آمار بسامدگرا به مثابهٔ نظریهٔ استنباط استقرایی•

دیبورا جی. مایو*، دی. آر. کاکس**

ترجمهٔ محمدرضا مشکانی

۱. آمار و فلسفهٔ استقرایی

۱.۱ فلسفهٔ آمار چیست؟

مبانی فلسفی آمار را می‌توان به مثابهٔ بررسی مسأله‌های شناخت‌شناسانه، مفهومی، و منطقی دربارهٔ کاربست و تفسیر روشهای آماری، به‌معنای کلی، تلقی کرد. همچون دیگر حوزه‌های فلسفهٔ علم، پژوهش در علم آمار تا حد زیادی بدون دغدغه دربارهٔ «مبانی فلسفی» به‌پیش می‌رود. با وجود این، حتی در کار حرفه‌ای آمار، بحث و جدلها دربارهٔ رویکردهای متفاوت به تحلیل آماری ممکن است بر مسائل کلّیِ ماهیتِ استنباط آماری-استقرایی اثر بگذارند و از آنها متأثر شوند، و بدین ترتیب با مباحث بنیادی و فلسفی سروکار داشته باشند. حتی کسانی که عمدتاً با کاربردها سروکار دارند اغلب علاقه‌مند به شناسایی آن اصلهای کلی‌اند که زیربنای شیوه‌های مورد استفاده و توجیه‌کنندهٔ آنها هستند و کاربران به‌دلایل کمابیش عملی به ارزش آنها واقف شده‌اند. دست‌کم در یک سطح از تحلیل، آمارشناسان و فیلسوفانِ علم بسیاری پرسشهای مشابه را مطرح می‌کنند.

- چه چیزی باید مشاهده شود و از داده‌های حاصل چه چیزهایی را به‌طور موجه می‌توان استنباط کرد؟
- داده‌ها تا چه حدّ مدلی را تأیید می‌کنند یا به آن می‌برازند؟
- آزمون خوب چیست؟
- آیا شکست در ردّ فرض H شاهدی بر «تأیید» H است؟
- چگونه می‌توان تعیین کرد که یک بی‌هنجاری ظاهری، واقعی است یا نه؟ چگونه می‌توان تقصیر یک بی‌هنجاری را به‌درستی به‌عامل آن نسبت داد؟
- اگر نگاه‌کردن به داده‌ها بر فرضی که باید امتحان شود تأثیر بگذارد، آیا این امر به‌رابطهٔ بین داده‌ها و فرض ربطی دارد؟
- چگونه می‌توان رابطه‌های کاذب را از نظمهای واقعی متمایز ساخت؟
- چگونه می‌توان یک تبیین علّی و یک فرض را توجیه و آزمون کرد؟
- چگونه می‌توان اختلاف بین داده‌های موجود و ادعاهای نظری را به‌شیوهٔ قابل اعتمادی رفع کرد؟

این پرسشهای بسیار عام که با مجادلات دیرپای فلسفهٔ علم درهم تنیده‌اند به تبیین این موضوع که چرا رشتهٔ آمار، خواه به تصریح یا به تلویح، گرایش به ورود به حیطهٔ فلسفه دارد، کمک می‌کند. برخی حتی ممکن است آمار را به نوعی «فلسفهٔ علم کاربردی» (فیشر [۱۰]، کمپتورن [۱۳])، و نظریهٔ آمار را نوعی «فلسفهٔ استنباط استقرایی کاربردی» تلقی کنند. چنانکه لیمن [۱۵] تأکید کرده است، نیمن کار خود را نه تنها کاری در آمار بلکه کاری در فلسفهٔ استقرایی نیز تلقی می‌کرد. پرسشی اساسی که در «فلسفهٔ استقرایی» هم در آمار و هم در فلسفه مطرح است این است که: ماهیت و نقش مفهومهای احتمالاتی، روشها، و مدلها در انجام استنباط، هنگامی که با داده‌های محدود، عدم حتمیت و خطا سروکار داریم، چیست؟

در این فرصتی که برای شرکت در جلسهٔ راجع به فلسفهٔ آمار در «دومین سمپوزیوم لیمن» به‌دست آورده‌ایم، توصیهٔ نیمن ([۲۲]، ص ۱۷) را به‌عنوان مبنای بحث خود برمی‌گزینیم و نظریهٔ آمار را اساساً «نظریهٔ بسامدگرای استنباط استقرایی» تلقی می‌کنیم. آنگاه این پرسش پیش می‌آید که چه تصور (یا تصوراتی) از استنباط استقرایی این گزینش را مجاز می‌سازد. برای بررسی اینکه آیا این روایت، تنها یا حتی رضایت‌بخش‌ترین روایت از استنباط استقرایی است یا نه، جالب است بررسی شود که چه مقدار پیشرفت به‌سوی روایتی از استنباط استقرایی، در مقابل رفتار استقرایی، ممکن است از آمار بسامدگرا (با عطف توجه به آزمون‌کردن و روشهای مربوط) عاید شود. این روشها اغلب برای هدفهای استنباطی، یعنی کسب آگاهی از جنبه‌های سازوکار زیربنایی مولد داده‌ها، به‌کار می‌روند، و در صورتی که شفافیت بیشتری در مورد نقشهای احتمال خطاهای فرضی در استنباط وجود داشته باشد، می‌توان از بسیاری از سردرگمی‌ها و انتقادها (مثلاً، در این مورد که آیا باید نرخ خطاها تعدیل شوند یا نه و چرا باید تعدیل شوند) جلوگیری کرد.

با پذیرش نظرات فیشر [۱۰]، دیدگاه لیمن [۱۵] دربارهٔ نیمن، و پویر [۲۶] راجع به استقرا به‌عنوان پیشینهٔ موضوع، نقش آزمونهای معنی‌داری را در پرکردن شکافهای استقرایی موجود در استنباط استقرایِ فرضیِ سنتی

مورد بحث قرار می‌دهیم. هدف ما شناسایی یک اصل اساسی دربارهٔ شواهد است که به‌وسیلهٔ آن بتوان احتمال خطاهای فرضی را برای انجام استنباط استقرایی از داده‌های خاص به‌کار بست، و نیز بررسی این نکته که این اصل چگونه می‌تواند موارد زیر را هدایت و توجیه کند: (الف) کاربردها و تفسیرهای متفاوت از سطحهای استنباط آماری در آزمون انواع مختلف فرضهای صفر، و (ب) چه موقع و چرا «اثرهای گزینش» باید در آزمون آماریِ داده‌وابسته به‌حساب آورده شوند.

۲.۱ نقش احتمال در استقرای بسامدگرا

ویژگی تعریف‌کنندهٔ استنباط استقرایی آن است که مقدمات گزاره‌های مبتنی‌بر شواهد ممکن است درست باشند ولی نتیجهٔ استنباط‌شده بدون تناقضی منطقی نادرست باشد: یعنی نتیجه «فراتر از شواهد است». احتمال در به‌دست آوردن این گونه استنباطهای فراتر از شواهد به‌طور طبیعی ظاهر می‌شود. اما بیش از یک راه برای استفاده از احتمال وجود دارد. پیرسون ([۲۴]، ص ۲۲۸) دو سنّت فلسفی متمایز برای به‌کار بستن احتمال در استنباط را جمع‌بندی کرده است:

«در نظر یکی از دو مکتب، درجهٔ اطمینان به یک گزاره، یعنی کمیتی که با ماهیت و میزان شواهد تغییر می‌کند، مفهومی پایه‌ای است که باید به آن مقیاس عددی نسبت داد.» مکتب دیگر به ارتباط بین آگاهی از بسامد نسبی رخ‌دادن رده‌ای خاص از پیشامدها در سلسله‌ای از تکرارها با زندگی عادی و با بسیاری از شاخه‌های علوم توجه دارد، و بر آن است که «از طریق پیوند احتمال با بسامد نسبی است که احتمال صریح‌ترین معنا را در ذهن انسان پیدا می‌کند.»

در استقرای بسامدی، به‌هر صورتی که باشد، احتمال به روال دوم به‌کار گرفته می‌شود. مثلاً، در آزمون معنی‌داری به احتمال متوسل می‌شویم تا نسبت مواردی را که در آنها فرض صفری مانند H_0 در یک نمونه‌گیری مکرر درازمدتِ مفروض رد خواهد شد، یعنی احتمال خطا، را مشخص سازیم. این تفاوت در نقش احتمال، متناظر با تفاوت در شکل استنباطی است که مقتضی انگاشته می‌شود: کاربرد قبلی احتمال معمولاً با این دیدگاه که توصیف احتمالاتی استقرا شامل کمّی‌سازیِ درجهٔ پشتیبانی یا تأیید موجود در ادعاها یا فرضهاست، رابطهٔ تنگاتنگ دارد.

برخی طرفداران رویکرد بسامدگرا، که ترجیح می‌دهند اصطلاح «رفتار استقرایی» را به‌کار برند، با این نظر موافق‌اند که نقش احتمال را در آمار بسامدی توصیف کنند. در این حال، استدلال‌کنندهٔ استقرایی «تصمیم می‌گیرد که نتیجه را استنباط کند»، و احتمال مخاطرهٔ خطای مربوط را کمّی می‌سازد. این اندیشه که یک نقش احتمال در علم عبارت است از مشخص‌کردن «مخاطره‌آمیزی» یا تأمین شواهد یا شدت آزمونهایی که فرضها در معرض آنها قرار داده می‌شوند، یادآور فلسفهٔ کارل پوپر [۲۶] است. به‌ویژه، لیمن ([۱۶]، ص ۳۲) شباهت زمانی و مفهومی اندیشه‌های پوپر و نیمن را دربارهٔ «ظرافت بخشی» به موضوع استقرا با درنظرگرفتن یک فرایند آزمون فرض به‌جای استدلال استقرایی، خاطر نشان کرده است.

این مطلب درست است که پوپر و نیمن رویکردهایی مشابه مبتنی بر این اندیشه داشته‌اند که می‌توانیم از فرضی سخن بگوییم که به یک معنا خوب آزمون شده است، و این کاملاً متمایز است با اینکه درجه‌ای از احتمال، باور، یا تأیید به آن اعطا شده باشد؛ این همان «ظرافت‌بخشی به استقرا»ست. آن دو همچنین کلاً این دیدگاه مشترک را داشتند که برای آنکه داده‌ها فرضی مانند H را «تأیید» یا «پشتیبانی» کنند، آن فرض باید در معرض آزمونی قرار گیرد و در صورت نادرست‌بودن با احتمال یا توان زیاد رد شود. اما برخلاف ارتباط نزدیک اندیشه‌ها، به‌ظاهر هیچ ارجاعی به پوپر در نوشته‌های نیمن وجود ندارد (لیمن [۱۶]، ص ۳) و ارجاعهای پوپر به نیمن اندک و به‌ندرت ذی‌ربط‌اند. افزون بر آن، چون پوپر انکار می‌کرد که هیچ‌گونه ادعای استقرایی قابل‌توجیه باشد، فلسفه‌اش وی را مجبور ساخت انکار کند که حتی روشی که از آن پشتیبانی می‌کرد (حدس و ابطال) قابل اعتماد است. پوپر روشن ساخت که اگرچه H ممکن است درست باشد، وی پشتیبانی را حداکثر به‌عنوان گزارشی از عملکرد گذشتهٔ H تلقی می‌کند: این امر هیچ ادعایی را دربارهٔ قابلیت اعتماد آن در کاربردهای آینده ایجاب نمی‌کند. برعکس، ویژگی بارز آمار بسامدی آن است که بتوان احتمال آن را که اگر فرضی نادرست باشد آزمون آن را رد کند، ارزیابی و کنترل کرد. این احتمالها از فرمول‌بندی فرایند مولد داده‌ها برحسب یک مدل آماری ناشی می‌شوند.

نیمن در سراسر تحقیقات خود براهمیت یک مدل احتمالاتی برای سامانهٔ تحت بررسی تأکید می‌کند و آمار بسامدی را به مثابهٔ مدل‌بندی پدیدهٔ پایداری بسامدهای نسبیِ نتایج «آزمایه‌های» مکرر توصیف می‌کند، در عین حال که می‌پذیرد امکانات دیگری برای مدل‌بندی پدیده‌هایِ روان‌شناختیِ مرتبط با شدتهای باور، یا مرتبط با آمادگی برای شرط‌بندی روی مبلغهای مشخص، و غیره وجود دارند، در این موارد از کارناپ [۲]، دوفینتی [۸]، و سه‌ویج [۲۷] نقل قول می‌کند. به‌ویژه نیمن از دیدگاه استنباط «بسامدی» مورد قبول کارناپ به‌خاطر نادیده گرفتن نقش اساسی مدل تصادفی پدیدهٔ مورد بررسی انتقاد می‌کند. دستاورد آماری کارناپ [۲] در ارتباط با فلسفهٔ استقرایی همان دستاورد کینز [۱۴]، و دارای تأثیری مستقیم‌تر بر کاربردهای آماری است (جفریز [۱۲]).

۳.۱ استقرا و استنباط فرضی-قیاسی

هر چند ممکن است گمان کنیم که «استنباط فرضی-قیاسی» استقرا را «ظرافت می‌بخشد»، در واقع استنباطهای استقرایی‌اند که در سراسر آزمونهای تجربی انجام می‌شوند. می‌توان ملاحظه کرد که ایده‌های آزمون آماری این شکافهای استقرایی را پر می‌کنند: اگر فرض قطعی باشد می‌توانیم تابعی ذی‌ربط از داده‌ها بیابیم که مقدارش: (i) ویژگیِ ذی‌ربطِ تحت آزمون را نمایش دهد و (ii) بتوان آن را با آن فرض پیشگویی کرد. مقدار تابع را حساب می‌کنیم و ملاحظه می‌کنیم که داده‌ها با پیشگویی موافق‌اند یا نیستند. اگر داده‌ها با پیشگویی در تعارض باشند، آنگاه یا فرض در معرض خطاست یا نوعی عامل کمکی یا پس‌زمینه‌ای دیگر را می‌توان در این بی‌هنجاری مقصر شناخت (مسألهٔ دوئم[1]).

