بسم الله الرحمن الرحیم

آمار کاربردی ۲

(در مدیریت، اقتصاد و حسابداری)

مؤلف:

احمد هژبر

تابستان ۱۳۹۰

سرشناسه	: هژبر، احمد
عنوان و پدیدآور	: آمارو کاربردی (۲) (در مدیریت، اقتصاد و حسابداری)/ مؤلف احمد هژبر.
مشخصات نشر	: تهران: پوران پژوهش، ۱۳۹۰.
مشخصات ظاهری	: ۳۱۷ ص.: جدول.
شابک	: 978-964-184-242-2
وضعیت فهرست‌نویسی	: فیپا
موضوع	: دانشگاهها و مدارس عالی - ایران -- آزمونها.
موضوع	: آمار -- راهنمای آموزشی (عالی)
موضوع	: آمار -- آزمونها و تمرینها (عالی)
موضوع	: آزمون دوره های تحصیلات تکمیلی -- ایران.
رده‌بندی کنگره	: ۱۳۸۹ ۸آ۲۷۶هـ/LB۲۳۵۳
رده‌بندی دیویی	: ۳۷۸/۱۶۶۴
شماره کتابخانه ملی	: ۲۰۲۹۲۵۳

انتشارات پوران پژوهش

نام کتاب:	آمار کاربردی (۲)
تألیف:	احمد هژبر
ناشر:	پوران پژوهش
حروفچینی:	پوران پژوهش
چاپ:	آرش
صحافی:	سیدالشهدا
شمارگان:	۳۰۰۰ نسخه
نوبت چاپ:	اول – تابستان ۱۳۹۰
قیمت :	۵۴۰۰۰ ریال
شابک:	۹۷۸-۹۶۴-۱۸۴-۲۴۲-۲
	ISBN: 978-964-184-242-2

دفتر مرکزی: میدان انقلاب – ابتدای کارگر جنوبی – کوچه مهدیزاده – پلاک ۹ – واحد ۴ تلفن: ۶۶۹۲۷۰۴۰

به نام تنها پرستیدنی که قلم را آفرید

پیش‌گفتار ناشر

نگاهی به شمار داوطلبان آزمون کارشناسی ارشد نشان می‌دهد که در این سال‌ها درخواست برای تحصیل در دوره‌های تحصیلات تکمیلی دانشگاه‌ها رشد چشمگیری داشته است.

دشواری پیش روی بیش‌تر داوطلبان، گوناگونی منابع درسی و دسترسی نداشتن به آنها هم‌چنین نمونه آزمون‌های مناسب برای تمرین و فهم بیش‌تر مفاهیم درسی است.

موسسه‌ی انتشاراتی پویان پژوهش با بیش از ۱۵ سال تلاش در راستای برآورده کردن نیاز آموزشی داوطلبان، آماده‌سازی و چاپ سه مجموعه‌ی گوناگون با سه هدف معین را در دستور کار داشته است.

مجموعه‌ی نخست با نام **کتاب ارشد (با جلد آبی رنگ)** که تاکنون به دست داوطلبان رسیده با استقبال چشمگیری همراه بوده است. در هر کتاب ارشد پس از شرح کامل درس در هر فصل، پرسش‌های چهار گزینه‌ای آزمون‌های سراسری و آزاد چند سال گذشته با پاسخ تشریحی آورده شده است. شرح درس در هر کتاب از این مجموعه به گونه‌ای است که برای دانشجویان ترم‌های پایین‌تر مفید بوده و نیز یک منبع درسی مناسب برای دانشجویان و استادان دانشگاه‌ها می‌باشد. کتاب ارشد نخستین بار در مهر ماه سال ۱۳۸۰ در قالب پانزده عنوان به داوطلبان شناسانده شد و اینک ۱۶۰ عنوان را در بر می‌گیرد.

مجموعه‌ی دوم با نام **چند آزمون ارشد (با جلد سیاه رنگ)** به گونه‌ای گردآوری شده است که دانشجو دفترچه‌ی آزمون‌های سراسری چند سال گذشته را با پاسخ‌های تشریحی در یک کتاب خواهد داشت.

