu}(t) = \sum_{j=1}^{2} \hat{\omega}_j(t)\hat{\mu}_j(t).$$

که در آن برآورد میانگین کل به صورت ترکیب خطی از برآورد میانگین‌های گروهی است و هر گروه با ضریب تاثیر $\hat{\omega}_i(t)$ در میانگین کل حضور دارند بعلاوه میانگین برای هر گروه و تابع وزن برای j-امین گروه به صورت زیر است:

$$\hat{\mu}_j(t) = J_j(t)N_j(t)^{-1} \sum_{i=1}^{n_j} O_{ji}(t)X_{ji}(t), \quad j \in \{1, 2\}, \quad \hat{\omega}_j(t) = \frac{\dfrac{N_j(t)}{\hat{r}_j^2}}{\displaystyle\sum_{k=1}^{K} \frac{N_k(t)}{\hat{r}_k^2}}.$$

بعلاوه احتمال این که در لحظه‌ی t مقدار تابع مشاهده شود و نیز احتمال این که در دو لحظه معین به‌صورت توام مقادیر تابع مشاهده شود به‌صورت زیر تعریف می‌شوند:

$$\pi_i(t) \equiv EO_i(t) = \Pr\{O_i(t) = 1\}, \quad \nu_i(s,t) = E[U_i(s,t)] = P[O_i(t) = 1, O_i(s) = 1].$$

و ضریب استاندارد ساز در $\hat{\omega}_j(t)$ نیز به‌صورت $\hat{r}_j = \int_0^1 \hat{\rho}_j(t,t)\, dt$ تعریف می‌شود.

نخست تعریف می‌کنیم:

$$SSR_n(t) = \sum_{i=1}^{2} n_i \left(\bar{y}_{i\cdot}(t) - \bar{y}_{\cdot\cdot}(t)\right)^2,$$

که همان تغییرات بین گروهی است که در تحلیل واریانس مورد استفاده قرار می‌گیرد. مشابه تعریف فوق مولفه‌ی تصادفی $Z_{n_j}(t)$ نیز مشابه جملات مجموع بالا و با تعدیل آن توسط دو عامل ذکر شده برای داده‌های تابعی جزئی مشاهده‌شده به‌صورت زیر ارائه می‌شود:

$$Z_{n_j}(t) = \hat{r}_j^{-1} N_j(t)^{1/2} \left[\hat{\mu}_j(t) - \hat{\mu}(t)\right].$$

اکنون برای سهولت در اعمال دو مولفه‌ی مهم یعنی، ساختار گمشدگی نامتقارن و تابع کوواریانس متمایز و دوری از پیچیدگی های غیر ضروری با تعریف یک عملگر قطری می‌توان $Z_{n_j}(t)$ را به‌صورت ضرب نقطه‌ای از عناصر $f : \mathcal{L}^2 \to \mathcal{L}^2$ ایجاد کرد که در آن برآورد عملگر قطری یعنی $\hat{\psi}_{jl}$ بر روی هر عضو دلخواه f_l متعلق به فضای L^2 به صورت زیر عمل می‌کند:

$$(\hat{\psi}_{jl} f_l)(t) = \hat{r}_j^{-1} \left\{\delta_{jl} - N_j(t)^{1/2} \hat{\omega}_l(t) J_l(t) N_l(t)^{-1/2}\right\} f_l(t)$$

بعلاوه تعریف می‌کنیم:

$$H_j(t) = N_j(t)^{1/2}[\hat{\mu}_j(t) - \mu(t)], \quad j = 1, 2$$

از این رو می‌توان $Z_{n_j}(t)$ را برای هر جامعه به‌صورت زیر بازنویسی کرد،

$$Z_{n_j}(t) = \hat{\psi}_{j1} H_1(t) + \hat{\psi}_{j2} H_2(t), \quad j = 1, 2 \tag{۱.۲}$$

اکنون عناصر $Z_{n_j}(t)$ به صورت زیرتعریف می‌شوند:

$$Z_{n_j}(t) = \hat{r}_j^{-1} N_j(t)^{1/2} \left[\hat{\mu}_j(t) - \hat{\mu}(t)\right] = \sum_{i=1}^{2} \hat{\psi}_{ji} H_i(t),$$

از این رو $Z_{n_j}(t)$ با اعمال عملگر قطری $\hat{\psi}_{jl}$ روی $H_l(t)$ به صورت زیر نمایش داده می‌شود:

$$(\hat{\psi}_{jl} H_l)(t) = \hat{r}_j^{-1} \left\{\delta_{jl} - N_j(t)^{1/2} \hat{\omega}_l(t) J_l(t) N_l(t)^{-1/2}\right\} N_l(t)^{1/2}[\hat{\mu}_l(t) - \mu(t)], \quad l = 1, 2$$