ملاحظات آماری به دو طریق وارد بحث می‌شوند. اگر H یک فرض آماری باشد، آنگاه معمولاً هیچ برامدی اکیداً آن را نقض نمی‌کند. درنظرگرفتن داده‌ها به‌صورت ناسازگار با H صرفاً به این دلیل که آنها بسیار نامحتمل‌اند، مشکلات عمده‌ای در بردارد؛ تک‌تک برامدها در صورتی که به تفصیل تشریح

1. Duhem's problem

شوند ممکن است احتمال وقوعِ خیلی کمی داشته باشند. اما اساساً طبق نظر پوپر ([۲۶]، صص ۸۶، ۲۰۳) موضوع این است که آیا برامدی که امکاناً بی‌هنجار باشد نمایشگر اثری نظام‌مند و بازآفریدنی است یا نه.

تمرکز نظریهٔ پوپر بر ابطال به‌عنوان هدف آزمونها، و ابطال به‌مثابهٔ ملاک تعریف‌کننده برای یک نظریه یا فرض علمی، به روشنی و قویاً یادآور تفکر فیشر است. در حالی که شواهدی با تأثیر مستقیم عملاً غایب باشد، دیدگاههای پوپر با این بیان فیشر ([۹]، ص ۱۶) موافق است که می‌توان گفت هر آزمایش تنها برای آن است که شانس ردکردن فرض صفر را به واقعیات اعطا کند. اما، چون موضع پویر این است که هیچ‌گاه دلیلی برای استنباط دربارهٔ قابلیت اعتماد نخواهیم داشت، وی قبول ندارد که هرگز بتوانیم دلایلی برای استنباط راجع به انحرافهای بازآفریدنی داشته باشیم.

مزیت چارچوب آماری نوین آن است که احتمالها از تعریف مدل احتمال برای نمایش پدیدهٔ مورد نظر ناشی می‌شوند. اگر پوپر از ایده‌های آزمون آماری که تقریباً در همان زمانها ارائه شدند بهره گرفته بود، امکان داشت بتواند توصیف خود از ابطال را مستدل سازد.

موضوع دوم مربوط به این مسأله است که وقتی داده‌ها با پیشگویی «توافق» دارند چگونه باید استدلال کرد. استدلال بر این اساس که H مستلزم داده‌های y است، و اینکه y مشاهده شده است، برای رسیدن به این استنباط که H درست است، البته از لحاظ قیاسی نامعتبر است. مسأله‌ای اساسی در توصیف استقرایی عبارت است از اینکه بتوان این استنباط H را به معنایی تضمین کرد. امّا، مسألهٔ کلاسیک، حتی در حالتهای قطعی، آن است که بسیاری فرضهای رقیب (برخی خواهند گفت به تعداد بینهایت از آنها) نیز y را پیشگویی خواهند کرد، و ازاین‌رو مانند H قابل قبول خواهند بود. برای آنکه آزمونی گواه‌آور [فراهم‌آورندهٔ شواهد] باشد، پیشگویی از روی H باید چیزی باشد که اگر H نادرست و رقیبان مهم H درست باشند، در همان حال این پیشگویی به معنایی بسیار شگفت‌انگیز باشد و به‌آسانی قابل توجیه نباشد. اینک بحث می‌کنیم که چگونه شکافهای موجود در آزمون استقرایی را می‌توان با یک نوع شیوهٔ آماری خاص، یعنی آزمون معنی‌داری، پر کرد.

۲. آزمونهای معنی‌داری آماری

اگرچه آزمون معنی‌داری آماری طی بیش از ۵۰ سال با مباحثاتی همراه بوده و در باتلاق سوءتفاهم‌های موجود در نوشتگان آماری فرو رفته است، شماری از ویژگیهای اساسی دیدگاه استقرای بسامدی را که ما درنظر داریم، به شکل ساده‌ای نشان می‌دهد. برای مثال، ر.ک. [۲۱] و [۱۱]. ما بحث را تا حد امکان با عناصر اصلی آزمون معنی‌داری آغاز می‌کنیم: روایت ما از این موضوع قویاً مرتبط، امّا از جهاتی متفاوت، با هر دو رویکرد فیشری و نیمن-پیرسونی، دست‌کم در فرمول‌بندی معمولی آنهاست.

۱.۲ تذکرات کلی و تعریف

فرض می‌کنیم داده‌هایی تجربی دردست داریم که به‌طور دسته جمعی با y نمایانده شده‌اند و آنها را مقدارهای مشاهده شدهٔ متغیر تصادفی Y درنظر می‌گیریم. داده‌های y را تنها تا آنجا مدّ نظر داریم که اطلاعاتی دربارهٔ توزیع احتمال Y به‌گونه‌ای که با مدل آماری ذی‌ربط تعریف شده، فراهم می‌سازند.

این توزیع احتمال را باید نمایشی غالباً انتزاعی و یقیناً آرمانی‌شده از فرایند زیربنایی مولد داده‌ها تلقی کرد. سپس فرضی دربارهٔ توزیع احتمال داریم، که گاهی به آن فرض تحت آزمون می‌گویند اما اغلب آن را فرض صفر می‌نامند و با H_0 نشان می‌دهند. بعداً شماری از انواع کاملاً متفاوت فرضهای صفر را بیان خواهیم کرد، اما فعلاً بین آنهایی که گاهی ساده خوانده می‌شوند، به این معنی که توزیع Y را به‌طور کامل (اصولاً به‌طور عددی) مشخص می‌سازند و آنهایی که گاهی مرکب نامیده می‌شوند، به این معنی که برخی جنبه‌های معین را کاملاً مشخص می‌کنند و دیگر جنبه‌ها را نامشخص باقی می‌گذارند، تمایز قایل می‌شویم.

مثالی که از بسیاری جهات ابتدایی‌ترین مثال، هر چند مثالی تکراری، است آن است که Y متشکل از n مؤلفهٔ مستقل و هم‌توزیع است که به‌طور نرمال با میانگین μ و انحراف معیار امکاناً مجهول σ توزیع شده‌اند. اگر مقدار σ معلوم و برابر با، مثلاً، σ_0، باشد و فرض صفر آن باشد که μ برابر با μ_0، یعنی ثابتی مفروض است، فرضی ساده حاصل می‌شود. یک فرض مرکب در همان چارچوب بالا ممکن است شامل σی مجهول باشد و باز هم مقدار μ را مشخص کند.

توجه کنید که در این فرمول‌بندی لازم است جنبه‌ای مجهول از توزیع، نوعاً یک یا چند پارامتر مجهول، دقیقاً مشخص شود. فرض اینکه، مثلاً، $\mu \leq \mu_0$، یک فرمول‌بندی پذیرفتنی برای فرض صفر در آزمون فیشری نیست؛ در حالی که این صورت کلی‌ترِ فرض صفر در فرمول‌بندی‌های نیمن-پیرسون مجاز است.

هدف فوری عبارت است از آزمودن سازگاری داده‌های خاصِ تحت تحلیل با H_0، از جنبه‌ای که قرار است مشخص شود. برای انجام این کار تابعی مانند $t = t(y)$ از داده‌ها می‌یابیم، که آمارهٔ آزمون نامیده می‌شود، به‌گونه‌ای که

- هر چه مقدار t بزرگ‌تر باشد داده‌ها با H_0 ناسازگارترند؛
- وقتی H_0 درست باشد، متغیر تصادفی متناظر $T = t(Y)$ دارای یک توزیع احتمالِ (به‌طور عددی) معلوم است.

این دو شرط با شرطهای قطعی متناظر هم‌ترازند. برای ارزیابی اینکه یک ناهمخوانی واقعی با H_0 (یا انحراف بازآفریدنی از H_0) وجود دارد یا نه، به‌اصطلاح p-مقدار متناظر با هر t را به‌صورت

$$p = p(t) = P(T \geq t; H_0)$$

تعریف می‌کنیم، که شاخصی از ناهمخوانی با H_0 از جنبهٔ مورد آزمون تلقی می‌شود. دست‌کم در فرمول‌بندی آغازین، فرضهای مقابل در لایه‌های زیرین پنهان‌اند اما از جنبهٔ احتمالاتی به‌طور صریح فرمول‌بندی نمی‌شوند؛ همچنین پیشاپیش هیچ پرسشی دربارهٔ تعیین یک مقدار آستانه‌ایِ از پیش تعیین شده و «رد» H_0 اگر و تنها اگر $p \leq \alpha$، مطرح نیست. افزون بر آن، برای توجیه آزمونها صرفاً به رفتار درازمدت متوسل نخواهیم شد بلکه یک ملاک عقلی استنباطی یا مبتنی‌بر شواهد را شناسایی خواهیم کرد. اینک در این باره بیشتر توضیح می‌دهیم.

۲.۲ رفتار استقرایی در برابر استنباط استقرایی

این استدلال را می‌توان روایتی آماری از استدلالِ موسوم به برهان خلف درنظر گرفت که در منطق قیاسی معتبر است. در این روش، انکار یک فرض H از ترکیب گزارهٔ H مستلزم E است با این اطلاع که E نادرست است، استنباط می‌شود. چون اگر $H_{\circ}$ درست می‌بود احتمال زیادِ $(p-1)$ وجود می‌داشت که نتیجه‌ای کمتر معنی‌دار رخ دهد، می‌توانیم انتخاب pـ مقدارهای کوچک را که درست حساب شده باشند، به‌عنوان شواهدی علیه $H_{\circ}$ توجیه کنیم. چرا؟ به دو دلیل عمده:

اولاً چنین قاعده‌ای باعث می‌شود که در درازمدت، وقتی $H_{\circ}$ درست باشد، نرخهای خطا (یعنی، ردکردن‌های نادرست) کوچک باشند، که استدلالی رفتارگرایانه است. هماهنگ با نگرش مبتنی‌بر ارزیابی خطا به آمار، می‌توانیم به هر مقدار خاص مثلاً p، تفسیر فرضی زیر را نسبت دهیم: فرض کنید که قرار است داده‌ها را تنها شواهد قاطع علیه $H_{\circ}$ قلمداد کنیم. دراین‌صورت، در تکرارهای فرضی، $H_{\circ}$ به‌نسبت درازمدت p از مواردی که در آنها واقعاً درست است، رد خواهد شد. اما، آگاهی از این احتمالهای خطای فرضی را می‌توان برای حمایت از توجیهی متمایز درنظر گرفت.

بدین‌گونه است که چنین قاعده‌ای راهی را برای تعیین اینکه مجموعه‌ای از داده‌های خاص شواهدی از ناهمخوانی با $H_{\circ}$ است یا نه، پیش پای، می‌گذارد.

به‌ویژه، یک pـ مقدارکوچک، به شرط آنکه درست حساب شود، شواهدی از اختلاف با $H_{\circ}$ را از حیث جنبهٔ مورد امتحان به دست می‌دهد، در حالی که pـ مقداری که کوچک نباشد شواهدی از همخوانی یا سازگاری با $H_{\circ}$ را در اختیار می‌گذارد (که در اینجا بین این حالت و شواهد مثبت به نفع $H_{\circ}$ آن‌گونه که در زیر در بخش ۲ـ۳ بررسی می‌شود، باید فرق گذاشت). در کاربردها توجه ما نوعاً به این است که آیا p در دامنه‌ای مانند $p \geq ۰٫۱$ واقع است یا نه، که می‌توان آن را به مثابهٔ همخوانی معقول با $H_{\circ}$ از جنبهٔ مورد آزمون تلقی کرد، یا اینکه p نزدیک به عددهای قراردادی از قبیل ۰٫۰۵، ۰٫۰۱، ۰٫۰۰۱ است یا نه. شیوهٔ معمول در اکثر کاربردها آن است که مقدار مشاهده شدهٔ p را به‌صورت تقریبی اعلام کنند. مقدارکوچک p بیانگر یکی از موارد زیر است: (i) $H_{\circ}$ نادرست است (اختلافی با $H_{\circ}$ وجود دارد) یا (ii) اساس آزمون آماری معیوب است، اغلب از این نظر که خطاهای واقعی کم‌براورد شده‌اند، مثلاً به‌دلیل مفروضات نامعتبر استقلال، یا (iii) نقش شانس فوق‌العاده بوده است.

بخشی از هدف طرح پژوهشی خوب و انتخاب مناسب روش تحلیل این است که با کسب اطمینان از اینکه ارزیابیهای خطا مناسب‌اند، از شقّ (ii)ی بالا پرهیز شود.

به هیچ وجه قصد آن نیست که بگوییم آزمون معنی‌داری نوعاً تنها تحلیلی خواهد بود که گزارش می‌شود. در واقع، یک اصل بنیادی مفهوم یادگیری استقرایی که بیشترین سازگاری را با فلسفهٔ بسامدگرا دارد آن است که استنباط استقرایی لازم می‌نماید که استدلالها و استنباطهای قاطع از طریق به‌هم پیوستن چند خرده نتیجهٔ متفاوت انجام شود. هر چند به‌خاطر پیچیدگی داستان، توصیف شسته رفته‌ای از آن، مثلاً چنانکه گویی تک‌الگوریتمی یگانه می‌تواند کل استنباط استقرایی را نمایش دهد، بسیار مشکل است، حاصل کارگزارشی است که به نوع استدلالهای نابی که دانشمندان به قصد کسب دانش قابل اعتماد و درک یک رشتهٔ علمی بنا می‌نهند نزدیک می‌شود.