مجموعه‌ی سوم با نام **بانک پرسش‌های چهار گزینه‌ای ارشد (با جلد نارنجی رنگ)** در دروس پایه و تخصصی هر رشته، یک **کتاب کار** به شمار می‌رود که در آن پرسش‌های طبقه‌بندی شده به همراه پاسخ‌های تشریحی آورده شده تا دانشجو با حل و بررسی پرسش‌های آن، مهارت لازم برای پاسخ‌گویی در آزمون‌ها را به دست آورد.

بسمه تعالی

مقدمه مؤلف

امروزه بر هیچکس پوشیده نیست که تعبیر و تفسیر مناسب ارقام و اطلاعات به منظور تصمیم‌گیریهای صحیح و منطقی مستلزم آشنایی با علم آمار است. در جامعه‌ی رو به تکامل کنونی، کارایی و اثربخشی هر برنامه‌ای به کمک روشهای آماری مشخص می‌شود و استفاده بهینه از آن موجب کاهش پیشداوری‌ها و تصمیم‌گیریهای نابجا از سوی برنامه‌ریزان و سیاستگذاران خواهد بود. همین ضرورت و استفاده‌ی روزافزون از روشهای آماری سبب شد تا دانشگاهها درس «آمار و کاربرد آن» را به‌عنوان درس پایه رشته‌های مدیریت، حسابداری و اقتصاد و سایر رشته‌ها منظور نمایند.

کتاب حاضر که براساس سرفصل مصوب ستاد انقلاب فرهنگی تدوین شده، حاصل بیش از ۱۵ سال تجربه‌ی مولف در تدریس دروس آمار برای رشته‌های اقتصاد، مدیریت و حسابداری است. از آنجا که تحصیلات مولف در حد دکتری تخصصی آمار و احتمال است، لذا وجه تمایز مهم این کتاب با کتاب‌های مشابه آن **پرداختن به مباحث آمار از طریق اصولی، همراه با بیان استدلال و ریشه مباحث است.**

در متن کتاب علاوه بر توضیح کامل مطالب، مثالهای مرتبط در هر قسمت ارائه و نسبت به حل آن اقدام شده است. پیشنهاد می‌گردد در برخورد با هر مثال ابتدا خودتان آن را حل نموده (یا اینکه در مورد آن فکر نمائید) و راه حل خودتان را با راه‌حل این کتاب مقایسه و نقاط ضعف خود را شناسایی و نسبت به رفع آن با حل تمرینات آخر هر بخش مرتفع نمائید. سعی شده است که مثال‌های ارائه شده دارای زمینه کاربردی باشد. بعضی مباحث تئوریک از جمله اثبات قضایا که در کادر خاکستری قرار دارد، صرفاً برای دانشجویانی آورده شده که خواستار درک عمیق‌تری از مباحث آمار بوده و از پشتوانه ریاضی بالاتری برخوردارند. در ضمن برای آن دسته از دانشجویانی که خود را برای امتحانات کارشناسی ارشد آماده می‌نمایند سؤالات چهارگزینه‌ای آزمون‌های کارشناسی ارشد سالهای گذشته به همراه پاسخ آن در انتهای هر بخش آورده شده است. سعی شده که نمادهای مورد استفاده در کتاب حالت عمومی داشته باشد بطوریکه با نمادهای مورد استفاده در امتحانات ورودی کارشناسی ارشد و امتحانات ورودی استخدامی همگنی داشته باشد. در ضمن در متن کتاب در هر بخش نکات لازم در هر مورد به‌صورت جداگانه ذکر شده که به دانشجویان عزیز در پاسخ به سؤالات تستی کمک خواهد نمود.

با اغتنام از فرصت ایجاد شده وظیفه خود می‌دانم از آقایان، محسن مختارنژاد و محمد اسلامی دانشجویان کارشناسی ارشد صنایع که در تألیف کتاب اینجانب را یاری کردند تشکر و قدردانی کنم. همچنین از آقای مسعود نیّری و خانم‌ها مینا نظری و طاهره خسروی که زحمت تایپ و صفحه‌آرایی کتاب را به عهده داشتند قدردانی ویژه دارم. از خوانندگان محترم تقاضا دارم نظرات و انتقادات خود را از طریق انتشارات پوران پژوهش به اینجانب انتقال دهند.

