فرهنگ و اندیشهٔ ریاضی
سال ۳۹، شمارهٔ ۶۷ (پاییز و زمستان ۱۳۹۹) صص. ۱۳ تا ۳۲

علم پایان‌ناپذیر آمار و تحوّلات آینده

عبداله جلیلیان و محمدقاسم وحیدی اصل

چکیده

شکل‌گیریِ مباحثی مانند مه‌داده‌ها، داده‌کاوی، و یادگیری ماشین و پیدایش علم داده‌ها، موجب پرسش‌ها و تضادهایی در جامعه‌های آماری و نگرانی‌هایی دربارهٔ آیندهٔ رشتهٔ آمار در جهان و ایران شده است. این تحوّلات و نگرانی‌ها تا آنجا پیش رفته است که پیشنهادهایی برای تغییر عنوان و محتوای رشتهٔ آمار یا بازنگری در سرفصل درس‌ها از سوی فعالان حوزهٔ آمار کشور مطرح شده است. از جملهٔ این پیشنهادها، کاهش مطالب نظری یا به‌تعبیری، ریاضی‌زدایی از رشتهٔ آمار و تأکید بر جنبه‌های کاربردی و محاسباتی آن در برنامه‌های دانشگاهی رشتهٔ آمار است. این نوشتار با مرور تاریخیِ پیدایش علم آمار، سعی در نقد این رویکرد دارد و به چالش‌های آمار در قرن بیست‌ویکم و نیاز به یک نظریهٔ وحدت‌بخش جدید برای داده‌های قرن جدید می‌پردازد. با ذکر چند نمونه از سیر تکامل نظریه‌ها در آمار، نشان داده می‌شود که چگونه هر کدام از آماردانان دوران‌ساز، نه از سر تفنّن و کنجکاوی، بلکه بنابر ضرورت حل یک مسئلهٔ کاربردی، بخشی از پیکرهٔ عظیم آمار ریاضی یا آمار نظری امروزی را پدید آورده‌اند. ماهیت و رسالت رشتهٔ آمار ایجاب می‌کند که هم جنبه‌های نظری و هم جنبه‌های کاربردی در هر تحلیل آماری، مورد توجه آماردانان قرار گیرد.

۱. سرآغاز

گسترش روزافزونِ الگوریتم‌ها و روش‌های محاسباتیِ نوین در تحلیل‌های آماری که از دههٔ ۱۹۷۰ به بعد آغاز شد، در دو دههٔ اخیر شتاب بیشتری گرفته است. لم‌ها و قضیه‌ها که زمانی بخشی عمده از مقاله‌های بنیادی و اثرگذار آمار بودند و شالودهٔ آمار کلاسیک بر پایهٔ آنها شکل می‌گرفت، در نوشته‌های آماریِ امروزی

عبارات و کلمات کلیدی. آمار ریاضی؛ تاریخچهٔ علم آمار؛ علم داده‌ها.

کمتر به چشم می‌آیند. حتی به نظر می‌رسد که بررسی رفتارهای بزرگ‌نمونه‌ایِ حدی و مجانبیِ برآوردگرها و آزمون‌ها، جای خود را به مطالعات شبیه‌سازی داده‌اند. در واقع، با چیره‌شدنِ جنبه‌های الگوریتمی و محاسباتی بر مبانی نظری و استنباطیِ تحلیل‌های نوین آماری، پرسش‌هایی دربارهٔ سرنوشت و آیندهٔ علم آمار مطرح می‌شود که از جملهٔ آنها می‌توان موارد زیر را برشمرد:

- آیا عمر «آمار کلاسیک»، «آمار ریاضی» و «آمار نظری» به سر آمده است؟
- منظور از «آمار ریاضی» چیست؟ آیا منظور، پرداختن به ریاضیات آماری بدون توجه به کاربرد در مسائل دنیای واقعی است؟
- آیا اکنون زمان کنار گذاشتن یا استفادهٔ حداقلی از ریاضیات در حل مسئله‌ها و توسّل بیشتر به الگوریتم‌ها (کامپیوتر و نرم‌افزارها) است؟
- آیا اصولاً باید آمار را کنار گذاشت و «طرحی نو» در انداخت؟

معرفی و گسترش سریع روش‌های یادگیری ماشین[۱]، داده‌کاوی[۲] و مه‌داده‌ها[۳]، و پیدایش علم داده‌ها[۴] مزید بر علت شده است و آیندهٔ آمار را با چالش‌هایی مواجه ساخته که ضرورت بحث و تفکّر دربارهٔ ماهیت و رسالت رشتهٔ آمار را جدّی‌تر کرده است.

در رویارویی با علم آمار، دو رویکرد را می‌توان اتخاذ کرد. در یک رویکرد، برای آمار ماهیتی مجرّد و ریاضیاتی در نظر گرفته می‌شود که منجر به حل مسئله در کاربردهای گوناگون می‌شود. در این رویکرد، اصالت با مباحث نظری و ریاضی است و کاربرد این مباحث برای حل مسئله‌ها، در مرحلهٔ بعدی اهمیت قرار می‌گیرد. رویکرد دیگر، ماهیتی تجربی مانند علوم تجربی برای آمار قائل می‌شود که برای انسجام مفاهیم خود، به مبانی نظری و ریاضی نیاز دارد. در این رویکرد، اصالت با حل مسئله در کاربردهای گوناگون و پاسخ به نیازهای کاربردی به هر شیوه‌ای است و توسعهٔ نظری، در فرع قرار می‌گیرد [۱۴].

گرچه آموزش دانشگاهی و کتاب‌های درسی آمار ممکن است اغلب با رویکرد اول به معرفی مفاهیم و تحلیل‌های آماری بپردازند، سیر تاریخیِ پیدایش و گسترش علم آمار نشان می‌دهد که در تکوین آمارِ امروزی و به‌ویژه آمار ریاضی، بیشتر از رویکرد دوم پیروی شده است و توسّل به ریاضیات از سرِ ضرورت و برای دستیابی به پاسخ بی‌شبهه و ماندگار برای مسئله‌های کاربردی بوده است. به‌عنوان شواهدی برای پشتیبانی از این ادعا، در بخش‌های ۲ و ۳ به مرور مختصری از تاریخ پدید آمدنِ علم آمار جدید و آمار ریاضی می‌پردازیم.

[۱] machine learning [۲] data mining [۳] big data [۴] data science

۲. شکل‌گیری و پذیرشِ علم آمار برای تحلیل داده‌ها

کارهای جمعیتیِ جان گرانت[۱] (۱۶۲۰–۱۶۷۴) را روی داده‌های حاصل از سیاهه‌های مرگ‌ومیر در میانهٔ قرن هفدهم، می‌توان نخستین تلاش‌ها برای تحلیل آماری داده‌ها انگاشت. اما بررسی و تحلیل داده‌ها در ستاره‌شناسی[۲] و نظریهٔ خطاها[۳] نیز نقشی مهم در ایجاد و توسعهٔ روش‌های آماری داشته است. تحلیل داده‌های رصدی مربوط به حرکت سیاره‌ها توسط یوهانِس کِپلِر[۴] (۱۵۷۱–۱۶۳۰) از جملهٔ نخستین کارها در این زمینه است. نظریهٔ خطاها که با تلاش‌های پی‌یِر–سیمون لاپلاس[۵] (۱۷۴۹–۱۸۲۷)، آدریان–ماری لژاندر[۶] (۱۷۵۲–۱۸۳۳)، و کارل فریدریش گاوس[۷] (۱۷۷۷–۱۸۵۵) شکل گرفت، به مطالعهٔ مقادیر حاصل از کمیت‌هایی که به‌صورت تقریبی اندازه‌گیری می‌شدند و خطای اندازه‌گیری به‌ویژه در مسئله‌های ستاره‌شناسی، می‌پرداخت. کارهای لاپلاس، لژاندر و گاوس در این زمینه به دستاوردهایی بزرگ مانند معرفی تابع چگالی نرمال، روش کمترین توان‌های دوم[۸] و قضیهٔ حدیِ مرکزی[۹] انجامید که تأثیری شگرف در توسعهٔ روش‌های آماری داشته‌اند و همچنان دارند.

به‌کار گیریِ تحلیل داده‌ها برای کشف قاعده‌ها و بررسی فرض‌ها، تنها مختص به علوم فیزیکی و تجربی نماند و به علوم اجتماعی نیز راه پیدا کرد. در ابتدای قرن نوزدهم، لامبرِ آدولف ژاک کِتله[۱۰] (۱۷۹۶–۱۸۷۴) تلاش کرد با تحلیل داده‌های متغیرهای اجتماعی و جمعیتی گوناگون، قوانین کلی را که بر رفتار انسان حاکم هستند، استنباط کند و نوعی فیزیک اجتماعی را بنا نهد. کارهای او منجر به گسترش رویکرد کمّی و استفاده از تحلیل‌های آماری در علوم اجتماعی شد. همزمان با کارهای کِتله، با ابداع نمودارهای میله‌ای، دایره‌ای، خطی و ناحیه‌ای توسط ویلیام پلی‌فیر[۱۱] (۱۷۵۹–۱۸۲۳)، گرافیک آماری پا به عرصهٔ وجود گذاشت و در قرن نوزدهم گسترش یافت. در سال ۱۸۵۴ بیماری وبا در محلهٔ سوهو[۱۲] شهر وِست‌مینستر[۱۳] در شرق لندن شیوع پیدا کرد. جان اِسنو[۱۴] (۱۸۱۳–۱۸۵۸) توانست با استفاده از نمودار نقشهٔ نقطه‌ای[۱۵]، متوجه الگوی نقطه‌ای–خوشه‌ای مکان زندگی موارد ابتلا به بیماری وبا در نزدیکیِ تلمبه‌های آب آشامیدنی شود. این نمودار ساده، باعث شد اِسنو پی ببرد که منبع شیوع وبا که تا آن زمان هوا در نظر گرفته می‌شد، در واقع آلودگی آب است؛ یافته‌ای علمی که نقشی مهم در بهداشت عمومی سراسر جهان داشت و باعث شد اِسنو در زمرهٔ پیشگامان علم همه‌گیرشناسی[۱۶] و تحلیل داده‌های فضایی[۱۷] قرار گیرد. از این دوران به بعد، بصری‌سازی داده‌ها در قالب نمودارها به‌عنوان نوعی از تحلیل آماری داده‌ها پذیرفته شد و نمودارهای آماری گوناگونی مانند بافت‌نگاشت[۱۸]، نمودار پراکنش، پارتو، چندک–چندک، جعبه‌ای، موزاییکی و ... معرفی و در تحلیل داده‌ها به‌کار گرفته شدند.