در میان این پیچیدگی، استدلال آزمون معنی‌داری مفهومی نسبتاً سرراست از ارزیابی شواهد ناهمخوان با $H_{\circ}$ را در زمینه‌ای آماری بازمی‌تاباند، مفهومی که شاید پوپر در ذهن خود داشت اما فاقد ابزارهایی برای تحقق بخشیدن به آن بود. ایدهٔ اساسی در اینجا آن است که احتمالهای خطا را می‌توان برای ارزیابی «مخاطره‌آمیز بودن» پیشگوییهایی که لازم است $H_{\circ}$ در آنها صدق کند به‌کار بست، این کار با سنجش قابلیت اعتماد آزمون در تمایز قائل شدن بین اینکه فرایند واقعی مولد داده‌ها با فرایند توصیف شده در $H_{\circ}$ همخوان است یا نه، انجام می‌شود. آگاهی از این ظرفیتِ گواه‌آوری امکان می‌دهد که تعیین کنیم آیا شواهد قوی مبنی‌بر ناهمخوانی وجود دارند یا نه. برای تعیین اینکه شواهدی علیه $H_{\circ}$ وجود دارند یا نه، استدلال برپایهٔ اصل بسامدگرایانهٔ زیر بنا نهاده می‌شود.

شواهد بسامدگرا (ش‌ب) (i). داده‌های y شواهدی (قوی) علیه $H_{\circ}$ اند، یعنی شواهدی (قوی) مبنی بر وجود اختلاف با $H_{\circ}$ اند اگر و تنها اگر هنگامی که $H_{\circ}$ توصیفی درست از سازوکار مولد y باشد، آنگاه با احتمال زیاد، این سازوکار نتیجه‌ای بدهد که کمتر از آنچه به‌وسیلهٔ y نشان داده می‌شود، ناهمخوان باشد.

یک فرع **ش‌ب** آن است که y شواهدی (قوی) علیه $H_{\circ}$ نیست اگر احتمال نتیجه‌ای ناهمخوان‌تر خیلی کم نباشد، حتی اگر $H_{\circ}$ درست باشد. یعنی، اگر احتمالی نسبتاً زیاد برای نتیجه‌ای ناهمخوان‌تر وجود داشته باشد، حتی اگر $H_{\circ}$ درست باشد، آنگاه $H_{\circ}$ از جنبهٔ مورد آزمون با y همخوانی دارد.

مطلبی که قدری بحث‌انگیزتر است تفسیر شکست در یافتن pـ مقداری کوچک است؛ امّا برپایهٔ صورت فوق از **ش‌ب** می‌توان شرح و تفسیری رسا برای آن عرضه کرد.

۳.۲ شکست و تأیید

مشکل درنظرگرفتن مقداری متوسط از p به مثابهٔ شاهدی به نفع $H_{\circ}$، آن است که حتی اگر رقیبان $H_{\circ}$ که جداً با $H_{\circ}$ تفاوت دارند درست باشند، همخوانی بین $H_{\circ}$ و y ممکن است رخ دهد. این موضوع به‌ویژه وقتی حاد است که اندازهٔ داده‌ها محدود باشد. اما، گاهی می‌توانیم شواهدی به‌نفع $H_{\circ}$ بیابیم، که به‌مثابهٔ ادعایی درک می‌شود مبنی براینکه اختلافی، عیبی، یا خطایی در بین نیست، و این کار را می‌توانیم به‌وسیلهٔ آزمونهایی انجام دهیم که با احتمال زیاد، در صورت وجود اختلاف، آن را نشان می‌دادند. همان‌قدرکه نام نیمن با فنون تصمیم‌گونهٔ خودکار همراه است، دست‌کم در عمل، هم او و هم پیرسون، انتخاب مقتضی احتمال خطاها را بازتاب‌دهندهٔ بافتار ویژهٔ مورد نظر تلقی می‌کردند. (نیمن [۲۳]، پیرسون [۲۴]).

در اینجا دو موضوع متفاوت مطرح است. یکی اینکه آیا مقداری خاص از p باید به مثابهٔ آستانه‌ای در هرکاربرد به کار رود یا نه. این همان شیوه‌ای است که اگر نه در همه ولی در اکثر شرحهای رسمی از نظریهٔ نیمن‌ـ پیرسون بیان می‌شود. موضوع دوم آن است که آیا کنترل نرخ خطاهای درازمدت، توجیهی برای آزمونهای بسامدگراست یا آیا توجیه نهایی آزمونها در نقش آنها

در تفسیر شواهد موجود در موارد خاص نهفته است یا نه. در شرحی که در اینجا عرضه می‌شود، مقدار تحقق‌یافتهٔ p، دست کم به تقریب، گزارش می‌شود، و شرحِ مبتنی‌بر «بپذیر-ردکن» صرفاً شرحی فرضی است که به p تفسیری عملیاتی ببخشد. مشهور است که پیرسون [۲۴] از تفسیر رفتارگرای تنگ‌نظرانه پشتیبانی نمی‌کرد (مایو [۱۷]). نیمن نیز، دست کم در بحث خود با کارناپ (نیمن [۲۳])، به‌نظر می‌رسد که به تمایزی بین تفسیرهای رفتاری و استنباطی اشاره دارد.

در این بحث، نیمن درکوششی برای روشن ساختن ماهیت آمار بسامدگرا، نگران اصطلاح «درجهٔ تایید» بود که کارناپ به‌کار برده است. نیمن در چارچوب مثالی که در آن یک آزمون بهینه نتوانسته بود H_0 را «رد کند»، به بررسی این مطلب پرداخت که آیا این آزمون H_0 را «تایید» کرده است یا نه. وی تذکر داد که این امر بستگی به‌معنای واژه‌هایی نظیر «تأیید» و «اطمینان» دارد و در بافتاری که در آن H_0 «ردنشده» بوده است تلقی آن به مثابهٔ تأیید H_0 در صورتی که این آزمون در واقع شانس اندکی برای کشف اختلافی مهم با H_0 داشته باشد، حتی اگر آن اختلاف وجود داشته باشد، «خطرناک» است. از سوی دیگر، اگر این آزمون توان قابل توجهی می‌داشت که آن اختلاف را کشف کند وضعیت «از بیخ و بن متفاوت» می‌بود.

نیمن بر مغلطه‌ای مرتبط با «نتایج منفی» تأکید می‌کند به این مضمون که اگر داده‌های y یک نتیجهٔ آزمونی به‌دست دهند که از لحاظ آماری تفاوت معنی‌داری با H_0 (مثلاً، فرض صفر «بی‌اثر» بودن) نداشته باشد، و با این حال آزمون مذکور احتمال کوچکی برای رد H_0، حتی زمانی که اختلاف جدی در بین است، داشته باشد آنگاه y شواهد خوبی برای استنباط اینکه H_0 از سوی y تایید می‌شود، نیست. طبق این استدلال، تنها اگر شانس آنکه آزمون به‌درستی توانسته باشد اختلافی را کشف کند زیاد باشد، می‌توان از نبودِ اختلاف مطمئن بود.

نیمن این وضعیت را با تفسیرهای مقتضی برای رفتار استقرایی مقایسه می‌کند. در اینجا تأیید و اطمینان را می‌توان برای توصیف انتخاب عمل به‌کار بست، مثلاً در خودداری از اعلام یک کشف یا تصمیمی مبنی بر اینکه H_0 را رضایت‌بخش درنظر بگیریم، منطق کار، منطق رفتارگرای عملگرایانه دربارهٔ کنترل خطاها در درازمدت است. این تمایز بدان معنی است که حتی درنظر نیمن، برای شواهد مربوط به تصمیم‌گیری ممکن است ملاکی متمایز از شواهد مربوط به باورداشتن لازم باشد، اما متاسفانه نیمن شق دوم را به صراحت تشریح نکرد. ما این نظر را در پیش می‌نهیم که اصل شواهدیِ مورد نیاز اقتباسی از **ش ب (i)** برای حالتی است که p-مقدار کوچک نیست:

ش ب (ii). مقدار متوسطی از p-مقدار شاهدی از نبودِ اختلاف بین δ و H_0 است تنها اگر احتمال زیادی وجود داشته باشد که آزمون، برازشی بدتر به H_0 (یعنی، p-مقدار کوچک‌تری) به‌دست می‌داد در صورتی که اختلافی مانند δ وجود می‌داشت. **ش ب (ii)** مخصوصاً در بافتار فرضهای «نشانده‌شدهٔ» (زیر) پیش می‌آید.

آنچه نوع استدلال فرضی را به حالت مورد نظر مربوط می‌سازد صرفاً یا عمدتاً مقدار پایین نرخ خطاهای درازمدتِ وابسته به استفاده از این ابزار (یا آزمون) با این روال نیست؛ بلکه چیزی است که آن نرخهای خطا دربارهٔ منبع یا پدیدهٔ مولد داده‌ها آشکار می‌سازند. محاسبات مبتنی‌بر خطا اطمینان خاطری می‌بخشند که از تفسیرهای نادرست شواهد در حالت خاص پرهیز می‌شود. برای تشخیص تفاوت بین این توجیه «شواهدیِ» استدلال آزمونهای معنی‌داری، و توجیه «رفتارگرا»، ممکن است بررسی یک مثال غیررسمی از کاربرد این استدلال در «حالت خاص» سودمند باشد. مثلاً، فرض کنید که افزایش وزن به‌وسیلهٔ روشهای کاملاً کالبیده[1] و پایدار، در صورت امکان با استفاده از چند ابزار اندازه‌گیری و مشاهده‌گر، اندازه گرفته می‌شوند و نتایج طی دورهٔ زمانی موردنظر آزمون، تغییر ناچیزی را نشان می‌دهند. این امر را می‌توان به مثابهٔ دلیلی برای این استنباط تلقی کرد که افزایش وزن افراد در محدوده‌ای که توسط حساسیت ترازوها تعیین می‌شود چشم پوشیدنی است. چرا؟

هر چند شخص با پیروی از چنین شیوه‌ای در درازمدت به‌ندرت افزایش وزنها را نادرست گزارش خواهد کرد، این تنها دلیل منطقی برای این استنباط خاص نیست. بلکه توجیه آن است که خواص احتمالاتیِ خطایِ شیوهٔ توزین، آنچه را که واقعاً در این مورد خاص مطرح است نشان می‌دهد. (باید بین این تفسیر و تفسیر شواهدیِ نظریهٔ نیمن-پیرسون که برنباوم [۱] آن را پیشنهاد کرده، و داده‌وابسته است، فرق گذاشت.)

آزمون معنی‌داری یک ابزار اندازه‌گیری برای همخوانی با فرضی ویژه است که، همچون ابزارهای اندازه‌گیری به‌طور عام، با عملکردش درکاربردهای مکرر، در این مورد نوعاً به‌طور نظری یا با شبیه‌سازی، کالبیده می‌شود. درست همانند کاربست ابزارهای اندازه‌گیری، که در موردی خاص به‌کار برده می‌شوند، ویژگیهای عملکرد را برای انجام استنباط دربارهٔ جنبه‌های چیز خاصی که اندازه‌گیری می‌شود، جنبه‌هایی که ابزار اندازه‌گیری به‌طور مقتضی قادر به آشکارسازی آنهاست، به‌کار می‌گیریم.

البته برای آنکه این وضع برقرار باشد، محاسبات درازمدت احتمالاتی باید برای مورد تحت بررسی مناسب و عملی باشند. اجرای این موضوع در نظریهٔ آمار در بحثهای استنباط شرطی، انتخاب توزیع مقتضی برای تعیین مقدار p، ظاهر می‌شود. به‌نظر می‌رسد مشکلاتی که این امر به همراه دارد بیشتر فنی‌اند تا مفهومی، و در اینجا به آنها نمی‌پردازیم جز آنکه تذکر دهیم که اقدام به‌کاربست (یا کوشش برای به‌کار بستن) **ش ب** می‌تواند به مشخص‌کردن آزمون مناسب کمک کند.

۳. انواع فرض صفر و استنباطهای استقرایی متناظر آنها

در تحلیلِ آماری داده‌های علمی و فن‌شناختی، تقریباً همیشه اطلاعاتی بیرونی وجود دارند که باید وارد بحث شوند تا از آنچه داده‌ها دربارهٔ پرسش اولیهٔ مورد نظر نشان می‌دهند، نتیجه‌گیری شود. معمولاً این ملاحظاتِ پس‌زمینه‌ای نه از طریق تخصیص احتمال بلکه از راه شناسایی پرسشی که باید مطرح شود، طراحی بررسی، تفسیر نتایج آماری، و ربط‌دادن آن استنباطها به استنباطهای علمی اولیه و کاربست آنها برای گسترش و پشتیبانی نظریهٔ زیربنایی وارد بحث می‌شوند. برای آنکه ابزارها به‌گونه‌ای عاری از مغلطه و طبق منظور به‌کار روند باید داوریهایی دربارهٔ اینکه چه چیزی مناسب و آگاهی بخش است در اختیار گذاشته شود. با وجود این، کاربست‌های نظام‌مندی وجود دارند که می‌توان آنها را متناظر با انواع آزمون و انواع فرضهای صفر، تشریح کرد.

1. well-calibrated

۱.۳ انواع فرض صفر

اکنون شماری از انواع فرض صفر را تشریح می‌کنیم. در بحث زیر، بحثی را که کاکس [(۴]، [۵]) و کاکس و هینکلی [۶] عرضه کرده‌اند گسترش می‌دهیم. هدف ما در اینجا آن نیست که راهنمایی برای مجموعهٔ همهٔ بافتارهایی که پژوهشگر ممکن است با آنها روبه‌رو شود، ارائه کنیم، بلکه برخی تفسیرهای متفاوت از نتایج آزمون و pـ مقدارهای مربوط را توضیح می‌دهیم. در بخش ۴.۳، تفسیر ژرف‌تر استنباطهای استقرایی متناظر را که، به‌نظر ما، با استدلال pـ مقداری مجازند (و مجاز نیستند)، بررسی می‌کنیم.