احمد هژبر

تابستان ۱۳۹۰

فهرست مطالب

پیشگفتار

ورود به استنباط آماری

هدف اصلی اغلب علوم، شناخت جهان و پدیده‌های طبیعی اطراف ماست. در بسیاری از موارد برای شناخت یک پدیده، الگو یا مدلی برای آن طراحی کرده و از آن پس برای پاسخ به سئوالات یا مسائلی که در ارتباط با آن پدیده مطرح می‌شود به مدل آن پدیده مراجعه می‌کنیم. بدیهی است که هر چه مدل ساخته شده با پدیده مورد نظر تطابق بیشتری داشته باشد پاسخ به سئوالات مربوطه درست‌تر صورت گرفته و پیش‌بینی‌های دقیق‌تری از رفتار آینده آن پدیده خواهیم داشت.

علم آمار به عنوان یکی از کاربردی‌ترین علوم در قرن اخیر از علوم دیگر مستثنا نیست. این علم برای مطالعه رفتار آماری پدیده‌ها یا جوامع، مدلی برای آنها طراحی می‌کند و به کمک آن مدل به سئوالات و پیش‌بینی‌های آماری آن پدیده‌ها یا جامعه پاسخ می‌گوید.

ابزار اصلی مدل‌سازی در علوم مختلف، ریاضیات است. علم آمار برای مدل‌سازی پدیده‌ها و جوامع آماری از علم احتمال، به عنوان ابزار ریاضی، نهایت استفاده را می‌برد. در فصل ۳، ۴، ۵ و ۶ جلد اول این کتاب به طور مختصر به مطالعه مبانی احتمالات و توزیع‌های احتمال پرداختیم. توزیع‌های احتمال گسسته و پیوسته که در فصل‌های ۴ و ۵ مطالعه شد در واقع مدل‌های ریاضی برای جوامع و پدیده‌های آماری هستند که در آنها متغیر آماری مورد مطالعه، گسسته و یا پیوسته است. توزیع‌های احتمال توام نیز که در فصل ۶ بیان شد مدل ریاضی جوامع یا پدیده‌های آماری هستند که در آنها دو متغیر با هم مورد مطالعه قرار می‌گیرند. در ادامه، به چند مسئله آماری و مدل‌های ریاضی به کار رفته در آنها و پرسش‌های مطرح شدهٔ آن اشاره می‌کنیم.

مسئله ۱- یک شرکت حمل نقل درون شهری می‌خواهد متوسط مجموع زمان‌های صرف شده توسط اتوبوس‌های خود را در ایستگاه‌ها محاسبه کند. ظرفیت اتوبوس‌ها ۲۰ نفر و مسیری که اتوبوس‌ها در طول آن مسافرها را جابجا می‌کنند دارای ۱۰ ایستگاه است. فرض بر این است که اتوبوس فقط در صورتی در ایستگاه توقف می‌کند که حداقل یک متقاضی قبل از رسیدن به ایستگاه تصمیم خود را با فشار دادن دکمه مخصوص اعلام کند. همچنین به تجربه معلوم شده که زمان صرف شده برای توقف و پیاده کردن مسافران در هر ایستگاه تقریباً ۳ دقیقه است.

پاسخ: مدل ریاضی را در نظر می‌گیریم که در آن یک اتوبوس با ظرفیت $n=۲۰$ مسافر شروع به حرکت کرده و بسته به وجود متقاضی در $k=۱۰$ ایستگاه، توقف و در هر ایستگاه $t=۳$ دقیقه

وقت صرف می‌کند. در این مدل فرض می‌کنیم که پیاده شدن مسافران از هم مستقل‌اند و در شروع حرکت نیز اتوبوس با ظرفیت کامل حرکت می‌کند. دو فرض اخیر ممکن است در جهان واقعی کاملاً برقرار نباشد ولی برای حل مسئله با رویکرد ریاضی، به آنها نیاز داریم.