---

[۱] John Graunt [۲] astronomy [۳] theory of errors [۴] Johannes Kepler [۵] Pierre-Simon Laplace [۶] Adrien-Marie Legendre [۷] Carl Friedrich Gauss [۸] least squares [۹] central limit theorem [۱۰] Lambert Adolphe Jacques Quetelet [۱۱] William Playfair [۱۲] Soho [۱۳] City of Westminster [۱۴] John Snow [۱۵] dot map [۱۶] epidemiology [۱۷] spatial data analysis [۱۸] histogram

تأسیس انجمن سلطنتی آمار[۱] در سال ۱۸۳۴، انجمن آمار آمریکا[۲] در سال ۱۸۳۹ و مؤسسهٔ بین‌المللی آمار[۳] در سال ۱۸۸۵ و راه‌اندازی مجله‌های علمی آماری توسط این نهادها نیز نقشی مهم در گسترش و تبیین حوزه‌های کاربردی علم آمار داشت. عضویت افرادی مانند آدولف کِتله، چارلز بَبِیج[۴] (۱۷۹۱–۱۸۷۱) ابداع‌کنندهٔ مفهوم کامپیوتر و توماس رابرت مالتوس[۵] (۱۷۶۶–۱۸۳۴)، نظریه‌پرداز شهیر اقتصاد سیاسی و جمعیت‌شناسی، در انجمن سلطنتی آمار و عضویت افرادی نظیر الکساندر گراهام بل[۶] (۱۸۴۷–۱۹۲۲) مخترع تلفن، هرمان هولریث[۷] (۱۸۶۰–۱۹۲۹) بازرگان و ابداع‌کنندهٔ ماشین جدول‌بندی کارت پانچ، اندرو کارنِگی[۸] (۱۸۳۵–۱۹۱۹) سرمایه‌دار و توسعه‌دهندهٔ صنعت فولاد ایالات متحده و مارتین ون بیورِن[۹] (۱۷۸۲–۱۸۶۲) یکی از بنیانگذاران حزب دموکرات و هشتمین رئیس جمهور ایالات متحدهٔ آمریکا، در انجمن آمار آمریکا بیانگر پذیرش نقش و اهمیت تحلیل داده‌ها در گسترهٔ وسیعی از رشته‌های علوم تجربی و اجتماعی و بخش‌های گوناگون جامعه از آغاز قرن نوزدهم است.

۳. تدوین نظریهٔ آمار و پیدایش آمار ریاضی

در پایان قرن نوزدهم، ریشه‌های نظریهٔ آمار از کارهای فرانسیس گالتون[۱۰] (۱۸۲۲–۱۹۱۱) و فرانسیس ایسیدرو اِج‌وُرث[۱۱] (۱۸۴۵–۱۹۲۶) شکل گرفت که با کارهای کارل پی‌یرسون[۱۲] (۱۸۵۷–۱۹۳۶) و جورج یودنی یول[۱۳] (۱۸۷۱–۱۹۵۱) در آغاز قرن بیستم ادامه یافت. اگرچه کار این افراد ریشه در زمینه‌های زیست‌شناسی، اقتصاد و علوم اجتماعی داشت، روش‌های آماری رسمی‌ای را پی‌ریزی کردند و توسعه دادند که نه‌تنها در زمینهٔ علمی خویش، بلکه در طیف وسیعی از علوم قابل به‌کارگیری بودند. کار این افراد روی مسئله‌های مختلف از رشته‌های مختلف، به پیدایش آمار جدید انجامید. در واقع، این دوران ابتدای قرن بیستم، یک نقطهٔ عطف در تبدیل آمار از مجموعه‌ای پراکنده از روش‌های شهودی به یک علم چارچوب‌مند و آمادهٔ به‌کارگیری برای استدلال علمی در علوم تجربی، زیستی و اجتماعی به‌شمار می‌رود.

در ادامهٔ این مسیر، کارهای اثرگذار و تلاش‌های رونالد فیشر[۱۴] (۱۸۹۰–۱۹۶۲)، جرزی نیمن[۱۵] (۱۸۹۴–۱۹۸۱)، و اِگون پی‌یرسون[۱۶] (۱۸۹۵–۱۹۸۰) به استقرار «آمار ریاضی» و برافراشتن پایه‌های آمار جدید کمکی شایان کردند. در مقابلِ رویکرد فراوانی‌گرای فیشر، نیمن و پی‌یرسون، برونو دفینیتی[۱۷] (۱۹۰۶–۱۹۸۵) و هارولد جِفریز[۱۸] (۱۸۹۱–۱۹۸۹) رویکرد بیزی به استنباط آماری را از دو منظرِ ذهنی و عینی مطرح کردند که به غنای بیشتر نگاه ریاضی به آمار و مستحکم‌تر شدن پایه‌های آمار ریاضی انجامید. همزمان با تلاش‌ها و کارهای اثرگذار این افراد و دیگران، در دههٔ ۱۹۳۰ با راه‌اندازی مجله‌های علمی جدید

[۱] Royal Statistical Society [۲] American Statistical Association [۳] International Statistical Intitute [۴] Charles Babbage [۵] Thomas Robert Malthus [۶] Alexander Graham Bell [۷] Herman Hollerith [۸] Andrew Carnegie [۹] Martin Van Buren [۱۰] Francis Galton [۱۱] Francis Ysidro Edgeworth [۱۲] Karl Pearson [۱۳] George Udny Yule [۱۴] Ronald Fisher [۱۵] Jerzy Neyman [۱۶] Egon Pearson [۱۷] Bruno de Finetti [۱۸] Harold Jeffreys

در حوزهٔ آمار و ایجاد گروه‌های آمار مجزا در دانشگاه‌های اروپا و ایالات متحدهٔ آمریکا، آمار ریاضی ریشه دوانید و رواج یافت.

آمار ریاضی سعی بر آن داشت که نظریهٔ احتمال را برای توصیف تغییرپذیری در داده‌ها و مقایسهٔ عملکرد روش‌های آماری به‌کار بگیرد و روش‌ها و فنون پراکندهٔ آماری را در قالب اصل‌ها و مفاهیم منسجم ریاضی، یکپارچه و متحد کند. در واقع، کار آمار ریاضی را می‌توان تبدیل ایده‌های تجربی، عقل سلیم و کاربست پسندیده[۱] در تحلیل داده‌ها به روش علمی دانست. برای مثال، برآورد میانگین جامعه با میانگین نمونه، یک کاربست پسندیده بود که آمار ریاضی با اصل بیشینهٔ درستنمایی، نااریبی و دارا بودنِ به‌طور یکنواخت کمترین واریانس، آن را به یک روش علمی تبدیل کرد.

در این دوران، نسل جدیدی از آماردانان پا به عرصه گذاشتند که آمار را بخشی از علوم ریاضی می‌دانستند. کارهای افرادی چون هنری شِفه[۲] (۱۹۰۷–۱۹۷۷)، اریک لِئو لی‌من[۳] (۱۹۱۷–۲۰۰۹)، دیوید بِلَک‌وِل[۴] (۱۹۱۹–۲۰۱۰)، کالیام‌پودی راداکریشنا رائو[۵] (۱۹۲۰–) و دِبابراتا باسو[۶] (۱۹۲۴–۲۰۰۱) دستاوردهایی خیره‌کننده در جستجو برای یافتن بهترین روش‌های آماری به همراه داشت و باعث پیشرفت و گسترش آمار ریاضی گردید. مقاله‌های چاپ‌شده در مجله‌های انجمن سلطنتی آمار و انجمن آمار آمریکا که در قرن نوزدهم اغلب توصیفی و متمرکز بر تحلیل یک مجموعه‌دادهٔ خاص بودند، در قرن بیستم رنگ و بوی ریاضی به خود گرفتند و پیدایش قضیه‌ها و گزاره‌ها در آنها رایج گردید [۲].

این نگاه ریاضی‌وار و احتمالاتی به تحلیل داده‌ها، ورود مباحث و مفاهیم پیشرفته‌تر از حساب دیفرانسیل و انتگرال مانند جبرخطی، فضاهای هیلبرت، نظریهٔ اندازه و آنالیز تابعی به تحلیل‌های آماری را در پی داشت. این دیدگاه جدید، تنش‌ها و کشمکش‌هایی بین طرفداران آمار ریاضی و علاقه‌مندان کار روی مسئله‌های کاربردی را به‌همراه داشت. با این حال، از آغاز تا میانهٔ قرن بیستم، آمار ریاضی از جوانه‌ای کوچک به درختی استوار و تنومند تبدیل و تحلیل آماری داده‌ها دارای مبانی نظریِ قابل دفاع شد.