۱. فرضهای صفر نشانده‌شده. در این مسأله‌ها نه تنها یک مدل احتمال برای فرض صفر فرمول‌بندی می‌شود، بلکه مدلهای دیگری نیز فرمول‌بندی می‌شوند که امکانهای دیگری را نمایش می‌دهند که در آنها فرض صفر نادرست است، و بنابراین، معمولاً امکانهایی را نمایش می‌دهند که اگر وجود داشته باشند مایلیم آنها را کشف کنیم. در بین شماری از وضعیتهای ممکن، در معمول‌ترین آنها خانواده‌ای پارامتری از توزیعها وجود دارد که با پارامتری مجهول مانند θ اندیس‌گذاری می‌شود؛ θ به مؤلفه‌های $\theta = (\phi, \lambda)$ افراز می‌شود به‌گونه‌ای که فرض صفر عبارت باشد از $\phi = \phi_0$، و λ یک پارامتر مزاحم مجهول است و دست‌کم در بحث آغازین، ϕ تک‌بُعدی است. توجه معطوف به فرضهای مقابل $\phi > \phi_0$ است.

فرمول‌بندی بالا این مزیت فنی را دارد که تا حد زیادی آمارهٔ آزمون مناسب $t(y)$ را با این شرط که حساس‌ترین آزمون ممکن با توجه به داده‌هایِ در دست باشد، تعیین می‌کند.

دو روایت متفاوت از فرمول‌بندی بالا وجود دارد. در یکی، خانوادهٔ کامل، یک فرمول‌بندی موقتی است که نه‌چندان به‌عنوان پایه‌ای ممکن برای تفسیر نهایی بلکه بیشتر به‌عنوان ابزاری برای تعیین یک آمارهٔ آزمون مناسب درنظر گرفته می‌شود. یک مثال، استفاده از مدلی درجهٔ دوم برای آزمون‌کردن رسایی یک رابطهٔ خطی است؛ به‌طور کلی رگرسیونهای چندجمله‌ای پایه‌ای ضعیف برای تحلیل‌نهایی‌اند اما برای کشف انحرافهای کوچک از یک صورت مفروض، مناسب و قابل تفسیرند. در حالت دوم، خانوادهٔ مذکور پایه‌ای محکم برای تفسیر است. بازه‌های اطمینان برای ϕ تفسیری معقول دارند.

یک امکان دیگر، که خیلی به‌ندرت پیش می‌آید، آن است که یک فرض صفر ساده و یک فرض مقابل سادهٔ تک وجود دارد، یعنی تنها دو توزیع ممکن تحت بررسی‌اند. اگر این دو فرض برپایه‌ای برابر درنظر گرفته شوند، بهتر است این تحلیل به‌عنوان یک تحلیل تشخیصی فرضی یا واقعی در نظر گرفته شود یعنی تعیین اینکه کدام یک از دو امکان (یا به‌طور کلی‌تر کدام یک از چند امکان محدود) مناسب‌تر است اگر با امکانهای مختلف برپایه‌ای از لحاظ مفهومی برابر رفتار شود.

در این حالت دو رویکرد کلی وجود دارند. یکی کاربستِ نسبت درستنمایی به‌عنوان شاخصی از برازش نسبی است، که امکان دارد همراه با کاربردی از قضیهٔ بیز باشد. دیگری، که با رویکرد احتمال خطا همخوانی بیشتری دارد، آن است که هر مدل را به‌نوبت به مثابهٔ فرض صفر و دیگری را به‌عنوان فرض مقابل می‌گیرد و منجر به ارزیابی این امر می‌شود که آیا داده‌ها همخوان با هر دو فرض، یکی از آنها، یا هیچ یک از آنهاست. با کاربست ش ب در این حالت، وقتی که در چارچوب نیمن-پیرسون درآمده باشد، اساساً یک تفسیر نتیجه می‌شود.

این سه حالت را به‌ترتیب خانواده‌ای صوری از فرضهای مقابل، خانواده‌ای خوش‌بنیاد از فرضهای مقابل، و خانواده‌ای از امکانهای گسسته می‌نامیم.

۲. تقسیم فرضهای صفر. در بسیاری از موارد، مخصوصاً درکاربردهای فن‌شناختی اما نه تنها در آنها، کانون توجه ما مقایسهٔ دو یا چند شرط، فرایند یا تیمار است بدون هیچ دلیل خاصی برای اینکه انتظار داشته باشیم آنها دقیقاً یا تقریباً همانند باشند، مثلاً دارویی جدید در مقایسه با دارویی استاندارد ممکن است نرخهای بقا را افزایش یا کاهش دهد.

در واقع، شخص دو آزمون را ترکیب می‌کند، نخست برای امتحان این امکان که مثلاً $\mu > \mu_0$، دیگری برای $\mu < \mu_0$. در این حالت، این آزمون دوطرفه هر دو آزمون یک طرفه را، هر یک با سطح معنی‌داری خود، ترکیب می‌کند. سطح معنی‌داری به‌دلیل «اثر گزینش»، دوبرابر pی کوچک‌تر است (کاکس و هینکلی [۶]، ص ۱۰۶). در بخش ۴ به این موضوع برمی‌گردیم. در این حال فرض صفرِ مبنی‌بر تفاوتِ صفر، وضعیتهای ممکن را به‌دو ناحیهٔ کیفیتاً متفاوت نسبت به ویژگی مورد آزمون تقسیم می‌کند، وضعیتهایی که در آنها یکی از تیمارها نسبت به دیگری برتر است و وضعیتهایی که در آنها همان تیمار بدتر است.

۳. فرضهای صفرِ مبنی بر نبودِ ساختار. در بسیاری از پژوهشهای کمابیش تجربی در رشته‌هاییَ که پایهٔ نظری محکمی ندارند، داده‌ها به امید یافتن ساختاری، اغلب به‌صورت وابستگیهایی بین ویژگیهای ناشناخته، گردآوری می‌شوند. این کار در همه‌گیرشناسی به‌صورت آزمونهایی از عاملهای خطر بالقوه برای بیماری‌ای که علّت ناشناخته دارد، در می‌آید.

۴. فرضهای صفر مبتنی‌بر رسایی مدل. حتی در حالت کاملاً نشانده شده که در آن خانواده‌ای کامل از توزیعهای تحت بررسی وجود دارند و این خانواده‌ها به‌طور بالقوه آنقدر غنی‌اند که خواه فرض صفر درست یا غلط باشد داده‌ها را تبیین می‌کنند، این امکان وجود دارد که اختلافی مهم با مدل وجود داشته باشد که گسترش، اصلاح، یا جانشین‌کردن کامل مدل به‌کاررفته برای تفسیر را توجیه کند. در بسیاری از رشته‌ها، مدلهای آغازین به‌کار رفته برای تفسیر کاملاً آزمایشی‌اند؛ در دیگر رشته‌ها، به‌خصوص در برخی زمینه‌های فیزیک، مدلها پایه‌ای محکم در نظریه و آزمایشهای گسترده دارند. اما در همهٔ موارد اگرچه به‌صورت غیررسمی، امکان مشخص‌کردن نادرست مدل را باید پذیرفت.

در این صورت انتخابی نه‌چندان آسان بین یک آمارهٔ آزمون نسبتاً خاص‌نگر که نسبت به انواع خاص نارسایی مدل (جهتهای انحراف خاص) حساس است، و آزمونهای موسوم به کلی‌نگر که دربارهٔ ماهیّت انحرافها مفروضات قوی ندارند، مطرح است. روشن است که آزمونهای اخیر نسبت به فرضهای مقابل خاص ناحساس، و اغلب فوق‌العاده ناحساس، خواهند بود. این دونوع به‌طور کلی با آزمونهای خی‌دو با درجه‌های آزادی کم و زیاد متناظرند. برای آزمون خاص‌نگر می‌توانیم یا آمارهٔ آزمونی مناسب، یا تقریباً به‌طور هم‌ارز، خانواده‌ای تصوری از فرضهای مقابل را برگزینیم. مثلاً برای امتحان توافق n مشاهدهٔ مستقل با یک توزیع پواسون در واقع می‌توانیم توافق واریانس نمونه با میانگین نمونه را به‌وسیلهٔ یک آزمون پراکنش خی‌دو (یا معادل دقیق آن) آزمون کنیم یا توزیع پواسون را مثلاً در یک خانوادهٔ دوجمله‌ای منفی بنشانیم.

۵. فرضهای صفر پرمایه. در برخی بافتارهای خاص، نتایج مبتنی برفرض صفر ممکن است شواهدی جدی برای ادعاهای علمی در بافتارهایی باشند

که شایستهٔ قرارگرفتن در یک رستهٔ پنجم‌اند. در اینجا، نظریه‌ای مانند $\mathcal{T}$ که برای آن شواهد نظری و/یا تجربی قابل ملاحظه‌ای وجود دارند، پیشگویی می‌کند که H_0، دست کم با تقریب بسیار خوبی، وضعیتِ واقعی است.

(الف) در یک روایت، ممکن است نتایجی به‌ظاهر بی‌هنجار برای $\mathcal{T}$ وجود داشته باشند، و آزمونی طراحی می‌شود که مجال کافی برای آشکار ساختن ناهمخوانی با H_0 در صورتی که این نتایج بی‌هنجار، واقعی باشند، داشته باشد.

(ب) در روایت دوم، نظریه‌ای رقیب مانند $\mathcal{T}^*$ ناهمخوانی مشخصی با H_0 را پیشگویی می‌کند و آزمون معنی‌داری برای تشخیص بین $\mathcal{T}$ و آن نظریهٔ $\mathcal{T}^*$ (در قلمرویی که تاکنون آزمون نشده) طراحی می‌شود.

به‌عنوان مثالی از (الف)، نظریهٔ فیزیکی حاکی از آن است که چون کوانتوم انرژی در میدانهای برقاطیسی غیریونده، مانند انرژیهای حاصل از خطهایِ انتقال فشارقوی، بسیار کمتر از حد لازم برای شکستن پیوند ملکولی است، نباید واقع شدن در معرض این‌گونه میدانها اثر سرطان‌زایی داشته باشد. بدین‌ترتیب، در یک آزمایش تصادفیده که در آن دو گروه موش تحت شرایط همانند هستند جز آنکه یک گروه در معرض چنین میدانی قرار داده می‌شود، فرض صفر مبنی بر اینکه نرخهای شیوع سرطان در این دو گروه یکسان‌اند ممکن است کاملاً درست باشد و در تحلیل داده‌ها در درجهٔ اول اهمیت و توجه باشد. البته فرض صفری از این نوع کلّی لازم نیست که مدلی با اثر صفر باشد؛ ممکن است اشاره به توافقی با یافته‌های تجربیِ کاملاً اثبات‌شدهٔ قبلی یا نظریه‌ای داشته باشد.

۲.۳ برخی نکته‌های کلی

در بالا اساساً آزمونهای یک‌طرف را توصیف کردیم. گسترش بحث به آزمونهای دوطرفه برخی مسائل را در زمینهٔ تعریف پیش می‌آورد اما در اینجا دربارهٔ آنها بحث نخواهیم کرد.

چند نوع از فرضهای صفر مستلزم مشخص کردن احتمال به‌طور ناقص‌اند. یعنی ممکن است تنها فرض صفر به روشنی مشخص شده باشد. می‌توان استدلال کرد که در فرمول‌بندی کامل احتمال باید همیشه هم فرض صفر و هم امکانات فرضهای مقابلِ شدنی را درنظر گرفت. این امر شاید از لحاظ اصولی معقول به‌نظر برسد اما به‌مثابهٔ راهبردی که مستقیماً به‌کار برده شود، اغلب شدنی نیست؛ در هر حالت، مدلهایی که همهٔ امکانات معقول را دربر خواهند گرفت باز هم ناکامل خواهند بود و با اثرهای جانبی زیان‌بخش خود حتی مسأله‌های ساده را پیچیده‌تر خواهند ساخت.

اما توجه داشته باشید که در همهٔ فرمول‌بندی‌هایی که در اینجا به‌کار گرفته شده‌اند نوعی مفهوم تبیین داده‌ها به‌صورت بدیلی برای فرض صفر از راه انتخاب آمارهٔ آزمون در بین است؛ موضوع این است که چه هنگام این انتخاب از طریق یک فرمول‌بندی احتمالاتی صریح صورت می‌گیرد. اصل کلّی شواهد **ش ب** کمک می‌کند بفهمیم که در بافتارهایی مشخص، حالت قبلی برای انجام یک ارزیابی شواهدی کفایت می‌کند (رک. بخش ۳.۳).

اما، گاهی استدلال می‌شود که انتخاب آمارهٔ آزمون را می‌توان برپایهٔ توزیع داده‌ها تحت فرض صفر به تنهایی بنا نهاد، در واقع با انتخاب منفیِ لگاریتم احتمال به‌عنوان آمارهٔ آزمون، و بدین‌ترتیب با جمع بستن احتمالها روی همهٔ نقاطی از نمونه که به اندازهٔ مقادیر مشاهده شده یا کمتر از آنها محتمل‌اند. هر چند این کار اغلب به نتایج معقول می‌انجامد در اینجا این مسیر را دنبال نخواهیم کرد.