برای حالتی که $J_j(t) = \mathbb{I}_{\{N_j(t)>0\}} = 1$، به این معنا که در هر گروه مشاهده داشته باشیم، $Z_{n_j}(t)$ به‌صورت زیر ساده می‌شود:

$$Z_{n_j}(t) = \sum_{l=1}^{2} \hat{r}_j^{-1} \left\{\delta_{jl} N_l(t)^{1/2} - N_j(t)^{1/2} \hat{\omega}_l(t)\right\} [\hat{\mu}_l(t) - \mu(t)],$$

اما با استفاده از تعریف دلتای کرونکر، داریم:

$$\sum_{k=1}^{2} \hat{r}_j^{-1} \delta_{jl} N_l(t)^{1/2} [\hat{\mu}_l(t) - \mu(t)] = \hat{r}_j^{-1} N_j(t)^{1/2} [\hat{\mu}_j(t) - \mu(t)],$$

در ادامه مولفه‌ی $Z_{n_j}(t)$ به صورت مجموع دو مولفه‌ی تصادفی تجزیه می‌شود و به همگرایی هر دو جز به صورت کلی اشاره می‌شود برای این منظور دو بخش تصادفی را بر اساس انتقال میانگین‌ها بازنویسی کرده و به همگرایی دو جز تصادفی به صورت کلی اشاره می‌کنیم.

با فرض این که میانگین‌های هر گروه به‌صورت $\mu_j(t) = \mu(t) + h_j(t)$ نمایش داده شوند، $Z_{n_j}(t)$ بر اساس این انتقال به اندازه‌ی هر تابع حقیقی مقداری مانند $h_j(t)$ به‌صورت زیر بازنویسی می‌شود:

$$Z_{n_j}(t) = \sum_{l=1}^{2} \hat{r}_j^{-1} \left\{ \delta_{jl} - N_j(t)^{1/2} \hat{\omega}_l(t) N_l(t)^{-1/2} \right\} N_l(t)^{1/2} \left[h_l(t) + (\hat{\mu}_l(t) - \mu_l(t)) \right]. \tag{۲.۲}$$

همگرایی مولفه‌های تجزیه شده $Z_{n_j}(t)$

با توجه به تعریف $Z_{n_j}(t)$ در رابطه (۲.۲) می‌توان نوشت: $Z_{n_j}(t) = \Delta_{n_j}(t) + S_{n_j}(t)$ که در آن

$$\Delta_{n_j}(t) = \sum_{l=1}^{2} \hat{r}_j^{-1} \left\{ \delta_{jl} - N_j(t)^{1/2} \hat{\omega}_l(t) N_l(t)^{-1/2} \right\} N_l(t)^{1/2} h_l(t), \tag{۳.۲}$$

$$S_{n_j}(t) = \sum_{l=1}^{2} \hat{r}_j^{-1} \left\{ \delta_{jl} - N_j(t)^{1/2} \hat{\omega}_l(t) N_l(t)^{-1/2} \right\} N_l(t)^{1/2} \left[\hat{\mu}_l(t) - \mu_l(t) \right], \tag{۴.۲}$$

قضیه ۱.۲. با قرار دادن $j = 1, 2$ $h_j(t) = \frac{d_j(t)}{n_j^{1/2}}$، که در آن $d_j(t)$ هر تابع حقیقی مقدار است، می‌توان نشان داد که همگرایی در توزیع برای $\Delta_{n_j}(t)$ و $S_{n_j}(t)$ به صورت زیر نتیجه می شود:

$$\Delta_{n_j}(t) = \hat{r}_j^{-1} N_j(t)^{1/2} \left\{ n_j^{-1/2} d_j(t) - \sum_{l=1}^{2} \hat{\omega}_l(t) n_l^{-1/2} d_l(t) \right\} \xrightarrow{d} r_j^{-1} \pi_j(t)^{1/2} \left\{ d_j(t) - a_j^{1/2} \sum_{l=1}^{2} \omega_l(t) a_l^{-1/2} d_l(t) \right\}, \tag{۵.۲}$$

$$S_{n_j}(t) \xrightarrow{d} r_j^{-1} \left[S_j^{\infty}(t) - \sum_{l=1}^{2} \omega_l(t) \left[\frac{\pi_j(t) a_j}{\pi_l(t) a_l} \right]^{1/2} S_l^{\infty}(t) \right], \tag{۶.۲}$$