۴. پیدایش کامپیوتر و تأثیر آن بر علم آمار

تا دههٔ ۱۹۶۰ و حتی ۱۹۷۰ اغلبِ روش‌ها و تحلیل‌های آماری رایج، برای نمونه‌های کوچک و با استفاده از ماشین حساب‌های دستی یا مکانیکی اجرا می‌شدند. عکس مشهور رونالد فیشر در حال کار با ماشین حساب مکانیکی خود «میلیونر[۷]» یادگاری از آن دوران است. با ورود کامپیوترها، محاسبات مربوط به همان روش‌ها و تحلیل‌های آماری با دقت، سرعت و کاراییِ بیشتر انجام شدند و اجرای تحلیل‌ها برای نمونه‌های با اندازهٔ بزرگتر هم امکان‌پذیر شد. اما معرفی نرم‌افزار آماری گلیم (GLIM[۸]) توسط بخش

[۱] good practice [۲] Henry Scheffé [۳] Erich Leo Lehmann [۴] David Blackwell [۵] Calyampudi Radhakrishna Rao [۶] Debabrata Basu [۷] Millionaire [۸] Generalized Linear Interactive Modelling

کاری انجمن سلطنتی آمار و به سرپرستی جان نِلدر[۱] برای برازش مدل‌های خطی تعمیم‌یافته به داده‌ها در سال ۱۹۷۴ غوغایی به پا کرد. نِلدر و رابرت وِدِربِرن[۲] با معرفی مدل‌های خطی تعمیم‌یافته، چارچوب نظری وحدت‌بخشی را برای ارائهٔ تعمیمی انعطاف‌پذیر از رگرسیون خطی معرفی کرده بودند که در آن، متغیر پاسخ (وابسته) می‌توانست نرمال نباشد و روش بیشینهٔ شبه‌درستنمایی[۳] را برای برآورد پارامترهای مدل خود ارائه کرده بودند. استفاده از نرم‌افزار گلیم ساده بود و هر کس که کامپیوتری در اختیار داشت، می‌توانست طیف وسیعی از مدل‌های آماری را به داده‌های خود در زمان اندکی برازش دهد. کاربر نرم‌افزار گلیم نیازی به شناخت کاملِ پس‌زمینهٔ نظری مدل‌های خطی تعمیم‌یافته نداشت و تنها با مطالعهٔ مقاله‌هایی که روش استفاده از گلیم را شرح می‌دادند، می‌توانست تحلیل آماری خود را اجرا کند. با استقبال از گلیم و پیدایش نرم‌افزارهای آماری جدیدتر، اجرای تحلیل‌های آماری توسط کامپیوترها در زمینه‌های گوناگون علمی و حتی در صنعت و تجارت به‌سرعت رایج شد.

به فاصلهٔ کمی از معرفی گلیم، در سال ۱۹۷۹ بردلی اِفرون[۴] روش بازنمونه‌گیری خودگردان[۵] را برای برآورد دقت تحلیل‌های آماری (اریبی و واریانس برآوردگرها، بازه‌های اطمینان، توان آزمون، خطای پیش‌بینی و ...) بر اساس مشاهدات نمونه معرفی کرد. کار اِفرون جنبهٔ دیگری از کاربرد کامپیوترها در اجرای روش‌ها و تحلیل‌های آماری را به نمایش گذاشته بود: این‌بار کامپیوترها نه صرفاً برای انجام محاسبات قدیمی با سرعت و دقت بیشتر، بلکه برای انجام روش‌ها و تحلیل‌های آماری جدید بر اساس شبیه‌سازی کامپیوتری به خدمت گرفته شده بودند. اِفرون علاوه بر معرفی روش خودگردان، مبانی نظری لازم برای توجیه این روش را نیز بیان کرد. در همان زمان، سایر روش‌های مبتنی بر شبیه‌سازی (معروف به مونت‌کارلو[۶]) به‌طرز انفجارآمیزی در حال غلیان بودند. اگرچه مفهوم‌های اولیهٔ روش مونت‌کارلوی زنجیر مارکُفی[۷] به دههٔ ۱۹۴۰ باز می‌گشت و استفاده از آن در فیزیک از دههٔ ۱۹۷۰ به بعد رایج شده بود، مقالهٔ آلن گِلفَند[۸] و آدریان اسمیت[۹] در سال ۱۹۹۰ بود که کاربرد روش اِم‌سی‌اِم‌سی را در پهنه‌ای گسترده از موقعیت‌ها به‌ویژه در آمار بیزی، به آماردانان نشان داد [۹].

ایده‌های هموارسازی[۱۰]، منظم‌سازی[۱۱]، و مدل‌های رگرسیونی تاوانیده[۱۲] نیز به‌مرور به تحلیل‌های آماری راه پیدا کردند که بدون استفاده از کامپیوترها، پیاده‌سازی جنبه‌های الگوریتمی آنها امکان‌پذیر نبود. مدل رگرسیونی لاسو (lasso) که در سال ۱۹۹۶ توسط رابرت تیبشیرانی[۱۳] معرفی شد، مثالی از این نوع روش‌ها است و تعمیم‌های بیشتری برای آن معرفی شده است. این تحلیل‌های آماری جدید نه‌تنها راه‌حل مسئله‌ها را تغییر دادند، بلکه شیوهٔ فکر کردن آماردانان دربارهٔ مسئله‌ها را نیز دستخوش تغییر کردند. با این حال، هم روش‌های اِم‌سی‌اِم‌سی و هم رگرسیون‌های تاوانیده مانند لاسو، دارای مبانی نظری مستدلی هستند.

---

[۱] John Ashworth Nelder [۲] Robert William Maclagan Wedderburn [۳] quasi-likelihood [۴] Bradley Efron [۵] bootstrap [۶] Monte Carlo [۷] Markov Chain Monte Carlo (MCMC) [۸] Alan Gelfand [۹] Adrian F. M. Smith [۱۰] smoothing [۱۱] regularization [۱۲] penalized regression models [۱۳] Robert Tibshirani

جنبه‌های استنباطیِ این تحلیل‌های آماری همترازِ جنبه‌های الگوریتمی مورد توجه قرار گرفت و پس از رشد و تکوین کافی، به‌سرعت در میان آماردانان سراسر دنیا به‌عنوان تحلیل‌های نوین آماری پذیرفته شدند.

بِردلی اِفرون و ترِور هَستی[۱] در فصل اول کتاب «استنباط آماری در عصر کامپیوتر[۲]» تأثیرگذاریِ پیدایش کامپیوترها و گسترش تفکّر محاسباتی بر تحلیل‌های آماری را به‌خوبی شرح داده‌اند. اِفرون و هَستی بیان می‌کنند که هر تحلیل آماری دارای دو جنبه است: جنبهٔ الگوریتمی و جنبهٔ استنباطی. جنبهٔ الگوریتمی یا محاسباتی، مشخص می‌کند که تحلیل آماریِ مورد نظر، چگونه روی داده‌ها اجرا می‌شود [۳]. امروزه جنبه‌های الگوریتمی در اغلب تحلیل‌های آماری در قالب بسته‌های نرم‌افزاری متعددی پیاده‌سازی می‌شوند و به‌صورت تجاری یا متن‌باز در دسترس همگان قرار دارند. جنبهٔ استنباطی به پشتوانهٔ مفاهیم نظری ریاضی، احتمالاتی و آماریِ مستدل، چراییِ اجرای تحلیل مورد نظر را تعیین می‌کند. برای مثال، هنگامی که برای مقایسهٔ میانگین دو جامعه از آزمون tی استیودنت دونمونه‌ای با تصحیح وِلچ[۳] استفاده می‌شود، جنبهٔ الگوریتمی این تحلیل به شیوهٔ محاسبهٔ آمارهٔ آزمون، p–مقدار و سرانجام تصمیم به رد یا پذیرش فرض صفر برابری میانگین‌های دو جامعه می‌پردازد، در حالی که جنبهٔ استنباطی، دلیل‌های موجه استفاده از این آزمون مانند بهینگی (نااریبی و پرتوان بودن) آن تحت شرط استقلال و نرمال بودن دو جامعه را مورد توجه قرار می‌دهد.

تقریباً از سال ۱۹۹۰ به بعد، توان محاسباتی و ذخیره‌سازی کامپیوترها افزایش سریعی یافت و حجم عظیم داده‌های خام حاصل از تصویربرداری‌های نجومی، تراکنش‌های بانکی، ترافیک شبکهٔ جهانی اینترنت و ... برای تحلیل به سمت مهندسان، مدیران، برنامه‌نویسان کامپیوتر و آماردانان سرازیر شدند. نیاز به تحلیل‌های خودکار، سریع، ساده و قابل‌فهم برای غیرآماری‌ها به‌طور کلی و برنامه‌نویسان کامپیوتری به‌طور خاص، باعث توجه بیشتر به روش‌هایی مانند شبکه‌های عصبی، جنگل تصادفی و ماشین بردار پشتیبان و شکل‌گیریِ مباحثی مانند مه‌داده‌ها، داده‌کاوی، یادگیری ماشین، هوش مصنوعی، و سرانجام، پیدایش علم داده‌ها شد. ریاضیاتِ به‌نسبت ساده (جبرخطی و حساب دیفرانسیل و انتگرال) در پس‌زمینهٔ این روش‌های نوین، گسترش متن‌باز کُدهای برنامه‌نویسی لازم برای پیاده‌سازی آنها، موفقیت دور از انتظار آنها، تمرکز و سرمایه‌گذاری شرکت‌های عظیم فناوری روی این روش‌ها و کاربست سریع آنها در دنیای فناوری باعث شد بسیاری از افراد مجذوب علم داده‌ها شوند و عامهٔ مردم نیز دست‌کم از طریق فناوری‌های تشخیص چهره و صوت و ترجمهٔ خودکار روی تلفن‌های همراه خود، از مزایای این روش‌های جدید بهره‌مند گردند.