۳.۳ استنباطهای استقرایی برپایهٔ برامدهای آزمونها

استدلال براساس آزمون معنی‌داری چگونه استنباطهای استقرایی یا ارزیابیهایِ شواهدی را در حالتهای گوناگون تأیید می‌کند؟ تفسیر عملیاتیِ فرضیِ pـمقدار روشن است اما پیامدهای ژرفترِ مربوط به یکی از دومقدار متوسط یا کوچک pـمقدار چیست‌اند؟ این پیامدها به‌شدت به (i) نوع فرض صفر، و (ii) ماهیت انحراف یا فرض مقابل تحت بررسی، وابسته‌اند و همچنین (iii) به اینکه آیا به تفسیر مجموعه‌هایی خاص از داده‌ها علاقه‌مندیم، چنانکه در اکثر کارهای آماری تفصیلی پیش می‌آید، یا اینکه مدلی کلی برای تحلیل و تفسیر در یک رشتهٔ علمی را بررسی می‌کنیم. حالت اخیر به فرمول‌بندی سنتی نیمن-پیرسون مبنی بر تثبیت یک سطح بحرانی و به‌معنایی، پذیرفتن H_0 اگر $p > \alpha$ و ردکردن H_0 در غیر این‌صورت، نزدیک است. برخی نواقص آشنای کاربستِ طبق عادت یا مکانیکیِ pـمقدار را بررسی می‌کنیم.

۴.۳ کاربستِ عادت‌وارِ pـمقدارها

تصورکنید که $\alpha = 0.05$ گرفته می‌شود و نتایج در صورتی به مقاله‌ای قابل انتشار می‌انجامد که تنها به‌ازای pی ذی‌ربط، داده‌ها $p < 0.05$ را نتیجه دهند. منطق کار از نوع منطق رفتارگراست که پیش‌تر توضیح داده شد. اما در اکثریت عمدهٔ بحثهای آماری در این زمینه، که سابقه‌اش به‌کار تیتس [۳۲] و قبل از آن برمی‌گردد، چنین رویکردی محکوم می‌شود، هم به‌واسطهٔ این نگرانی که شیوه‌های مکانیکی، خودکار، و بدون تفکر را تشویق می‌کند، و هم به‌خاطر تأکید این رویکرد بر براوردِ اثرهای ذی‌ربط علاوه بر آزمون فرضها. در واقع چند مجلهٔ علمی در برخی رشته‌ها عملاً استفاده از pـمقدار را ممنوع کرده‌اند. در دیگر رشته‌ها، مانند تعدادی از زمینه‌های همه‌گیرشناسی، معمول آن است که بر بازه‌های اطمینان ۹۵ درصدی تأکید شود که با جریان اصلی بحث آماری هماهنگ است. البته، این کار شخص را از نیاز به ارائهٔ توضیح مقتضی بسامدگرایانه دربارهٔ کاربست و تفسیر سطحهای اطمینان، که در اینجا ارائه نمی‌کنیم، نمی‌رهاند (رک. بخش ۶.۳).

با وجود این، کاربستِ نسبتاً مکانیکی pـمقدارها، در عین حال که در معرض ریشخند است، خیلی دور از عمل رایج در برخی رشته‌ها نیست؛ و این روش به‌عنوان یک ابزار غربالگری، با به رسمیت‌شناختن امکان خطا، وکاستن از امکان انتشار نتایج گمراه‌کننده، به‌کار گرفته می‌شود. یک نقش مشابه آزمونها در کار بنگاههای ناظر، به‌ویژه FDA (بنگاه دارویی فدرال آمریکا)، دیده می‌شود. ملزم ساختن بررسیها به نشان دادن مقداری از p که کوچک‌تر از یک سطح تخصیص یافته به‌وسیلهٔ آزمونی از پیش‌تعیین شده باشد، ممکن است باعث انعطاف‌ناپذیری شود، و انتخاب سطح بحرانی دلخواه باشد. با این حال، چنین شیوه‌هایی دارای محاسن بی‌طرفی و استقلال نسبی از دستکاریهای نامعقول‌اند. هر چند پایبندی به یک pـمقدار ثابت ممکن است باعث سوگیری نوشتگان به سمت نتیجه‌گیری‌های مثبت شود، اطمینانی خوشایند از برخی خواص درازمدت معلوم و مطلوب را به‌همراه می‌آورد. ملاحظه خواهد شد که این خواص به‌ویژه برای مثال ۳ی بخش ۲.۴ مناسب‌اند.

۵.۳ کاربستِ شواهدی ـ استقراییِ p ـ مقدارها

اکنون به کاربست آزمونهای معنی‌داری روی می‌آوریم که هر چند رایج‌ترند، در عین حال، به‌عنوان یک ابزارِ تنها برای کمک به تحلیل مجموعه‌های خاص از داده‌ها، و/یا بنانهادن استنباطهای استقرایی براساس داده‌ها، بحث‌انگیزترند. در این بحث پیشاپیش فرض می‌شود که توزیع احتمال مورد استفاده برای ارزیابی p ـ مقدار تاحد ممکن برای داده‌های خاصِ تحت تحلیل مناسب است.

اصل بسامدگرای کلی برای استدلال استقرایی، **ش ب**، یا چیزی مانند آن، می‌تواند راهنمایی برای رسیدن به حکم مقتضی دربارهٔ شواهد یا استنباط راجع به هر نوع فرض صفر باشد. تقریباً همان‌گونه که شخص دربارهٔ تغییرات وزن بدن برپایهٔ مشخصات عملکردیِ ترازوهای مختلف استنباطهایی می‌کند، می‌تواند از نتایج آزمون معنی‌داری با استفاده از خواص نرخ خطای آزمونها نیز استنباط به‌عمل آورد. این خواص ظرفیت آزمون خاص را به‌دلیل آنکه توانسته است ناسازگاریها و اختلافهای موجود در جنبه‌های تحت بررسی را آشکار کند، نشان می‌دهند، و این امر به نوبهٔ خود ربط‌دادن p ـ مقدارها را به فرضهایی دربارهٔ فرایندی که به‌طور آماری مدل‌بندی شده است، مجاز می‌سازد. نتیجه می‌شود که برای ارائهٔ یک شرح بسامدگرای رسا از استنباط باید در تأمین اطلاعات به‌منظور اجرای **ش ب** اهتمام ورزید.

فرضهای صفر نشانده‌شده. در حالت فرضهای صفر نشانده‌شده، استفاده از p ـ مقدارهای کوچک به‌عنوان شواهدی از اختلاف با فرض صفر در جهت فرض مقابل، سرراست است. اما، فرض کنید که معلوم شود داده‌ها موافق با فرض صفرند (p کوچک نباشد). ممکن است شخص، با استفاده از همان منطق آزمون معنی‌داری، این امر را، به مثابهٔ شاهدی برای اینکه هر اختلاف با فرض صفر کوچک‌تر از δ است تلقی کند. در چنین حالتهایی، همخوانی با فرض صفر ممکن است شواهدی از نبود اختلاف با فرضهای صفرِ دارای اندازه‌های گوناگون فراهم سازد، چنانکه در **ش ب** (**ii**) تصریح شد.

برای استنباطِ نبودِ اختلافی به بزرگی δ با H_0 می‌توانیم $\beta(\delta)$، احتمال مشاهدهٔ برازشی بدتر به H_0 را در صورتی که $\mu = \mu_0 + \delta$، بیازماییم، اگر این احتمال نزدیک به یک باشد، براساس **ش ب** (**ii**) ، داده‌ها شواهد خوبی برای $\mu < \mu_0 + \delta$اند. بدین‌ترتیب، $\beta(\delta)$ را می‌توان به‌عنوان جدّیت یا شدت آزمون در سنجش اختلاف δ در نظر گرفت؛ به‌عبارت دیگر می‌توان گفت که $\mu < \mu_0 + \delta$ آزمون شدیدی را گذرانده است (مایو [۱۷]).

این کار از تفسیر تضمین نشدهٔ سازگاری با H_0 در آزمونهای غیرحساس جلوگیری می‌کند. یک چنین ارزیابی برای داده‌های خاص مناسب‌تر است تا مفهوم توان، که نسبت به یک مقدار بحرانیِ از پیش تعیین‌شده حساب می‌شود که فراتر از آن، آزمونْ فرض صفر را «رد کند». یعنی، توان به یک ناحیهٔ رد از پیش مشخص‌شده مربوط می‌شود، نه به داده‌های خاص تحت تحلیل.

هر چند بیش‌حساسی کمتر محتمل است که مشکلی باشد، اگر آزمونی آنچنان حساس باشد که p ـ مقداری برابر یا حتی کوچک‌تر از مقدار مشاهده شده، محتمل باشد حتی وقتی $\mu < \mu_0 + \delta$، آنگاه مقداری کوچک از p شاهدی برای انحراف از H_0 به‌اندازهٔ بیش از δ نیست.

اگر خانواده‌ای از فرضهای مقابل صریح وجود داشته باشد، می‌توان مجموعه‌ای از بازه‌های اطمینان برای پارامتر مجهول تعریف‌کنندهٔ H_0 به‌دست داد و با این کار، پایه‌ای گسترده‌تر برای نتیجه‌گیری‌ها دربارهٔ پارامتر تعریف‌کننده به‌دست می‌آید.

فرضهای صفر تقسیم‌کننده و فرضهای مبنی بر نبود ساختار. در حالت فرضهای صفر تقسیم‌کننده، ناهمخوانی با فرض صفر (با به‌کارگیری مقدار دو طرفهٔ p) بیانگر جهت انحراف است (یعنی، اینکه کدام یک از دو تیمار برتر است)؛ همخوانی با H_0 حاکی از آن است که این داده‌ها شواهدی رسا حتی برای جهت اختلاف به‌دست نمی‌دهند. اغلب انتقادهایی شنیده می‌شود که آزمودن فرض صفری که می‌دانیم نادرست است بی‌فایده است، اما حتی اگر انتظار نداشته باشیم که دو میانگین مثلاً برابر باشند، این آزمون برای تقسیم انحرافها به انواعی که از لحاظ کیفی متفاوت‌اند آگاهی بخش است. وقتی فرض صفر مانند فرض صفر نبودِ ساختار است تفسیر مشابهی انجام می‌شود. مقدار متوسط p بیانگر آن است که داده‌ها حساسیت کافی برای کشف ساختار ندارند. اگر داده‌ها محدود باشند این امر ممکن است فقط هشداری علیه تفسیر بیش از حد باشد و نه شاهدی بر اینکه فکر کنیم در واقع ساختاری وجود ندارد. یعنی آزمون ممکن است ظرفیت اندکی برای کشف هرگونه ساختاری داشته بوده باشد. اما مقدار کوچک p شواهدی مبنی بر وجود اثری واقعی را نشان می‌دهد؛ که جستجو برای تفسیر واقعیت‌بنیاد از چنین اثری ذاتاً مستعد خطا نخواهد بود.

هنگامی که ارزیابیهای غیررسمی دربارهٔ گواه‌آوری یا حساسیت آزمونها صورت می‌گیرد، استدلال مشابهی مصداق دارد. اگر داده‌ها آن‌قدر گسترده باشند که همخوانی با فرض صفر دال بر نبود اثری با اهمیت عملی باشد، و یک p ـ مقدارِ به‌قدر معقول بزرگ به‌دست آمده باشد، آنگاه می‌توان آن را به‌عنوان شواهدی مبنی بر نبود اثری با اهمیت عملی گرفت. همین‌طور، اگر داده‌ها از چنان دامنهٔ محدودی باشند که بتوان فرض کرد که داده‌های همخوان با فرض صفر با انحرافهای واجد اهمیتِ علمی نیز سازگارند، آنگاه بزرگ بودن p ـ مقدار، استنباطی مبنی بر نبود انحرافهایِ مهم از فرض صفر را تضمین نمی‌کند.

فرضهای صفر مبنی بر رسایی مدل. وقتی فرضهای صفر ادعاهایی مبنی بر رسایی مدل‌اند، تفسیر نتیجه‌های آزمون بستگی به آن خواهند داشت که آیا آمارهٔ آزمون نسبتاً خاص‌نگر داریم که نسبت به انواع خاص نارسایی مدل حساس است، یا آزمونهایی به اصطلاح کلی‌نگر داریم. همخوانی با فرض صفر در حالت اول شواهدی مبنی بر نبود آن نوع انحرافی که آزمون برای کشف آن حساس است به‌دست می‌دهد، درصورتی که، با آزمون کلی‌نگر، این همخوانی کمتر آگاهی بخش است. در هر دو نوع آزمون، p ـ مقدار کوچک شواهدی از نوعی انحراف است، اما تا آنجا که مدلهای بدیل گوناگون بتوانند تخطی مشاهده‌شده را توجیه کنند (یعنی، تا آنجا که این آزمون توانایی اندکی برای تمیز دادن بین آنها داشته باشد)، این داده‌ها به‌تنهایی، فقط می‌توانند پیشنهادهایی موقتی دربارهٔ مدلهای بدیلی که باید امتحان شوند، فراهم سازند.

فرضهای صفر پرمایه. در حالتهای پیشین، همخوانی با یک فرض صفر حداکثر می‌توانست، برطبق توانایی آزمون در آشکارسازی اختلاف، شواهدی برای کنار گذاشتن اختلافهای با مقدار یا نوع مشخص به دست دهد.