که در آن

$$a_j = \lim_{n \to \infty} \frac{n_j}{n}.$$

نسبت حدی اندازه نمونه هر گروه به اندازه نمونه کل گروه‌ها است، از این رو میانگین مجانبی فرآیند گاوسی با انتقال میانگین همان توزیع حدی بخش غیر تصادفی در (۵.۲) است و تابع کوواریانس آن مرتبط با بخش تصادفی در (۶.۲) است. که در آن $S_j^{\infty}(t) \stackrel{d}{=} \mathcal{GP}(0, \mathcal{K}_j)$ به صورت فرآیند گاوسی با میانگین صفر و هسته عملگر کوواریانس زیر است:

$$\kappa_j(s,t) = \left[\pi_j(t) \right]^{-1/2} \left[\pi_j(s) \right]^{-1/2} \nu_j(s,t) \rho_j(s,t).$$

در ادامه هسته مربوط به توزیع مجانبی عملگر $Z_{n_j}(t)$ به‌صورت زیر ارائه می‌شود:

$$
\begin{aligned}
\nu_{jk}^{\infty}(s,t) &= \operatorname{cov}\left[Z_j^{\infty}(t), Z_k^{\infty}(s)\right] \\
&= \sum_{l=1}^{2} \psi_{jl}(s)\kappa_l(s,t)\psi_{kl}^{*}(t) \\
&= \sum_{l=1}^{2} r_j^{-1}\left\{\delta_{jl} - \pi_j(s)^{1/2} a_j^{1/2}\omega_l(s)\pi_l(s)^{-1/2}a_l^{-1/2}\right\} \\
&\quad \times \kappa_l(s,t)\left\{\delta_{kl} - \pi_k(t)^{1/2} a_k^{1/2}\omega_l(t)\pi_l(t)^{-1/2}a_l^{-1/2}\right\} r_k^{-1},
\end{aligned}
$$

که در آن $\|\hat{\psi} - \psi\|_{\infty} \xrightarrow{p} 0$ ، $\psi_{kl}^{*}(t)$ به‌صورت عملگر الحاقی تعریف می‌شود، از این رو توزیع مجانبی $Z_{n_j}(t)$ به صورت زیر است:

$$
Z_{n_j}(t) \xrightarrow{d} \mathcal{GP}\left(\Delta_j^{\infty}, \nu_{jj}^{\infty}\right),
$$

که یک فرآیند گاوسی با میانگین مجانبی زیر است:

$$
\Delta_j^{\infty}(t) = r_j^{-1}\pi_j(t)^{1/2}\left\{d_j(t) - a_j^{1/2}\sum_{l=1}^{2}\omega_l(t)a_l^{-1/2}d_l(t)\right\}.
$$

قضیه (۱.۲) به‌عنوان یک نتیجه مهم در تقریب توزیع آماره‌ی آزمون مورد استفاده قرار می گیرد.

مراجع

[1] Bugni, F. A., Smith, J., and Lee, K. (2012). Specification test for missing functional data, *Econometric Theory*, 28(5), 959-1002.

[2] Gromenko, O., Kokoszka, P., and Sojka, J. (2017). Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves, *Ann. Appl. Stat*, 11(2): 898-918.

[3] Liebl, D., Rameseder, S. (2019). Partially observed functional data: The case of systematically missing parts, *Computational Statistics Data Analysis*, 131, 104-115.

[4] Stefanucci, M., Sangalli, L. M., and Brutti, P. (2018). PCA-based discrimination of partially observed functional data, with an application to AneuRisk65 data set, *Statistica Neerlandica*, 72(3), 246-264.

[5] Mojirsheibani, M., Shaw, C. (2018). Classification with incomplete functional covariates. *Probability Letters*, 139, 40-46.

[6] Delaigle, A., Hall, P. (2016). Approximating fragmented functional data by segments of Markov chains, *Biometrika*,103(4), 779-799.

[7] Qiu, Z., Chen, J., and Zhang, J. T. (2021). Two-sample tests for multivariate functional data with applications, *Computational Statistics Data Analysis* , 157, 107160.

[8] Kraus, D. (2015). Components and completion of partially observed functional data, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(4), 777-801.

[9] Kneip, A., Liebl, D. (2020). On the optimal reconstruction of partially observed functional data, *The Annals of Statistics*, 1692-1717.

[10] Descary, M. H., Panaretos, V. M. (2019). Recovering covariance from functional fragments, *Biometrika* , 106(1), 145-160.

[11] *Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.*