این تحوّلات سریع باعث شد این تصوّر به وجود آید که با وجود الگوریتم‌هایی که عملکرد بسیار مناسبی دارند، بررسی جنبه‌های استنباطیِ تحلیل‌ها در دنیای امروزیِ داده‌های با حجم بسیار بزرگ، اهمیتی ندارد. در واقع، چنین استدلال می‌شد که تعریف ملاک‌های بهینگی و جستجو برای یافتن تحلیل آماری بهینه

[۱] Trevor Hastie [۲] computer age statistical inference [۳] Welch's correction

(برآوردهای نااریب با به‌طور یکنواخت کمترین واریانس، پرتوان‌ترین آزمون‌های یکنواخت و کوتاه‌ترین بازه‌های اطمینان نااریب) که در بررسی جنبه‌های استنباطی صورت می‌گیرد، تنها برای داده‌های با حجم محدود کارایی دارد و برای داده‌های با حجم عظیم کاری عبث است. حتی این پرسشِ ضمنی در میان آماردانان مطرح شد که بهتر نیست برای جا نماندن از متخصصان کامپیوتر و سایر رشته‌ها، تأکید بر جنبه‌های استنباطی را کنار گذاشته و تنها به جنبه‌های الگوریتمی بپردازیم؟ حتی کمی صریح‌تر، آیا اصلاً دنیای آینده به آماردان نیاز دارد یا نه؟ آیا آمار، قافیه را به علم داده‌ها باخته است؟

این پرسش‌ها ما را به پرسشی که در ابتدا مطرح کردیم بازمی‌گردانند: آیا عمر آمار ریاضی و آمار نظری سرآمده است؟ برای پاسخ به این پرسش، نخست باید دید که جایگزین آمار نظری چیست؟ منظور از آمار کاربردی چیست؟ آیا آمار نظری و کاربردی جدا از یکدیگرند و می‌توان یکی را بدون توجه به دیگری برگزید؟

۵. جنبه‌های نظری و کاربردی تحلیل‌های آماری

با پیشرفت و گسترش آمار ریاضی، مفاهیم ریاضی پیشرفته که فراتر از ریاضیات دبیرستان بودند، به‌مرور به مباحث نظری در اغلب تحلیل‌های آماری وارد شدند. با ورود این مفاهیم به آمار، اگرچه نظریهٔ آمار شکل مدوّن و زیبایی گرفت، شکایت‌هایی از قبیل اینکه افراد کمی بخش عمدهٔ این مباحث را می‌فهمند، این مفاهیم فاصلهٔ زیادی با کاربردها دارند و غیره، رفته رفته به گوش می‌رسید. گروهی از آماردانان شیفتهٔ زیبایی و یکدستیِ ریاضیات حاکم بر نظریهٔ آمار شدند و به سمت مباحث نظری شتافتند و کاربردها را در درجهٔ اهمیت پایین‌تری قرار دادند. نگرانی‌ای که دربارهٔ این رویکرد وجود داشت و همچنان پابرجا است، این است که غرق شدن در ریاضیات باعث نشود از ماهیت آمار که با مسئله‌های عالم واقعی پیوند دارد، دور شویم. از دیوید کاکس[1] نقل است که برخی از مقاله‌های امروزی آماری را فاقد ارزش می‌داند، زیرا صرفاً به‌دلیل ریاضیات موجود در آنها، از مرحلهٔ داوری جان به‌در می‌برند.

از سوی دیگر، با گسترش استفاده از کامپیوترها و ابداع روش‌های جدید مبتنی بر الگوریتم‌های محاسباتی پیشرفته که ذکر آن در بخش ۴ رفت، حجم قابل توجهی از تحلیل‌های آماری به سمت واسپاری داده‌ها به الگوریتم‌ها و دریافت نتایج بدون توجه به منطق حاکم بر این روش‌ها رفتند. برخی نوشته‌های آماری به‌وجود آمدند که فاقد اندیشهٔ آماری و روح نظری بودند و با خروجی الگوریتم‌ها یا جدول‌های شبیه‌سازی حجیم، سعی بر موجه جلوه دادنِ کار خود داشتند. اکنون شکایت‌هایی به گوش می‌رسد مبنی بر اینکه توجه به مبانی نظری کم شده است، افراد کمی هستند که بتوانند سنّت نظری قویِ شکل گرفته در دورهٔ ۱۹۰۰ تا ۱۹۷۰ را ادامه دهند، آمار ریاضی در قرن بیستم جامانده است و یا حتی نظریهٔ آمار نابود شده است.

[1] David Cox

روشن است که یک اختلاف دیدگاه در مورد وضعیت علم آمار در حال حاضر و در رقابت با علم داده‌ها بین صاحب‌نظران وجود دارد. یک دیدگاه، آمار ریاضی و مباحث نظری را عامل عدم استقبال از آمار در زمان حاضر و فرار داده‌ها به دامان الگوریتم‌ها می‌داند. دیدگاه دیگر، کم‌توجهی و بی‌مهری به مباحث نظری و در نتیجه رشد ضعیف و ناکافی آمار ریاضی همگام با تحوّلات جدید را مسبب کم‌فروغ شدن آمار می‌داند. در دیدگاه اول، آمار ریاضی، سمّی بود که سرزندگی و چابکی آمار را در آغاز قرن بیستم از آن گرفت و دیدگاه دوم، پرورش مبانی نظری را نوشدارو و اکسیر جوانی برای آمار می‌داند. اگرچه در توصیف ما از این دو دیدگاه، قدری اغراق شده است، این دو دیدگاه کم یا بیش طرفدارانی دارند و باید اعتراف کرد که رگه‌هایی از حقیقت در هر دو دیدگاه وجود دارد.

مرور تاریخ پیدایش و رشد علم آمار به شکل امروزیِ آن، نشان می‌دهد که تقریباً همهٔ مفاهیم و روش‌های آماری در آغاز به‌صورت ابتدایی و عمل‌گرایانه برای حل مسئله‌های کاربردی خاص طرح شده‌اند، سپس مورد بررسی‌های نظری و ریاضی موشکافانه قرار گرفته‌اند و سرانجام در قالب تحلیل‌های آماری چارچوب‌مند برای استفاده‌های آتی تدوین شده‌اند [۱۵]. یکی از ویژگی‌های برجستهٔ علم آمار این است که اگرچه هر روش یا تحلیل آماری معیّن، برای پاسخ به مسئله‌ای مشخص در یک زمینهٔ علمی خاص ابداع شده است، می‌توان پس از پرورش نظری و بالغ شدن، آن را برای مسئله‌های مشابه در زمینه‌های علمی گوناگون به‌کار برد. به عبارت دیگر، آمار به‌دنبال یافتن پاسخ‌های رضایت‌بخش بلندمدت است که بتوان آنها را بارها و بارها در زمینه‌های گوناگون به‌کار گرفت. این رویکرد حل مسئلهٔ آمار در تضاد با رویکرد رایج مثلاً در گرایش‌های مهندسی است که اغلب، در جستجوی یافتن پاسخ‌های کوتاه‌مدت و مختص به مسئلهٔ فعلی هستند. اما رضایت‌بخش بودن و استفادهٔ بلندمدت از پاسخ‌های آماری، مرهون بررسی نظری و ریاضی روش‌ها و تحلیل‌های آماری یا همان آمار ریاضی است. آمار از یک سو، ریشه در مسئله‌های کاربردیِ سایر زمینه‌های علمی برای تولید ایده‌ها و راه‌حل‌های جدید دارد و از سوی دیگر، برای مستدل کردن این ایده‌ها و راه‌حل‌های جدید، در چارچوب ریاضی زیست می‌کند. به همین دلیل، همهٔ تحلیل‌های آماری دارای جنبه‌های نظری و کاربردی درهم‌تنیده‌ای هستند و نمی‌توان آنها را از یکدیگر جدا دانست؛ اگرچه آماردانان بر اساس نوع آموزش، محیط کاری و علاقهٔ شخصی ممکن است تنها بر یکی از جنبه‌های نظری یا کاربردی در حوزهٔ کاری خود، متمرکز شوند و فعالیت کنند.