در مورد فرضهای صفر پرمایه [جدی]، چیز بیشتری می‌توان گفت. اگر فرض صفر نشان‌دهندهٔ یک پیش‌بینی از نظریه‌ای باشد که از لحاظ قابلیت کاربرد عام تحت بررسی است، سازگاری با فرض صفر را می‌توان به‌عنوان

شواهدی اضافی برای این نظریه تلقی کرد، به‌ویژه اگر آزمون و داده‌ها به‌قدر کافی حساس باشند که انحرافهای عمده از نظریه را کنار بگذارند. یک جنبهٔ موضوع در اندرز فیشر (کوکران، [۳]) خلاصه شده است که می‌گوید برای کمک به اینکه بررسیهای مشاهداتی ما تفسیر علمی دقیق‌تری داشته باشند، باید نظریه‌هایمان را تفصیلی‌تر سازیم، که منظورش این بود که باید آزمونهای گوناگونی برای پیامدهای متفاوت نظریه طراحی کنیم، تا امتحانی جامع از استلزامهای آن به‌دست آوریم. این نتیجهٔ محدود که مجموعه‌ای از داده‌ها با نظریه همخوان است یک جزء به شواهدی می‌افزاید که اعتبار آنها از روی‌هم انباشتن توانایی ردِ شقوق دیگر ناشی می‌شود.

نخستین نوع مثال تحت این رهنمود، ممکن است چنین باشد که نتیجه‌های آشکارا بی‌هنجار برای نظریه یا فرضی مانند $\mathcal{T}$ وجود داشته باشند، در حالی که $\mathcal{T}$ بازرسی دقیق نظری و/یا تجربی قابل توجهی را با موفقیت پشت سر گذاشته است. اگر نتیجه‌های آشکارا بی‌هنجار برای $\mathcal{T}$ واقعی باشند، انتظار می‌رود که H_0 رد شود، به‌طوری که اگر رد نشود، این نتیجه‌ها شواهد مثبت علیه واقعیت بی‌هنجاری‌اند. در دومین نوع این مورد، باز هم یک نظریهٔ کاملاً آزمون شدهٔ $\mathcal{T}$ داریم، و نظریه‌ای رقیب مانند $\mathcal{T}^*$ تعیین می‌شود که در حوزه‌ای تاکنون آزموده نشده از حیث یک اثر با $\mathcal{T}$ در تضاد است. اگر پیشگویی حاصل از $\mathcal{T}$ را با فرض صفر یکی بگیریم، هرگونه اختلاف در جهت $\mathcal{T}^*$ شانس خوبی دارد که کشف گردد، به‌گونه‌ای که اگر هیچ انحراف معنی‌داری یافت نشود، این امر شاهدی به نفع $\mathcal{T}$ از جنبهٔ مورد آزمون است.

اگرچه نظریهٔ نسبیت عام (GTR) در دههٔ ۱۹۶۰ با بی‌هنجاریهایی مواجه نبود، رقبای GTR شکست اصل هم‌ارزی ضعیف، WEP[1]، را برای اجسام خودگرانشی سنگین، مثل سامانهٔ زمین-ماه پیشگویی می‌کردند: این اثر که اثر نوردوت[2] نامیده می‌شود، برای GTR برابر صفر است (که با فرض صفر یکی گرفته می‌شود) و برای رقبا ناصفر خواهد بود. اندازه‌گیری زمانهای سفر رفت و برگشت بین زمین و ماه (بین سالهای ۱۹۶۹ و ۱۹۷۵) امکان داد که دربارهٔ وجود چنین بی‌هنجاری برای GTR تحقیق شود. پیدانشدن شواهدی علیه فرض صفر، کرانهای بالایی را برای تخطّی ممکن از WEP تعیین کرد، و چون آزمونها به‌قدر کافی حساس بودند، این اندازه‌گیری‌ها شواهد خوبی فراهم ساختند که اثر نوردوت وجود ندارد و بدین‌ترتیب شواهدی به نفع فرض صفر در دست است (ویل، [۳۱]). توجه کنید که یک چنین نتیجهٔ منفی شواهدی به نفع همهٔ GTR (در همهٔ حوزه‌های پیشگویی) فراهم نمی‌سازد، بلکه شواهدی به نفع درستی آن از حیث این اثر به‌دست می‌دهد. منطق کار چنین است: نظریهٔ $\mathcal{T}$ گزارهٔ H_0 را دست‌کم با تقریبی نزدیک به وضعیّت واقعی پیشگویی می‌کند؛ نظریهٔ رقیت $\mathcal{T}^*$ اختلاف مشخصی را با H_0 پیشگویی می‌کند، و احتمال زیاد دارد که اگر $\mathcal{T}^*$ صحیح باشد آزمون چنین اختلافی از $\mathcal{T}$ را کشف کند. پس کشف نشدن هیچ اختلافی، شاهدی بر نبودِ آن است.

۶.۳ بازه‌های اطمینان

چنانکه در بالا اشاره شد، در بسیاری از مسأله‌ها تدارک بازه‌های اطمینان، علی‌الاصول در دامنه‌ای از سطحهای احتمال، ثمربخش‌ترین تحلیل فراوانی‌گرا را به‌دست می‌دهد. اگر چنین باشد، تحلیل بازهٔ اطمینان نیز باید ذیلِ اصل بسامدگرای کلی ما قرار گیرد و قرار می‌گیرد. در آزمون یک‌طرفهٔ $\mu = \mu_0$ در مقابل $\mu > \mu_0$، p-مقدار کوچک متناظر با این است که μ_0 (صرفاً) از بازهٔ اطمینان متناظر $(1 - 2p)$ (دوطرفه) (یا $1 - p$ برای بازهٔ اطمینان یک‌طرفه) حذف شود. اگر μ را μ_L یعنی کران اطمینان پایینی، قرار می‌دادیم، آنگاه نتیجه‌ای ناهمخوان‌تر با احتمال زیاد $(1 - p)$ رخ می‌داد. بدین‌ترتیب، **ش‌ب** اجازهٔ قبول این وضع را به مثابهٔ شواهد ناسازگاری با $\mu = \mu_L$ (در جهت مثبت) می‌دهد. افزون بر آن، این استدلال مزیت درنظرگرفتن چند بازهٔ اطمینان را در دامنه‌ای از سطحها، به‌جای آنکه صرفاً گزارش شود که مقدار پارامتری مفروض درون بازه‌ای در یک سطح اطمینانِ ثابت قرار دارد یا خیر، نشان می‌دهد.

نیمن نظریهٔ بازه‌های اطمینان را از ابتدا یعنی تنها با اتکای تلویحی و نه تصریحی بر کار قبلی‌اش با پیرسون در نظریهٔ آزمونها ابداع کرد. این امر تا حدودی بستگی به‌نحوهٔ عرضهٔ مطلب دارد که براورد بازه‌ای را اصولاً آن قدر متفاوت از آزمون فرضها بینگاریم که [فکر کنیم] در بهترین حالت، به‌طور مجزا برای حفظ تمایز مفهومی [بین این دو] پدید آمده است یا نه. از سوی دیگر مزایای قابل‌توجهی برای درنظرگرفتن یک حد اطمینان، بازه یا ناحیه‌ای به‌عنوان مجموعهٔ مقدارهای پارامتر سازگار با داده‌ها در سطحی مشخص، وجود دارد که با آزمودن هر مقدار ممکن به‌وسیلهٔ شیوه‌های متقابلاً همخوان تعیین می‌شود. به‌خصوص، این رویکرد، در سطح معنی‌داریِ مشخصی، به‌راحتی با بازه‌های اطمینانی که تهی هستند یا مرکب از همهٔ مقادیر ممکن پارامترند مواجه می‌شود. چنین ناحیهٔ تهی یا نامتناهی دالّ بر آن است که داده‌ها با همهٔ مقادیر ممکن پارامترها ناسازگارند یا با همهٔ مقادیر ممکن سازگارند. ساختن مثالهایی که در آنها این نتایج کاملاً مناسب به‌نظر برسند، آسان است.

۴. برخی پیچیدگیها: اثرهای گزینش

فرمول‌بندی آرمانی که در تعریف آغازین آزمون معنی‌داری به‌چشم می‌خورد علی‌الاصول با یک فرض و یک آمارهٔ آزمون شروع می‌شود، و مراحل بعدی، کسبِ داده‌ها، به‌کار بستن آزمون و بررسی برامد است. پس شیوهٔ فرضی مستتر در تعریف آزمون درحد قابل قبولی با آنچه انجام شده تطابق دارد؛ براوردهای ممکن، مقدارهای ممکن متفاوت از آمارهٔ آزمونِ مشخص هستند. این امر امکان می‌دهد که ویژگیهای توزیع آمارهٔ آزمون در کسب اطلاع از ویژگیهای متناظر سازوکار مولد داده‌ها به‌کار آید. شیوه‌ای که عملاً از آن پیروی می‌شود به‌دلایل گوناگونی ممکن است متفاوت باشد و اینک یک جنبهٔ عام آن را بررسی می‌کنیم.

اغلب اتفاق می‌افتد که یا فرض صفر یا آمارهٔ آزمون از وارسی مقدماتی داده‌ها تأثیر می‌پذیرد، به‌طوری که شیوهٔ واقعی تولید نتیجهٔ آزمون نهایی تغییر می‌یابد. این امر به‌نوبهٔ خود ممکن است توانایی آزمون را در کشف اختلافات با فرض صفر به‌گونه‌ای قابل اعتماد، تغییر دهد، و تعدیلهایی را در احتمالهای خطای آن لازم بنماید.

تا آنجا که p به‌عنوان جنبه‌ای از رابطهٔ منطقی یا ریاضی بین داده‌ها و مدل احتمالی در نظر گرفته شود چنین گزینشهای مقدماتی بی‌ربط‌اند. این امر برای حصول اطمینان از اینکه p-مقدارها آن منظوری را براورده سازند که

1. Weak Equivalence Principle 2. Nordvedt effect

برای آنها در استنباط بسامدگرا، خواه در بافتهای رفتاری یا شواهدی، درنظر گرفته شده است، کفایت نمی‌کند. درحدی که شخص محاسباتی مبتنی بر خطا بخواهد که به آزمون معنایی حاکی از قابلیت کاربرد در کارهای آمار بسامدگرا ببخشد، تحلیل مقدماتی و گزینش ممکن است بسیار مناسب باشد.

نکتهٔ کلی مورد نظر هم در نوشتگان فلسفی و هم نوشتگان آماری بسیار مورد بحث قرار گرفته است؛ در مباحث فلسفی تحت عنوانهایی مانند لزوم نوبودن یا پرهیز از فرضهای مورد ویژه، و در مباحث آماری به‌عنوان قواعدی علیه دزدکی نگاه‌کردن به‌داده‌ها یا جستجوی معنی‌داری، و در نتیجه لزوم به‌حساب آوردن اثرهای گزینش، بحث شده‌اند. موضوع کلّی این است که آیا وقتی فرض صفر $H_\circ$ به‌طریقی برای آزمون‌شدن ساخته یا انتخاب می‌شود که به رابطهٔ مشاهده شدهٔ خاصی بین $H_\circ$ و y، خواه توافق یا عدم توافق، منجر شود، آیا تأثیر شواهدیِ داده‌های y بر استنباط یا فرض $H_\circ$ تغییر می‌یابد یا نه. کسانی که با رویکردهای منطقی به تأیید نظر مساعد دارند می‌گویند نه (مثلاً، میل [۲۰]، کینز [۱۴]، در صورتی که کسانی که به مفهوم آماری خطا عنایت بیشتری دارند می‌گویند آری (ویول [۳۰]، پیرس [۲۵]). به پیروی از این فلسفهٔ اخیر، پوپر خواستار آن بود که دانشمندان پیشاپیش مشخص سازند چه برامدهایی را به‌مثابهٔ ابطال $H_\circ$ درنظر خواهند گرفت، شرطی که حتی خود او وادار به رد آن شد؛ کل موضوع در فلسفه لاینحل باقی مانده است (مایو [۱۷]).

ملاحظات آماری خطا امکان پیشروی را با تدارک ملاکهایی برای زمانی که گزینشهای داده‌وابستهٔ گوناگون اهمیت دارند و نحوهٔ به‌حساب آوردن تأثیر آنها بر احتمالهای خطا، فراهم می‌سازد. به‌ویژه، اگر به‌دلیل آنکه آمارهٔ آزمون بزرگ است فرض صفر برای آزمون انتخاب شود، احتمال یافتن چنین ناهمخوانی یا ناهمخوانیهای دیگر حتی تحت فرض صفر ممکن است بزرگ باشد. بدین‌ترتیب، بنا به **ش ب (i)**، شواهدی واقعی مبنی‌بر ناهمخوانی با فرض صفر نخواهیم داشت، و جز آنکه pـمقدار به‌طرزی مقتضی اصلاح شود، استنباط گمراه‌کننده خواهد بود. در حدی که شخص از محاسبات مبتنی بر خطا انتظار دارد که معنایی به آزمون ببخشد، معنایی حاوی اطمینان‌دهی محدود در این باره که ناسازگاری ظاهری در این حالت خاص واقعی است و صرفاً معلول شانس نیست، تعدیل pـمقدار لازم است.

چنین تعدیلهایی اغلب در مواردی شامل گزینشهای داده‌وابسته یا در مدل گزینی یا در مدل‌سازی پیش می‌آیند؛ مسألهٔ تعدیل p اغلب در مواردی شامل آزمون فرضهای چندگانه رخ می‌نماید، اما مهم است که این حالتها صرفاً به این دلیل که داده‌وابستگی یا آزمون فرض چندگانه وجود دارد با هم اجرا نشوند. اکنون چند حالت خاص را به اجمال شرح می‌دهیم تا نکته‌های اساسی در روایتهای مختلف را آشکار سازیم. سپس بررسی می‌کنیم که در نظرگرفتن گزینش در هر حالت لازم است یا نه.

۱.۴ مثالها

مثال ۱. پژوهشگری مثلاً ۲۰ مجموعهٔ داده‌های مستقل دارد، که هر یک حاکی از اثری متفاوت اما بسیار مرتبط با بقیه است. پژوهشگر همهٔ ۲۰ آزمون را انجام می‌دهد و تنها کوچک‌ترین p را، که در واقع حدود ۰٫۰۵ است، و فرض صفر متناظر آن را گزارش می‌کند. نکته‌های اساسی عبارت‌انداز: استقلال آزمونها و گزارش نکردن نتیجه‌های حاصل از آزمونهای غیرمعنی‌دار.

مثال ۲. روایتی بسیار آرمانی از آزمون مربوط به تطبیق DNA با نمونه‌ای‌مفروض، که شاید از آنِ یک بزهکار باشد، این است که جستجو در سراسر یک پایگاه داده‌های تطبیق‌های ممکن، هر بار با یک فرد، صورت گرفته امتحان می‌شود که فرض توافق با نمونهٔ مفروض رد می‌شود یا نه. فرض کنید که حساسیت و ویژگی هر دو خیلی زیاد هستند. یعنی، احتمالهای منفی‌های کاذب و مثبت‌های کاذب هر دو خیلی کوچک‌اند. نخستین فرد، از پایگاه داده‌ها که فرض در مورد آن رد شود، در صورتی که چنین فردی وجود داشته باشد، تطبیق واقعی اعلام می‌شود و این فرایند همانجا متوقف می‌گردد.

مثال ۳. در یک بررسی ریزآرایه[1] چندهزار ژن از لحاظ تجلی بالقوهٔ، مثلاً، تفاوتی بین وضعیت بیماری نوع ۱ و نوع ۲ آزمون می‌شوند. بنابراین چند هزار فرض وجود دارند که هر یک با فرض صفر مربوط در یک گام تحت بررسی قرار می‌گیرند.

مثال ۴. برای بررسی وابستگی یک متغیر پاسخ یا برامدی مانند y به متغیری تبیینی مانند x، قصد آن است که از تحلیل رگرسیون خطی y برحسب x استفاده شود. وارسی داده‌ها حاکی از آن است که بهتر است از رگرسیون $\log y$ برحسب $\log x$ استفاده شود، مثلاً به این دلیل که رابطه به‌حالت خطی نزدیک‌تر است یا چون مفروضات ثانوی، مانند ثابت‌بودن واریانس خطا، با تقریب بهتری، صادق‌اند.

مثال ۵. برای بررسی وابستگی یک متغیر پاسخ یا برامدی مانند y به تعداد قابل توجهی از متغیرهای تبیینی بالقوهٔ x، شیوهٔ داده‌وابستهٔ گزینش متغیر به‌کار گرفته می‌شود تا نمایشی به‌دست آید که سپس با روشهای استاندارد برازانده شده و فرضهای ذی‌ربط آزمون شوند.

مثال ۶. فرض کنید وارسی مقدماتی داده‌ها نوعی اثر کلاً نامنتظر یا نوعی نظم را نشان دهد که در مرحله‌های آغازین به آن فکر نشده بود. در نتیجهٔ آزمونی رسمی، این اثر «به‌شدت معنی‌دار» است. معقول است چه نتیجه‌ای گرفته شود؟

۲.۴ نیاز به تعدیلهایی به‌خاطر گزینش

برای بحث عمیق دربارهٔ همهٔ این مثالها مجال کافی در این مقاله نیست. یک موضوع اساسی ناظر به این است که کدام یک از این وضعیتها نیاز به تعدیل به‌خاطر آزمون کردن چندگانه یا گزینش داده‌وابسته دارد و آن تعدیل چه باید باشد. تصور کلی از سرشت یک نظریهٔ تحلیل و تفسیر بسامدگرا چگونه کمک می‌کند که به جوابها برسیم؟

پیشنهاد ما این است که این امر به روال زیر صورت گیرد: نخست باید درنظر گرفته شود که آیا بافتار مورد نظر بافتاری است که در آن موضوع اصلی مورد نظر کنترل نرخهای خطا در یک سلسله ازکاربردهاست (هدف رفتاری)، یا انجام استنباط استقرایی خاص یا ارزیابی شواهد خاص (هدف استنباطی). احتمالهای خطای مربوط را برای حالت اول می‌توان تغییر داد و برای حالت اخیر نمی‌توان. دوم، لازم است دنبالهٔ تکرارهای ذی‌ربطی را که باید بسامدها را برپایهٔ

1. microarray

آنها بنا نهاد شناسایی کرد. شرط کلی آن است که ناهمخوانی با یک فرض صفر را به کمک شیوه‌ای ارائه نکنیم که ناهمخوانیها را با فراوانی نسبتاً بیشتری مطرح می‌کند حتی اگر فرض صفر درست باشد. تعیین سریهای فرضی ذی‌ربطی که براساس آنها بسامد خطا باید حساب شود، ایجاب می‌کند که طبیعت مسألهٔ استنباط در نظر گرفته شود. به‌طور مشخص‌تر، موانع ویژه‌ای را که به‌منظور استنباطی قابل‌اعتماد در حالت خاص باید از آنها پرهیز کرد، و ظرفیت آزمون را، به‌عنوان یک ابزار اندازه‌گیری برای آشکارساختن وجود موانع، باید شناسایی کرد.

هنگامی که هدف اصلی ما ارزیابی شواهد خاص است، **ش‌ب** برخی رهنمودها را در اختیار می‌گذارد. به‌طور مشخص‌تر، مشکل وقتی رخ می‌نماید که برای گزینش یک فرض برای آزمون کردن یا تغییر مشخصات یک مدل زیربنایی، داده‌ها به‌طریقی به‌کار گرفته می‌شوند که یا از **ش ب** عدول می‌شود یا نمی‌توان تعیین کرد که آیا **ش ب** صادق است یا خیر (مایو و کروس [۱۸]).

مثال ۱ (جستجوی معنی‌داریِ آماری). شیوهٔ آزمون بسیار متفاوت با حالتی است که در آن تک‌فرض صفری که معلوم‌شده از حیث آماری معنی‌دار است از قبل به‌عنوان فرض برای آزمون تعیین‌شده باشد؛ شاید آن فرض $H_{۰,۱۳}$، سیزدهمین فرض صفر از بین ۲۰ فرض، باشد. در مثال ۱، نتیجه‌های ممکن عبارت‌اند از عاملهایی که ممکن است از لحاظ آماری معنی‌دار باشند و معلوم شود که انحراف معنی‌دار آماری «محاسبه شده»ای از فرض صفر را نشان می‌دهند. بنابراین، احتمال خطای نوع ۱ احتمال یافتن دست‌کم یک چنین تفاوت معنی‌دار در بین ۲۰ تاست، هر چند که فرض صفر فراگیر درست است (یعنی، همهٔ بیست تفاوت مشاهده‌شده به واسطهٔ شانس رخ داده‌اند). احتمال آنکه این شیوه باعث ردکردن خطاآمیز شود متفاوت با، و خیلی بزرگ‌تر از، ۰٫۰۵ (و تقریباً ۰٫۶۴) است. راههای متفاوت، و در واقع راههای بسیار بیشتری، برای اتکاب خطا در این مثال نسبت به زمانی که یک فرض صفر از قبل مشخص شده، وجود دارد و این امر در pـ مقدار تعدیل‌شده بازتاب می‌یابد.

این قدر از مطلب کاملاً معلوم است، اما آیا این امر باید بر تفسیر نتیجه در یک چارچوب استنباط استقرایی تأثیر بگذارد؟ بنابر **ش‌ب** باید تأثیر بگذارد. اما دغدغه این نیست که از اعلامِ اشتباهی و مکرر اثرهای واقعی در یک سری پرهیز شود، نگرانی آن است که این آزمون به‌عنوان ابزاری برای تمیز دادن اثرهای واقعی از اثرهای شانسی در این حالت خاص به‌طور ضعیف عمل می‌کند. زیرا می‌دانیم که دست‌کم یک انحراف چشمگیر از این نوع، حتی اگر همهٔ انحرافها به واسطهٔ شانس باشند، رایج است؛ آزمون به سختی به ما اطمینان داده است که با پرهیز از ارتکاب به‌چنین اشتباهی در این حالت، کار خوبی انجام داده است. حتی اگر دلیلهایی دیگر برای باورداشتن به اصالت اثری که پیدا شده است وجود داشته باشند، انکار می‌کنیم که این آزمون به تنهایی چنین شواهدی را فراهم ساخته است.

گفته‌ایم که محاسبات بسامدگرایانه برای آن به‌کار می‌روند که حالت خاص را بیازمایند؛ و این امر با مشخص‌سازی ظرفیت آزمونها در عیان ساختن اشتباه در استنباط صورت می‌گیرد؛ بر این مبنا «شیوهٔ جستجو» ظرفیت اندکی هشداردادن در این مورد دارد که عملاً از اشتیاق خود بکاهیم حتی وقتی چنین کاهشی ضرورت دارد. از سوی دیگر، اگر شخص pـ مقدار را تعدیل کند تا نرخ خطای کلی را بازتاب دهد، بازهم آزمون ابزاری می‌شود که در خدمت این هدف است.

مثال ۱ را می‌توان با یک آزمایش عاملی استاندارد که برای تحقیق اثرهای چند متغیر تبیینی به‌طور همزمان طراحی شده است، قیاس کرد. در اینجا شماری از پرسشهای متمایز، هر یک با فرض مربوط و هر یک با pـ مقدار مخصوص به‌خود، وجود دارند. اینکه پرسشها را از طریق یک مجموعهٔ داده‌ها مورد بررسی قرار می‌هیم به‌جای آنکه از طریق مجموعه‌های داده‌های جداگانه بررسی کنیم، به یک معنا یک تصادف فنی است. هر p به‌درستی در زمینهٔ پرسش خودش تفسیر می‌شود. مشکلات برای استنباطهای خاص در صورتی پیش می‌آیند که عملاً بسیاری از پرسشها را دور بیندازیم و توجه خود را تنها بر یکی، یا به‌طور کلی‌تر، برشماری اندک، به‌دلیل آنکه دارای کوچک‌ترین pاند، متمرکز کنیم. زیرا در این صورت ظرفیت آزمون را در تغییردادن نظرمان به‌وسیلهٔ pـ مقداری که به‌درستی محاسبه شده، تغییر داده‌ایم، خواه شواهدی بر استنباط مورد نظر داشته باشیم یا نه.

مثال ۲ (تبیین اثری معلوم با استقرای حذفی). مثال ۲ شباهت سطحی با مثال ۱ دارد؛ یافتن یک تطبیق DNA تا حدی مثل یافتن یک انحرافِ از لحاظ آماری معنی‌دار از یک فرض صفر است: شخص درون داده‌ها جستجو می‌کند و بر یک حالت که در آن «تطبیق» با DNAی بزهکار یافت می‌شود تمرکز کرده، عدم تطبیق‌ها را نادیده می‌گیرد. اگر شخص برای «جستجو» در مثال ۱ تعدیل انجام می‌دهد، آیا نباید به‌طور کلی به همان طریق در مثال ۲ تعدیل انجام دهد؟ نه.

در مثال ۱ نگرانی عبارت است از استنباط کردن یک اثر «بازآفریدنی» واقعی، در حالی که چنین اثری وجود ندارد؛ در مثال ۲، اثری معلوم یا پیشامدی خاص، یعنی DNAی بزهکار، وجود دارد و شیوه‌هایی قابل اعتماد به‌کار گرفته می‌شوند تا علت یا منبع خاص (چنانکه با نرخ پایین «تطبیق اشتباه» بیان می‌شود) ردیابی گردد. احتمال زیاد دارد که اگر i بزهکار نباشد تطبیقی با فرد i به‌دست نیاوریم؛ پس بنابر **ش‌ب**، یافتن یک تطبیق در سطحی کیفی، شاهد خوبی بر بزهکاربودن i است. افزون بر آن، هر عدم تطبیق یافت شده، بنا به تصریحات مثال، عملاً آن شخص را حذف می‌کند؛ از این‌رو، هر قدر چنین نتیجه‌های منفی زیادتر یافت شوند، «تطبیق» استنباط‌شده محکم‌تر است؛ در صورتی که در مثال ۱ چنین نیست.

چون حداکثر یک فرض صفرِ مبنی‌بر بی‌گناهی نادرست است، شواهد بی‌گناهی برای یک نفر، اگرچه به‌قدر اندک، شانس مجرم‌بودن دیگری را افزایش می‌دهد. پس از آنکه شیوهٔ نمونه‌گیری برای آزمون‌کردن مشخص شود، ارزیابی نرخهای خطا حتماً میسر است. در اینجا از ارائهٔ جزئیات خودداری می‌کنیم.

موردی که شباهت کلی با این وضعیت دارد به بی‌هنجاری مدار عطارد مربوط می‌شود: چون تلاشهای پرشمار برای ارائهٔ تفسیری نیوتنی از این پدیده بی‌ثمر از آب درآمد، وقتی که نظریهٔ اینیشتین توانست نتیجه‌های بی‌هنجار را با دقّت و بدون هیچگونه تعدیلهای موردی پیشگویی کند، بیش از پیش اعتباریافت.