در بخش بعدی، به ذکر چند نمونه از رویکرد حل مسئله توسط آماردانان دوران‌سازی می‌پردازیم که نقش آنها را در تکامل علم آمار و آمار ریاضی در بخش‌های ۲ و ۳ بیان کردیم. این نمونه‌های تاریخی نشان می‌دهند که چگونه هر کدام از این افراد با هدف حل یک «مسئلهٔ واقعی» اصیل و کاربردی، هم دستاوردهای عمده‌ای در زمینه‌های علمی مورد بررسی به‌دست آورده‌اند و هم بخش‌هایی از آنچه را که پیکرهٔ عظیم آمار ریاضی یا آمار نظری نامیده می‌شود، خلق کرده‌اند. باید توجه داشت که حرفهٔ اصلی هیچ‌کدام

از این افراد، آمار به معنای مصطلح آن نبوده است و قصد همهٔ آنها رسیدن به هدف اصلی با هر مقدار لازم از ریاضیات بوده است. به عبارت دیگر، این افراد نه به‌طور تصنّعی مطالب خود را در هاله‌ای از پیچیدگی‌های ریاضی پوشانده‌اند و نه برای کاربردی جلوه دادنِ کار خود، سعی بر طفره رفتن از ریاضیات مورد نیاز کرده‌اند.

۶. چند نمونهٔ تاریخی از تحلیل داده‌ها و تکامل نظریه‌ها در آمار

همان‌گونه که فیشر بیان می‌کند [۱۰]

> «توجه بیشتر به تاریخ علم از سوی دانشمندان، به همان میزان توجه تاریخدانان ضروری است ... و این باید به معنای تلاشی سنجیده برای فهمیدن اندیشه‌های استادان بزرگ گذشته باشد که ببینند ایده‌های آنها در چه شرایطی و در کدام محیط فکری شکل گرفته است، کجاها مسیر نادرست را انتخاب کرده‌اند یا مسیر درست را تا به آخر نرفته‌اند.»

در این بخش، با ذکر چند مورد از کارهای افراد اثرگذار در علم آمار، شرایط محیطی و قلمرو فکری را که منجر به پرورش ایده‌های آنها شده است، مرور می‌کنیم. یافتن پاسخ دقیق و رضایت‌بخش برای مسئله‌های عمدهٔ علمی زمانهٔ خود، نقطهٔ اشتراک موارد زیر است.

۱.۶. جان گرانت. تاجر پارچه، وسایل خیاطی و لباس، اهل لندن بود و از تلاش‌های علمی او به‌عنوان نخستین کارها در شکل‌گیریِ علم جمعیت‌شناسی[۱]، همه‌گیرشناسی و آمار یاد می‌شود. کتاب گرانت دربارهٔ سیاهه‌های مرگ‌ومیر لندن منجر به عضویت او در انجمن سلطنتی شد. گرانت در چندین جای کتاب خود، به نداشتنِ تحصیلات دانشگاهی و استفاده‌اش از حساب مغازه‌داران به‌جای ریاضیات رسمی اشاره می‌کند. با وجود این، شکی نیست که او به‌طور کامل از اصالت رویکرد و نتایج خود آگاه بوده است [۵].

مسئله: بررسی مرگ‌ومیرهای ناشی از بیماری‌های همه‌گیر به‌ویژه طاعون و تغییرات نسبت‌های جمعیتی.

داده‌ها: سیاهه‌های هفتگی مرگ‌ومیر که حاوی خاکسپاری‌ها، غسل تعمیدها و عروسی‌های مربوط به افراد تابع کلیساهای بخش‌های لندن از سال ۱۶۰۴ تا سال ۱۶۶۱ که به‌صورت هفتگی ثبت و سیاهه‌ای برای کل سال در پایان هر سال، منتشر می‌شدند.

روش‌ها: اعتبارسنجی داده‌ها، صورت‌بندی یک مسئلهٔ آزمون فرض آماری «خام» بدون آشنایی با احتمال.

---

[۱] demography

یافته‌ها: تعیین سال‌های همه‌گیری و نرخ شیوع بیماری‌ها و پایداری برخی نسبت‌های جمعیتی مانند نسبت تولد نوزادان پسر به کل زاد و ولدها.

۲.۶. یوهانس کِپلر. ستاره‌شناس، ریاضیدان، ستاره‌بین[۱] آلمانی و یکی از چهره‌های کلیدی در انقلاب علمی قرن هفدهم بود. کارهای او دربارهٔ حرکت سیاره‌ها که منجر به قوانین کِپلر شد، نقشی تعیین‌کننده در شکل‌گیریِ رویکرد علمی و اهمیت به تحلیل داده‌های مشاهده‌شده در علوم تجربی داشت. البته قوانین کِپلر زمانی مورد قبول دانشمندان و به‌ویژه ستاره‌شناسان و فیزیکدانان قرار گرفت که نبوغ یکی از بزرگترین ریاضیدانان عالم به یاری آمد و نیوتون با ارائهٔ نظریهٔ گرانش خود، قوانین کِپلر را به‌صورت ریاضی مدلل کرد. قوانین کِپلر را می‌توان از جملهٔ نخستین مدل‌های آماری توصیفی در نظر گرفت [۷].

مسئله: چگونگی حرکت سیاره‌ها، مدار و شتاب حرکت سیاره‌ها، موقعیت مکانی یک سیاره به‌عنوان تابعی از زمان.

داده‌ها: حجم کلانی از مشاهده‌های رصدی موقعیت ستاره‌ها و سیاره‌ها در آسمان که تیکوبراهه با دقت و پشتکاری بی‌نظیر، طی حدود ۲۵ سال گردآوری و ثبت کرده بود.

روش‌ها: روش تجربی دسته‌بندی و توصیف داده‌های مشاهده‌شده برای کشف الگوهای موجود در داده‌ها (داده‌کاوی).

یافته‌ها: سه قانون دوران‌ساز کِپلر که نظریهٔ «خورشیدمرکزی کوپرنیک» را تأیید می‌کردند.

۳.۶. کارل فریدریش گاوس. ریاضیدان و فیزیکدان آلمانی بود که سهمی عمده و تأثیری شگرف در بسیاری از زمینه‌های ریاضی و علوم داشت. با اینکه او را (در کنار ارشمیدس و نیوتون) یکی از سه بزرگترین ریاضیدانانی می‌نامند که نوع بشر به خود دیده است، مهم‌ترین دستاوردهای او در ستاره‌شناسی و در نظریهٔ خطاهای مشاهداتی ناشی از رصدهای ستاره‌شناسی بوده است [۱۰].

مسئله: دقیق نبودن اندازه‌گیری‌ها و وجود خطاهای مشاهداتی در داده‌های ثبت‌شده برای کمیت‌های فیزیکی، برآورد مدار سیاره‌ها بر اساس مشاهده‌های محدود و تعیین موقعیت سیاره‌ها در مدار خود در طی زمان.

داده‌ها: داده‌های رصدی گوناگون و اندازه‌گیری‌های متنوع.

روش‌ها: کمترین توان‌های دوم، کمترین توان‌های دوم بازگشتی، رگرسیون ساده و چندگانه، برازش منحنی[۲]، توزیع تجربی خطاهای مشاهداتی.

یافته‌ها: بررسی احتمالاتی خطاهای مشاهداتی، معرفی تابع چگالی احتمال نرمال (گاوسی)، معرفی ضمنی مدل خطی مشهور $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$ به‌تعبیر امروزی آن، برآورد شش ثابت

[۱] astrologer [۲] curve fitting

(پارامتر) مدارهای بیضوی[۱] بر اساس $n > ۶$ مشاهده، برآورد مدار سیارهٔ کوتوله‌ی سِرِس[۲] و پیش‌بینی موقعیت آن برای رصدهای آتی، ورود به مسئلهٔ برآورد آماری پارامترهای مجهول با روش تجربی و حتی اشاره به برآوردی که تحت محتمل‌ترین حالتی به‌دست می‌آید که منجر به تولید داده‌ها می‌شود (روش بیشینهٔ درستنمایی).

۴.۶. فرانسیس گالتون. جامعه‌شناس، روان‌شناس، انسان‌شناس، مخترع، جغرافیدان، هواشناس، جستجوگر مناطق حارّه‌ای، از پیشگامان علم ژنتیک و آماردان انگلیسی بود. او یکی از پیشگامان علم به‌نژادی[۴] (اصلاح نژاد انسان) بود. گالتون از اولین افرادی بود که استفاده از پرسشنامه و آمارگیری[۵] را برای گردآوری داده‌ها در مورد جامعه‌های انسانی معرفی کرد. همچنین از نخستین افرادی بود که از روش‌های آماری برای مطالعهٔ تفاوت‌های ویژگی‌های انسان‌ها و وراثت هوش استفاده کرد. گالتون در سال ۱۸۸۵ تحلیل چندمتغیره را معرفی کرد. همچنین شواهدی وجود دارد که نشان می‌دهد او در سال ۱۸۷۷ تلاشی ناکام برای اجرای یک تحلیل بیزی برای توزیع نرمال با الگوریتم نمونه‌گیری با رَدّ[۶] داشته است [۱۱].

مسئله: تغییرات مشخصه‌هایی مانند میانگین و واریانس یک ویژگی ارثی (قد یا هوش) در یک جامعه.

داده‌ها: قدهای والدین و فرزندان و قدهای برادران.

روش‌ها: رورسیون یا نیاکان‌گرایی، رگرسیون.

یافته‌ها: رگرسیون[۷] (روپس‌گرایی) مشخصه‌های یک ویژگی ارثی مانند قد از فرزندان به سوی والدین و به‌عکس.