مثال ۳ (داده‌های ریزآرایه). در تحلیل داده‌های ریزآرایه، یک فرض آغازی معقول این است که شمار بسیار بزرگی از فرضهای صفر باید آزمون شوند که کسر نسبتاً اندکی از آنها (اکیداً) نادرست‌اند، و فرض صفر فراگیری که حاکی از نبود هیچ اثر حقیقی باشد، اغلب نامعقول است. دراین‌صورت، مسأله از نوع انتخاب جایگاههایی است که در آنها اثری می‌تواند ثابت‌شده تلقی گردد. در اینجا، نیاز به تعدیل به‌خاطر آزمون کردن چندگانه، عمدتاً به واسطهٔ این نگرانی

عملگرایانه که از «نوفهٔ بسیار زیاد در شبکه» پرهیز شود، موجه است. توجه عمده معطوف به این است که چگونه باید نرخهای خطا را به بهترین وجه تعدیل کرد تا به مؤثرترین صورت بیانگر فرضهای ژنیِ قابل پیگیری باشند. در این‌صورت، تحلیل مبتنی‌بر خطا از موضوعها از طریق نرخ کشف نادرست است، یعنی از طریق نسبت درازمدت جایگاههایی که به‌عنوان مثبت برگزیده شده‌اند و در آنها اثری وجود ندارد. یک فرمول‌بندی بدیل از طریق مدل بیزِ تجربی انجام می‌شود و نتیجه‌گیری‌های حاصل از آن را می‌توان به نرخ کشف نادرست ربط داد. روش اخیر ممکن است مرجح باشد زیرا یک نرخ خطای خاص برای هر ژن انتخاب‌شده می‌توان یافت؛ در برخی موارد محتمل است که شواهد بسیار قوی‌تر از سایر موارد باشند و این تمایز در نرخ کشف نادرست سراسری محو می‌شود. برای ملاحظهٔ بررسی نظام‌مندی از این موضوع، ر.ک. شافر [۲۸].

مثال ۴ (بازتعریف آزمون). اگر آزمونها با مشخصات متفاوت اجرا شوند، و آزمونی برگزیده شود که معنی‌داری آماریِ کرانگین را می‌دهد، آنگاه تعدیل به‌خاطر گزینش لازم است، هر چند که تعیین تعدیل دقیق ممکن است مشکل باشد. اگر شخصی تأثیر گذاشتن نتیجه بر انتخاب مشخصات آزمون را مجاز بداند، شیوه‌ای را که منجر به pـمقدارشده تغییر می‌دهد، و این کار ممکن است غیرقابل قبول باشد. در حالی که موضوع اساسی و فرض بی‌تغییر باقی می‌مانند تعیین مشخصات دقیق مدل احتمالی با تحلیل مقدماتی داده‌ها به‌گونه‌ای هدایت شده است که سازوکار تصادفی‌ای را که عملاً مسئول برامد آزمون بوده، تغییر دهد.

این وضع مانند آزمون توانایی یک تک‌تیرانداز است از این طریق که به سوی صفحه‌ای تیراندازی کند و سپس دایرهٔ هدف به دور جایی که بیشترین تعداد اصابت به آن صورت گرفته رسم شود، و به اصطلاح اصل نشانه‌زن تکزاسی به‌کار گرفته شود. مهارتی که شخص به‌ظاهر آزمون می‌کند و دربارهٔ آن استنباط انجام می‌دهد عبارت است از توانایی وی در زدن به‌هدف وقتی هدف داده شده و ثابت است، در حالی که آن همان مهارتی نیست که عملاً باعث به‌دست آمدن امتیاز زیاد شده است.

برعکس، اگر انتخاب مشخصات نه براساس ملاحظات معنی‌داری آماریِ انحراف از فرض صفر، بلکه به آن دلیل صورت گیرد که با توجه به داده‌ها تغییراتی به‌منظور دستیابی به خطی‌بودن یا ثابت بودن واریانس خطا مجاز باشند، به‌نظر می‌رسد که هیچ تعدیلی به‌خاطر گزینش لازم نیست. بلکه دقیقاً خلاف آن: انتخاب مشخصات مناسب‌تر از لحاظ تجربی، اطمینان می‌دهد که pـمقدار حساب‌شده برای تفسیری از شواهد که قابل اعتماد باشد، مناسبت دارد. (مایو و اسپانوس [۱۹]). این کار را می‌توان به‌طور رسمی‌تر با درنظر گرفتن انتخاب مشخصات به‌عنوان یک تحلیل ماکسیمم درستنمایی توجیه کرد، یعنی ماکسیمم‌سازی نسبت به پارامتری متعامد با پارامترهایی که فرض صفر مورد نظر را مشخص می‌کنند.

مثال ۵ (داده‌کاوی). این مثال شبیه مثال ۱ است، اگرچه چگونگی انجام تعدیل به‌خاطر گزینش ممکن است روشن نباشد زیرا شیوهٔ به‌کاررفته درگزینش متغیر ممکن است پر پیچ‌وخم باشد. در اینجا نیز برای اجتناب از مشکلات گزارش گزینشی، به‌جای انتخاب تنها یک مدل، همهٔ مدلهای سازگار با داده‌ها را که در حد معقولی ساده‌اند مشخص می‌کنیم (کاکس و اسنل [۷]). بخشی از مشکلات اجرای یک چنین راهبردی محاسباتی‌اند نه مفهومی. مثالهایی از این نوع در تحلیلهای آماری پیچیده اهمیت دارند از این لحاظ که سلسله‌ای از انتخابها را که به‌طور غیررسمی مشخص شده‌اند دربارهٔ فرمول‌بندی مدلی که بهترین مدل برای تحلیل و تفسیر است می‌توان انجام داد (اسپانوس [۲۹]).

مثال ۶ (اثر کلاً نامنتظر). این وضع مشکلات عمده‌ای را پیش می‌آورد. در علوم آزمایشگاهی که داده‌ها با سرعت معقولی به‌دست می‌آیند، کوشش برای تکرار نتیجه‌گیری‌ها عملاً اجباری خواهد بود. در دیگر زمینه‌ها جستجو برای داده‌های دیگر که بر موضوع تأثیر دارند لازم می‌آید. تفسیر معنی‌داری آماری شدید به‌خودی خود بسیار مشکل خواهد بود، اساساً به این دلیل که گزینش رخ داده است و با هر نوع واقعگرایی، مشخص‌کردن مجموعه‌ای که در آن گزینش رخ داده است نوعاً سخت یا ناممکن است. اما ملاحظاتی که در مثالهای ۱ تا ۵ مورد بحث واقع شدند می‌توانند راهنما باشند. مثلاً، اگر وضعیت مانند وضعیت مثال ۲ (تبیین اثری معلوم) باشد منبع را می‌توان به‌طور قابل اعتماد، به‌شیوه‌ای که شواهد را تقویت کند به‌جای آنکه از ارزش آنها بکاهد، شناسایی کرد. در حالتی شبیه به مثال ۱، یک اثر گزینش وجود دارد، اما در حد قابل‌قبولی روشن است که مجموعهٔ امکاناتی که در آن این گزینش رخ داده چیست، که تصحیح pـمقدار را مجاز می‌نماید. در دیگر مثالها، یک اثر گزینش وجود دارد، اما ممکن است روشن نباشد که تصحیح را چگونه باید انجام داد. کوتاه سخن آنکه، بسیار غیرمعقول خواهد بود که امکان یادگیری چیزی جدید از داده‌ها در جهتی کلاً نامنتظر را انکار کنیم، اما باید بین بافتارهای مختلف به‌منظور کسب راهنمایی دربارهٔ اینکه چه تحلیل دیگری ممکن است لازم باشد، تمایز قائل شد.

۵. تذکرات پایانی

استدلال کردیم که احتمالهای خطا در آزمونهای بسامدگرا را می‌توان برای ارزیابی قابلیت اعتماد یا ظرفیت آزمون در تمیزدادن این موضوع به‌کار برد که آیا فرایند واقعی پدیدآورندهٔ داده‌ها با آنچه در H_0 توصیف شده همخوان است یا نیست. با آگاهی از این ظرفیت «گواه‌آوری» می‌توان مشخص کرد که شواهدی قوی علیه H_0، برپایهٔ اصل بسامدگرای **ش‌ب** که شرح دادیم، وجود دارد یا خیر. چیزی که باعث می‌شود این نوع استدلال فرضی در موردِ دردستِ بررسی به کار آید، نرخهای پایین خطای درازمدت در استفاده از این ابزار (یا آزمون) به این روال نیست؛ بلکه این است که آن نرخهای خطا چه‌چیزی را دربارهٔ منبع یا پدیدهٔ زایندهٔ داده‌ها آشکار می‌سازد. سعی نکرده‌ایم که رابطهٔ بین تحلیلهای بسامدگرا و بیزی از موضوعات بسیار مشابه را مورد بحث قرار دهیم. یک اصلِ بنیادیِ مفهوم استنباط استقرایی که خیلی با فلسفهٔ بسامدگرا سازگار است، آن است که استنباط استقرایی نیاز به اقامهٔ استدلالها و استنباطهای قاطع از طریق کنار هم‌گذاشتن چند خرده نتیجهٔ متفاوت دارد؛ ملاحظاتی را برای هدایت این خرده نتیجه‌ها شرح دادیم. البته پیچیدگی موضوعها شرح شسته‌رفته‌ای از آنها را مشکل می‌سازد، شرحی که شخص می‌توانست ارائه کند اگر الگوریتمی واحد، کل استنباط استقرایی را در بر می‌گرفت. با این حال، حاصل کارگزارشی است که به نوع استدلالهایی که دانشمندان بنا می‌نهند تا دانش و درک قابل اعتمادی از یک رشتهٔ علمی کسب کنند نزدیک می‌شود.

مراجع

1. Birnbaum, A. (1997). The Neyman-Pearson Theory as decision theory, and as inference theory; with a criticism of the Lindley-Savage argument for Bayesian theory. *Synthese* **36**, 19-49. MR0652320
2. Carnap, R. (1962). *Logical Foundations of Probability.* University of Chicago Press. MR0184839
3. Cochran, W. G. (1965). The planning of observational studies in human populations (with discussion). *J. R. Statist. Soc.* A **128**, 234-265.
4. Cox, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357-372. MR0094890
5. Cox, D. R. (1977). The role of significance tests (with discussion). *Scand. J. Statist.* **4**, 49-70. MR0448666
6. Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London. MR0370837
7. Cox, D. R. and Snell, E. J. (1974). The choice of variables in observational studies. *J. R. Statist. Soc.* C **23**, 51-59. MR0413333
8. De Finetti, B. (1974). *Theory of Probability*, 2 vols. English translation from Italian. Wiley, New York.
9. Fisher, R. A. (1935a). *Design of Experiments.* Oliver and Boyd, Edinburgh.
10. Fisher, R. A. (1935b). The logic of inductive inference. *J. R. Statist. Soc.* **98** 39-54.
11. Gibbons, J. D. and Pratt, J. W. (1975). *P*-values: Interpretation and methodology. *American Statistician* **29**, 20-25.
12. Jeffreys, H. (1961). *Theory of Probability*, Third edition. Oxford University Press. MR0187257
13. Kempthorne, O. (1976). Statistics and the philosophers. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science.* Harper and Hooker (eds.) Vol. 2, 273-314. MR0488407
14. Keynes, J. M. [1921] (1952). *A Treatise on Probability.* Reprint. St. Martin's Press, New York. MR1113699
15. Lehmann, E. L. (1993). The Fisher and Neyman-Pearson theories of testing hypotheses: One theory or two? *J. Amer. Statist. Assoc.* **88**, 1242-1249. MR1245356
16. Lehmann, E. L. (1995). Neyman's statistical philosophy. *Probability and Mathematical Statistics* **15**, 29-36. MR1369789
17. Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge.* University of Chicago Press.
18. Mayo, D. G. and M. Kruse (2001). Principles of inference and their consequences. In *Foundations of Bayesianism*, D. Cornfield and J. Williamson (eds.). Kluwer Academic Publishers, Netherlands, 381-403. MR1889643
19. MayoAYO, D. G. and Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal of Philosophy of Science* **57**, 323-357. MR2249183
20. Mill, J. S. (1988). *A System of Logic*, Eighth edition. Harper and Brother, New York.
21. Morrison, D. and Henkel, R. (eds.) (1970). *The Significance Test Controversy.* Aldine, Chicago.
22. Neyman, J. (1955). The problem of inductive inference. *Comm. Pure and Applied Maths* **8**, 13-46. MR0068145
23. Neyman, J. (1957). Inductive behavior as a basic concept of philosophy of science. *Int. Statist. Rev.* **25**, 7-22.
24. Pearson, E. S. (1955). Statistical concepts in their relation to reality. *J. R. Statist. Soc.* B **17**, 204-207. MR0076234
25. Pierce, C. S. [1931-5]. *Collected Papers*, Vols. 1-6, Hartshorne and Wiess, P. (eds.). Harvard University Press, Cambridge. MR0110632
26. Popper, K. (1959). *The Logic of Scientific Discovery.* Basic Books, New York. MR0107593
27. Savage, L. J. (1964). The foundations of statistics reconsidered. In *Studies in Subjective Probability*, Kyburg, H. E. and H. E. Smokler (eds.). Wiley, New York, 173-188. MR0179814
28. Shaffer, J. P. (2005). This volume.
29. Spanos, A. (2000). Revisiting data mining: 'hunting' with or without a license. *Journal of Economic Methodology* **7**, 231-264.
30. Whewell, W. [1874] (1967). *The Philosophy of the Inductive Sciences. Founded Upon Their History*, Second edition, Vols. 1 and 2 Reprint. Johnson Reprint, London.
31. Will, C. (1993). *Theory and Experiment in Gravitational Physics.* Cambridge University Press. MR0778909
32. Yates. F. (1951). The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. *J. Amer. Statist. Assoc.* **46**, 19-34.

******

• Deborah G. Mayo and D. R. Cox, "Frequentist statistics as a theory of inductive inference", *2nd Lehmann Symposium-Optimality*, IMS Lecture Notes-Monograph Series, **49** (2006) 77-97.

* دیبورا جی. مایو، بخش فلسفه و اقتصاد، و یرجینیا تک، آمریکا

mayod@vt.edu

** دی. آر. کاکس، کالج نافیلد، آکسفرد، انگلستان

david.cox@nuffield.ox.ac.uk