۵.۶. کارل پیرسون. ریاضیدان و زیست‌آماردان انگلیسی بود که در رشته‌های متعدد علمی تحصیل و پژوهش کرد اما کارهای او نشان‌دهندهٔ علاقهٔ خاص او به زیست‌شناسی تکاملی است. همکاری علمی و دوستی نزدیکش با والتر وِلدون[۸] تأثیر زیادی در کارهای پیرسون در زمینهٔ زیست‌سنجی[۹] و نظریهٔ تکامل داشته است. وِلدون جانورشناس داروینی بود و مسائل جالبی در ذهن داشت که نیازمند راه‌حل‌های کمّی بودند. وِلدون پیرسون را به فرانسیس گالتون معرفی کرد. پیرسون با همراهی وِلدون و گالتون، مجلهٔ بیومتریکا[۱۰] را برای توسعهٔ آمار نظری راه‌اندازی کرد و تا پایان عمر سردبیر آن ماند. پایه‌ریزی آمار ریاضی و معرفی بافت‌نگاشت، به او نسبت داده می‌شود [۱۳].

مسئله: استخراج قواعد و الگوهای پنهان در وراث و تکامل گونه‌ها بر اساس اندازه‌گیری مشخصه‌های زیستی گونه‌ها.

---

[۱] elliptic orbits [۲] dwarf planet [۳] Ceres [۴] eugenics [۵] survey [۶] rejection sampling [۷] regression [۸] Walter Frank Raphael Weldon [۹] biometrics [۱۰] Biometrika

داده‌ها: مشاهدات گوناگون در زمینه‌های زیست‌شناسی، همه‌گیرشناسی، تن‌سنجی[۱]، پزشکی و روانشناسی؛ برای مثال، داده‌های گسستهٔ مِندِل[۲] در مورد اَلِل[۳]های غالب و مغلوب.

روش‌ها: آزمون نیکوییِ برازش خی‌دو[۴]، ضریب همبستگی، روش برآورد گشتاوری، تحلیل مؤلفه‌های اصلی.

یافته‌ها: پایه‌گذاری مکتب زیست‌سنجی برای توصیف تکامل و وراثت جمعیتی و پدید آوردن روش‌های آماری برای زیست‌سنجی [۱].

۶.۶. رونالد فیشر. آماردان و ژنتیک‌شناس[۵] انگلیسی بود. او را به‌دلیل کارهایش در آمار، «نابغه‌ای که تقریباً به‌تنهایی پایه‌های علوم آماری جدید را ایجاد کرد» و «مهم‌ترین چهره در آمار قرن بیستم» توصیف کرده‌اند. فیشر از ریاضیات برای ترکیب ژنتیک مِندِلی و انتخاب طبیعی داروین، استفاده کرد که به احیای داروینیسم در آغاز قرن بیستم انجامید. به همین دلیل، او را «بزرگترین جانشین داروین» نامیده‌اند. ترویج روش بیشینهٔ درستنمایی و استخراج ویژگی‌های آن، پایه‌ریزی اصول طرح آزمایش‌ها، استنباط اتکایی[۶] و استخراج توزیع‌های نمونه‌ای متعدد، از عمده‌ترین کارهای فیشر در آمار است. در سال ۱۹۲۵ کتاب «روش‌های آماری برای پژوهشگران[۷]» را منتشر کرد که یکی از تأثیرگذارترین کتاب‌های قرن بیستم در روش‌های آماری بود. در این کتاب بود که به مفهوم p–مقدر عمومیت داده شد؛ مفهومی که نقشی اساسی در رویکرد فیشر دارد [۸].

مسئله: تأثیر عوامل مختلف بر میزان محصول کشاورزی.

داده‌ها: حجم کلانی از داده‌ها که طی «آزمایش‌های میدانی کلاسیک» روی محصولات کشاورزی از دههٔ ۱۸۴۰ به بعد در ایستگاه آزمایشگاهی روتامستِد[۸] گردآوری شده بود.

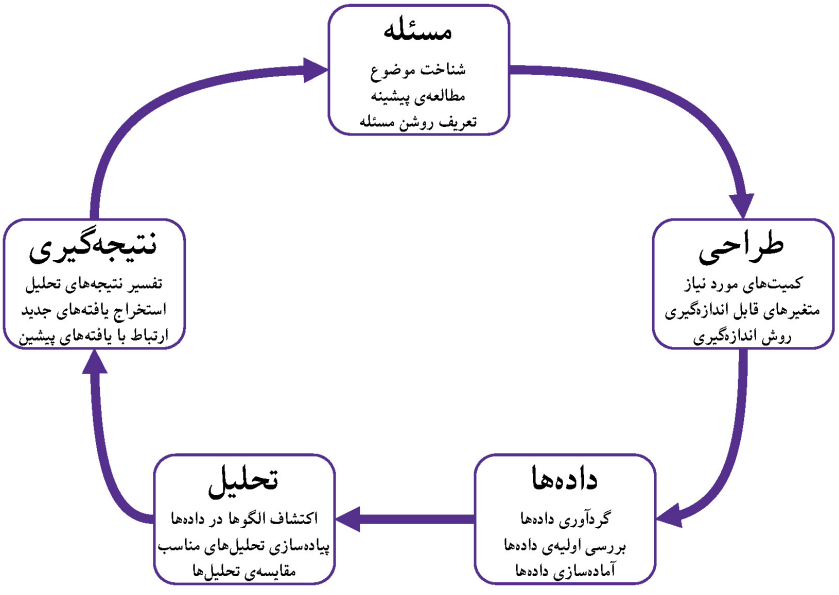
روش‌ها: تحلیل واریانس، برآورد بیشینهٔ درستنمایی، p–مقدار.

یافته‌ها: مقایسهٔ میانگین محصول‌ها و نرخ ضایعات در قطعه‌های زمین شامل سطح‌های مختلف نیتروژن، سولفات پتاسیم، سولفات سدیم و سولفات منیزیم.

۷. انتظارها از تحلیل‌های آماری و متخصصان آمار

امروزه می‌توان روش‌ها و ایده‌های آماری را تقریباً در هر زمینهٔ فکری که با داده‌ها و استدلال کمّی در مورد داده‌ها سروکار دارد، یافت. آمار به‌طور گسترده در علوم تجربی مانند ستاره‌شناسی، زیست‌شناسی، مهندسی، جغرافیا، پزشکی، بهداشت عمومی و علوم اجتماعی از جمله علوم سیاسی، حقوق، جامعه‌شناسی، روان‌شناسی، انسان‌شناسی، باستان‌شناسی و تاریخ به‌کار گرفته می‌شود. اما همان‌طور که اشاره شد، آمار

[۱] anthropometry [۲] Mendel [۳] allele [۴] chi-square goodness of fit test [۵] geneticist [۶] fiducial inference
[۷] Statistical Methods for Research Workers [۸] Rothamsted experimental station



شکل ۱. چرخهٔ بررسی و ارائهٔ راه‌حل آماری برای یک مسئله.

با همهٔ این زمینه‌های علمی در تعامل است. پس از هر تحلیل آماری و پاسخ‌گویی به مسئلهٔ مورد بررسی در یک زمینهٔ علمی، نتایج به‌دست آمده، اغلب محرکی برای بازتعریف مسئله یا طرح مسئله‌های جدیدتری هستند و به تحلیل‌های آماری بیشتر و راه‌حل‌هایی جدیدتر می‌انجامند. در واقع، همان‌طور که در شکل ۱ نشان داده شده است، فرایند جستجوی پاسخ و ارائهٔ راه‌حل آماری برای یک مسئله، چرخه‌ای است که همواره در جریان است. این چرخه به پژوهش‌های امروزی آمار، ماهیت بین‌رشته‌ای داده است و باعث پویایی علم آمار شده است.

۱.۷. انتظارها از یک تحلیل آماری. ماهیت بین‌رشته‌ایِ پژوهش‌های امروزی آمار همراه با پیدایش داده‌کاوی، یادگیری ماشین و علم داده‌ها، باعث طرح این پرسش می‌شود که آیا هر پردازش یا تحلیلی را که روی داده‌ها اجرا می‌شود، می‌توان یک «کار آماری» دانست؟ پاسخ به این پرسش نیز مستلزم یک تعریف روشن برای آمار و نقش آماردانان است. چندین مقاله با عنوان «آمار چیست؟» توسط آماردانان شهیر در یک قرن اخیر به نگارش درآمده است که نشان‌دهندهٔ دشواری پاسخ به این پرسش ساده و همچنین کافی نبودن پاسخ‌ها است. برخی از تعریف‌های آمار در نوشتگان آماری که توسط فاین‌برِگ در [۴] گردآوری شده است، از این قرار هستند:

(الف) یک روش استدلال است به‌همراه ابزارها و روش‌هایی که برای کمک به فهم بهتر جهان طراحی شده‌اند.

(ب) هنر ایجاد حدس‌های عددی برای پرسش‌های گیج‌کننده است.

(پ) مجموعه‌ای از مفاهیم، قواعد و روش‌ها برای گردآوری داده‌ها، تحلیل داده‌ها و نتیجه‌گیری از داده‌ها است.

(ت) رویکردی نظام‌مند برای یافتن پاسخ‌های مستدل همراه با ارزیابی اعتبار آنها در موقعیت‌هایی است که اطلاعات کامل، به‌موقع موجود نیست، یا غیرقابل دستیابی است.

(ث) علم و هنر گردآوری، تحلیل و استنباط از داده‌ها است.

(ج) هنر یادگیری از داده‌ها است و به گردآوری داده‌ها، توصیف متعاقب داده‌ها و تحلیل داده‌ها می‌پردازد و اغلب به نتیجه‌گیری منجر می‌شود.

(چ) گردایه‌ای از رویّه‌ها و اصول برای به‌دست آوردن و پردازش اطلاعات به‌منظور تصمیم‌گیری در هنگام مواجهه با عدم قطعیت است.

(ح) پیکره‌ای از روش‌ها برای تصمیم‌گیری‌های خردمندانه در مواجهه با عدم قطعیت است.

(خ) رشته‌ای است که به مطالعهٔ تغییرپذیری و عدم قطعیت می‌پردازد.

صرف‌نظر از اینکه کدام تعریف را بپذیریم، سه اصل را می‌توان اصولی دانست که در هر تحلیل آماری باید از آنها پیروی کرد. اول اینکه داده‌های مورد بررسی بر اساس یک مدل احتمالاتی تولید شده‌اند؛ دوم اینکه این مدل احتمالاتی با استفاده از داده‌ها، برازش می‌شود و سوم اینکه با شبیه‌سازیِ یک مدل برازش‌یافته، داده‌هایی به‌دست می‌آیند که از داده‌های واقعی غیر قابل تشخیص هستند. هر تحلیلی را که به‌گونه‌ای از سه اصل بالا پیروی کند، می‌توان یک تحلیل آماری نامید. اما برای مدل‌سازی احتمالاتی داده‌ها نیز رویکردها و دیدگاه‌های متفاوتی وجود دارد. در اینجا برای نمونه، اصول مدل‌سازی کِپلر به‌صورت خلاصه آورده می‌شود.

- ساختن مدل، فرایندی پیاپی است که از مدل‌های ساده‌تر شروع و به موارد پیچیده‌تر ختم می‌شود؛
- پیچیدگیِ یک مدل هم به ساختار ریاضی آن و هم به تعداد پارامترها بستگی دارد؛
- مدل‌ها باید از لحاظ ریاضی، زیبا و هماهنگ باشند؛
- مدل‌ها باید بر یک نظریهٔ وابسته به جهان فیزیکی در مورد پدیده‌ها مبتنی باشند؛
- معیار نهایی برای انتخاب از بین چندین مدل، سازگاری آن با مشاهدات، یعنی همان نیکوییِ برازش است.

این اصول باید در مراحل مختلف فرایند ساختن مدل، با یکدیگر تعامل داشته باشند. به عبارت دیگر، مدلی با پارامترهای متعدد که بر اساس یک نظریهٔ ریاضی ساخته شده، ممکن است به‌نفع مدلی که تعداد کمتری پارامتر دارد و مبتنی بر نظریهٔ ریاضی دیگری است، کنار گذاشته شود. برای یک آماردان، اصول مدل‌سازی کِپلر حتی امروزه هم آموزنده و ارزشمند هستند [۶].

۲.۷. انتظارها از یک متخصص آمار. یک کارورز[۱] آمار در حوزه‌های زیست‌شناسی، پزشکی، بهداشت، اقتصاد، علوم اجتماعی، مهندسی، ... نیازی به پرداختن به جنبهٔ استنباطی تحلیل‌های آماری ندارد و تنها جنبه‌های الگوریتمی آنها که اغلب توسط بسته‌های نرم‌افزاری کامپیوتری قابل اجرا و پیاده‌سازی هستند، مورد توجه وی است. اما یک کارشناس و متخصص آمار نمی‌تواند تنها به جنبهٔ الگوریتمی روش‌ها و تحلیل‌های آماری بسنده کند و از او انتظار می‌رود که علاوه بر جنبهٔ الگوریتمی، از جنبهٔ استنباطی تحلیل‌های خود نیز آگاهی داشته باشد و دلیل و توجیه قابل قبولی برای انتخاب تحلیل مورد نظر خود برای یک مجموعه از داده‌ها و مقایسهٔ آن با سایر تحلیل‌های جایگزین ارائه کند.

برای مثال، تحلیل آماری مسئلهٔ پیش‌بینی مقادیر آتی یک سری زمانی مشاهده شده را با استفاده از مدل‌های فضای وضعیت نوآوری‌ها برای هموارسازی نمایی[۲] در نظر بگیرید. جنبه‌های الگوریتمی این تحلیل آماری در بستهٔ نرم‌افزاری `forecast` در R پیاده‌سازی شده و به‌طور متن‌باز در دسترس همگان است[۳]. یک مهندس، کارشناس اقتصاد، فعال بخش درمان، ... می‌تواند با یک جستجوی ساده و صرف زمانی اندک، با مطالعهٔ راهنمای این بستهٔ نرم‌افزاری، جنبهٔ الگوریتمی این تحلیل را به‌خوبی فراگیرد. اما آیا می‌توان او را یک کارشناس یا متخصص آمار دانست؟ آنچه وجه تمایز یک کارشناس آمار از یک کارورز آمار است، آشنایی و تسلط او بر جنبه‌های استنباطی تحلیل‌های آماری است. از یک کارشناس آمار انتظار می‌رود که علاوه بر جنبه‌های الگوریتمی تحلیل، جنبه‌های استنباطی آن مانند ویژگی‌های پیش‌بینی‌های به‌دست آمده را بشناسد و از عملکرد آنها در مقایسه با پیش‌بینی‌هایی که ممکن است با استفاده از مدل‌های ARIMA به‌دست آمده باشند، آگاه باشد.

بنابراین از دانش‌آموختگان مقطع‌های کارشناسی، کارشناسی ارشد و دکتری تخصصی آمار انتظار می‌رود که با جنبه‌های الگوریتمی و استنباطی تحلیل‌های آماری آشنا باشند تا بتوانند در مواجهه با داده‌ها و مسئله‌های واقعی، تحلیل‌های آماری مناسب را انتخاب کنند و شیوهٔ مناسبی به اجرا درآورند. در این راستا، چارچوب کلی انتظاراتی که به باور نگارندگان باید از دانش‌آموختگان آمار داشت، عبارت‌اند از:

کارشناس آمار: آشنایی با جنبه‌های الگوریتمی و استنباطی تحلیل‌های آماری رایج و پُرکاربرد و توانایی کاربست خلاقانه و آگاهانهٔ این روش‌ها در کاربردهای گوناگون؛

---

[۱] practitioner [۲] innovations state space models for exponential smoothing [۳] https://cran.r-project.org/web/packages/forecast/index.html

کارشناسی ارشد آمار: ژرف شدن در جنبه‌های استنباطی تحلیل‌های آماری رایج و آشنایی با جنبه‌های الگوریتمی و استنباطی تحلیل‌های آماری نوین برای یک مسئلهٔ معیّن؛

دکتری تخصصی آمار: ابداع و توسعهٔ تحلیل‌های آماری نوین برای یک مسئلهٔ معیّن که مستلزم نوآوری در جنبه‌های استنباطی و الگوریتمی است.

۸. آیندهٔ علم آمار

فیشر از مجموعه‌داده‌های گل زنبق[۱] که توسط اِدگار اَندِرسون[۲] برای مسئلهٔ رده‌بندی گونه‌ها[۳] گردآوری شده بود، به‌عنوان مثالی برای تحلیل تشخیصی خطی[۴] در سال ۱۹۳۶ استفاده کرد. در این مجموعه‌داده‌ها، $p = ۴$ متغیر همبستهٔ درازای کاسبرگ، پهنای کاسبرگ، درازای گلبرگ، و پهنای گلبرگ برای $n = ۵۰$ از هر یک از $c = ۳$ گونه از گل زنبق، اندازه‌گیری شده است. تحلیل تشخیصی خطی برای این مجموعه‌داده‌ها و موارد مشابه با آن، عملکردی مطلوب در رده‌بندی گونه‌ها دارد. اما عملکرد تحلیل تشخیصی خطی برای تعداد $n = ۵۰$ تومور سرطانی از هر یک از $c = ۳$ رسته که برای هر یک از آنها $p = ۴۰۰۰۰۰۰$ متغیر همبسته (برای مثال، ۲۰۰۰ طیف جرمی[۵] در ۲۰۰۰ زمان مختلف) اندازه‌گیری شده‌اند، رضایت‌بخش نیست. این در حالی است که برخی از الگوریتم‌های یادگیری بانظارت[۶] امروزی در یادگیری ماشین، عملکردی خیره‌کننده در تحلیل چنین داده‌هایی و پاسخ به مسئلهٔ رده‌بندی دارند.

در قرن بیست‌ویکم، با مجموعه‌داده‌هایی سر و کار داریم که هزاران بار یا میلیون‌ها بار بزرگتر از مجموعه‌داده‌هایی هستند که بخش اعظم مبانی نظری تحلیل‌های آماری بر پایهٔ آنها بنا شده است. امروزه چنین مجموعه‌داده‌های بزرگی تنها در پروتئومیک[۷] (دانش بررسی پروتئین‌ها) یا ژنومیک[۸] (دانش بررسی ژن‌ها) ظاهر نمی‌شوند، بلکه همه‌جا با آنها مواجه می‌شویم. داده‌های میلیاردی صفحات وب، داده‌های رصدی چندین ترابایتی از یک تلسکوپ امروزی، تراکنش‌های تجارت الکترونیکی، تصویربرداری‌های مداوم ماهواره‌ای و انبوهی از دستگاه‌های جمع‌آوری داده‌های خودکار با توان عملیاتی بالا در علوم و فناوری، مثال‌هایی از داده‌های رایج قرن بیست‌ویکم هستند.

به نظر می‌رسد بسیاری از مبانی نظری علم آمار مناسب قرن بیست‌ویکم نیست و به همین دلیل، تحلیل‌های آماری مبتنی بر آنها، برای داده‌های امروزی «خوب کار نمی‌کنند». در مقابل، الگوریتم‌های یادگیری ماشین، داده‌کاوی و علم داده‌ها بدون آنکه بدانیم یا مطمئن باشیم چرا، «خوب کار می‌کنند». این باعث شده است که داده‌کاوان، مهندسان و کسان دیگری که تنها «به‌دنبال چیزی هستند که کار می‌کند»، با اشتیاق به سمت این‌گونه الگوریتم‌ها بروند.

---

[۱] Iris flower data set [۲] Edgar Anderson [۳] taxonomic [۴] linear discriminant analysis [۵] mass spectra
[۶] supervised learning [۷] proteomics [۸] genomics

مبانی نظری لازم برای اجرای تحلیل تشخیصی خطی شامل توزیع نرمال چندمتغیره و نسبت درستنمایی و ایدهٔ فیشر برای استفاده از آنها در تحلیل چنین داده‌هایی، قابل درک است. اما برخی از الگوریتم‌های بانظارت یادگیری ماشین، مانند شبکه‌های عصبی ژرف[1] که عملکرد موفقی دارند، فاقد مبانی نظری کافی هستند و جنبهٔ استنباطی آنها بررسی نشده است. پس از فروکش کردن هیجان‌های اولیه و معرفی انواع گوناگونی از شبکه‌های عصبی ژرف برای مسئله‌های متعدد، بحث‌هایی دربارهٔ لزوم پیشنهادهایی برای تعیین روش مناسب و به‌ویژه تعیین پارامترهای تنظیمی این روش‌ها مطرح شده است. بی‌تردید برای حرکت در این مسیر، چاره‌ای جز توجه به جنبه‌های استنباطی و مبانی نظری لازم برای این روش‌ها نیست. حرکت در این مسیر از مهندسان، برنامه‌نویسان و دیگران برنمی‌آید و از آماردانان انتظار می‌رود که جنبه‌های استنباطی لازم برای داده‌های حجیم عصر جدید را ارائه دهند.

از این‌رو شرایط رشتهٔ آمار در آغاز قرن بیست‌ویکم، مشابه شرایط این رشته در آغاز قرن بیستم است. اکنون زمان آن است که مبانی نظری آمار توسعه داده شود تا مناسبِ شرایط و تحوّلات قرن بیست‌ویکم گردند. مطالعهٔ جنبه‌های استنباطی الگوریتم‌های یادگیری ماشین به‌ویژه شبکه‌های عصبی ژرف، ممکن است نقطهٔ شروع خوبی باشد. پیدایش مبانی نظری و جنبه‌های استنباطی جدید در آینده دور از انتظار نیست. شاهدی بر این ادعا، کتاب «استنباط آماری در عصر کامپیوتر» نوشتهٔ اِفرون و هَستی است که هر دو از آماردانان پیشگام عصر کامپیوتر هستند. این دو نفر در کتاب خود بیان می‌کنند که روش‌های امروزی یادگیری ماشین، داده‌کاوی، مه‌داده‌ها و هوش مصنوعی در حال حاضر تنها الگوریتم‌هایی هستند که برای حل مسئله‌ها و تحلیل داده‌های حجیم عصر کامپیوتر مناسب هستند. اما این روش‌ها برای تبدیل شدن به تحلیل‌های آماری نیازمند مبانی نظری مستحکم و جنبه‌های استنباطی هستند. همان‌گونه که فیشر، پی‌یرسون و دیگران با توسعهٔ جنبه‌های استنباطی دقیق در چند دههٔ نخست قرن بیستم، روش‌ها و الگوریتم‌های پراکندهٔ پیش از خود را در قالب تحلیل‌های آماری متقن تدوین کردند، به فیشرها و پی‌یرسون‌های جدیدی نیاز است که با ارائهٔ مباحث نظری جدید و بررسی جنبه‌های استنباطی روش‌ها و الگوریتم‌های جدید، آنها را به تحلیل‌های آماری بالغی تبدیل کنند. به همین دلیل، آماردانان در عصر کامپیوتر و علم داده‌ها، نباید رسالت خود را که توجه به هر دو جنبهٔ الگوریتمی و استنباطی تحلیل‌های آماری است، فراموش کنند. درست زمانی که برخی مرگ آمار ریاضی و نظریهٔ آمار را اعلام می‌کنند، بیش از هر زمان دیگری به آمار ریاضی و یک نظریهٔ وحدت‌بخش جدید برای داده‌های قرن جدید نیاز داریم.

جنبه‌های نظری و کاربردی آمار قابل تفکیک نیستند. توجه به جنبه‌های نظری در دوران کنونی لازم است اما کافی نیست. نمونه‌های تاریخی بیان‌شده در بخش ۶ نشان می‌دهند که گسترش و پویاییِ علم آمار از یک سو به‌واسطهٔ تعامل آن با زمینه‌های علمی گوناگون و تحوّلات زمانه مانند رنسانس، انقلاب

[1] deep neural networks

صنعتی، مطرح شدن نظریهٔ تکامل و شور و شوق اولیه در برخورد با علم ژنتیک، سرشار از ایده‌های نو شده است؛ و از سوی دیگر، به پشتوانهٔ مبانی نظری و آمار ریاضی توانسته است این ایده‌ها را به روش‌های علمی بالغ و مستدلی تبدیل کند. از این‌رو تعامل بیشتر با سایر زمینه‌های علمی و یافتن و پاسخ‌گویی به مسئله‌های اصیل و اثرگذار علمی در زمینه‌های کاربردیِ گوناگون، نقشی تعیین‌کننده در آیندهٔ آمار دارد.

۹. سخن آخر

باور نگارندگان بر این است که اگرچه جنبه‌های کاربردی علم آمار در ایران مغفول مانده و کمتر مورد توجه برنامه‌های دانشگاهی قرار گرفته است، وجود یا تأکید برنامه‌های درسی بر مباحث نظری آمار علت این مشکل نبوده است و ریشهٔ اصلی این مشکل را باید در خارج از محیط دانشگاه و در شرایط اقتصادی و اجتماعی جامعه جُست. سطح نازل توسعهٔ صنعتی در ایران در کنار انحصاری و غیررقابتی بودنِ بخش عمده‌ای از اقتصاد کشور، از جملهٔ دلایل عدم اقبال به جنبه‌های کاربردی علوم به‌طور کلی و آمار به‌طور خاص، است. بررسی این موضوع و پرداختن به اینکه نیاز واقعی جامعه به آمار تا چه اندازه است، از حوصلهٔ این نوشتار خارج است و نوشتاری دیگر را می‌طلبد. در پایان، تنها متذکر می‌شویم که عدم توجه به مبانی نظری روش‌های آماری، نه‌تنها مشکلی را رفع نمی‌کند، بلکه قابلیت‌های علمی کشور را که به دلایل متعددی، بالفعل نشده است، کاهش خواهد داد.

مراجع

[۱] وحیدی اصل، محمدقاسم، *تاریخ آمار*، انتشارات مبتکران، تهران، ۱۴۰۰.

[2] Billard, Lynne, The role of statistics and the statistician, *The American Statistician*, **52** (1998), no. 4, 319–324.

[3] Efron, Bradley and Hastie, Trevor, *Computer Age Statistical Inference*, Cambridge University Press, Cambridge, 2016.

[4] Fienberg, Stephen E., What is Statistics? *Annual Review of Statistics and Its Application*, **1** (2014), 1–9.

[5] Glass, David Victor, John Graunt and his natural and political observations, *Proceedings of the Royal Society of London, Series B (Biological Sciences)*, **159** (1963), no. 974, 2–37.

[6] Hald, Anders, *A History of Probability and Statistics and Their Applications Before 1750*, John Wiley & Sons, New York, 2003.

[7] Lehmann, Eric L., Model specification: the views of Fisher and Neyman, and later developments, In: *Selected Works of EL Lehmann*, pp. 955–963, Springer-Verlag, Berlin, 2012.

[8] Rao, C. Radhakrishna and others, RA Fisher: The founder of modern statistics, *Statistical Science*, **7** (1992), no. 1, 34–48.

[9] Robert, Christian and Casella, George, A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data, *Statistical Science*, **7** (1992), 102–115.

[10] Sprott, David A., Gauss's contributions to statistics, *Historia Mathematica*, **5** (1978), no. 2, 183–203.

[11] Stigler, Stephen M., Darwin, Galton and the statistical enlightenment, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **173** (2010), no. 3, 469–482.

[12] Vapnik, Vladimir, *The Nature of Statistical Learning Theory*, Springer-Verlag science & business media, 2013.

[13] Walker, Helen M., The contributions of Karl Pearson, *Journal of the American Statistical Association*, **53** (1958), no. 281, 11–22.

[14] Wild, Christopher J., Utts, Jessica M., Horton, Nicholas J., What is statistics? In: *International Handbook of Research in Statistics Education*, pp. 5–36, Springer-Verlag, 2018.

[15] Wild, Chris J., Pfannkuch, Maxine, Statistical thinking in empirical enquiry,*International statistical review*, **67** (1999), no. 3, 223–248.

عبداله جلیلیان: دانشگاه رازی، دانشکدهٔ علوم، گروه آمار
تارنما: https://jalilian.github.io/
رایانامه: jalilian@razi.ac.ir

محمدقاسم وحیدی اصل: دانشگاه شهید بهشتی، دانشکدهٔ علوم ریاضی، گروه آمار
رایانامه: m-vahidi@sbu.ac.ir