اگر متوسط تعداد ایستگاه‌هایی که اتوبوس در طول مسیر خود توقف می‌کند را به دست آوریم، با ضرب آن در t=۳ دقیقه، زمان کل صرف شده در طول مسیر در ایستگاه‌ها به دست می‌آید. برای به دست آوردن متوسط تعداد توقف‌ها متغیرهای برنولی X_1 و X_2 و ... و X_k را برای ایستگاه‌های اول تا kام چنین تعریف می‌کنیم:

$$X_i = \begin{cases} ۱ & \text{اگر در ایستگاه iام توقف انجام شود} \\ ۰ & \text{اگر در ایستگاه iام توقف انجام نشود} \end{cases} \quad i = ۱,۲,۳,\ldots,k$$

احتمالات صفر و یک برای X_iها به صورت زیر محاسبه می‌شود:

$P(X_i = ۰) = P$ (توقفی در ایستگاه iام انجام نشود)

P(همه‌ی n مسافر در k−۱ ایستگاه دیگر پیاده شوند) $= \frac{k-۱}{k} \times \frac{k-۱}{k} \times \ldots \times \frac{k-۱}{k} = (\frac{k-۱}{k})^n$

$$P(X_i = ۱) = ۱ - P(X_i = ۰) = ۱ - (\frac{k-۱}{k})^n$$

توجه کنید که استقلال مسافران در محاسبه $P(X_i = ۰)$ در نظر گرفته شده است.

با توجه به نحوه تعریف X_iها، حاصل $\sum_{i=۱}^{k} X_i$ نمایانگر تعداد توقف در ایستگاه‌هاست، بنابراین متوسط تعداد توقف‌ها برابر است با $E\sum_{i=۱}^{k} X_i$. اما با استفاده از خاصیت خطی امید، و اینکه امید ریاضی هر یک از X_iها برابر $۱-(\frac{k-۱}{k})^n$ است (زیرا X_iها برنولی‌اند و امید آنها برابر $P(X = ۱)$ است)، خواهیم داشت.

$$\text{متوسط تعداد توقف‌ها} = E(\sum_{i=۱}^{k} X_i) = E(X_۱ + X_۲ + \ldots + X_k) = EX_۱ + \ldots + EX_k$$

$$= ۱-(\frac{k-۱}{k})^n + ۱-(\frac{k-۱}{k})^n + \ldots + ۱-(\frac{k-۱}{k})^n = k\left[۱-(\frac{k-۱}{k})^n\right]$$

بنابراین،

$$\text{متوسط کل زمان صرف شده در ایستگاه‌ها} = ۳ \times k\left[۱-(\frac{k-۱}{k})^n\right] = ۳\times۱۰\left[۱-(\frac{۹}{۱۰})^{۲۰}\right] = ۲۶/۳۵$$

مقدار به دست آمده، مدت زمانی است که هر اتوبوس مجموعاً در ایستگاه‌ها صرف توقف می‌کند.

■

در مسئله بالا مشاهده کردیم که چگونه از مفاهیم استقلال، احتمال، متغیرهای تصادفی برنولی و توزیع احتمال گسسته برنولی برای مدل‌سازی و حل مسئله استفاده شد. در این مسئله، همه پارامترهای مدل، از جمله، n، k و t معلوم بودند و خللی در مسئله پدید نیامد. در مسائل بعدی با مواردی روبرو می‌شویم که بعضی پارامترهای مدل مجهول‌اند و برای حل مسئله نیاز داریم که به جای مقدار واقعی آن پارامترها تخمین یا برآورد پارامترها را جایگزین کنیم.

مسئله ۲- مسئول مرکز تلفن یک موسسه تجاری می‌خواهد تعداد منشی‌های پاسخگو را تا اندازه‌ای افزایش دهد که درصد تماس‌هایی که با خطوط اشغال مواجه می‌شوند حداکثر ۰/۰۱ باشد. چه تعداد خط تلفن باید مستقر کند؟

پاسخ: برای حل این مسئله مدلی ریاضی با فرض‌های قراردادی در نظر می‌گیریم. فرض اول این که زمان پاسخ‌گویی منشی‌ها به تلفن تقریباً یک دقیقه است. فرض دوم این که هر تماس را یک رخداد در یک نقطه از زمان تلقی می‌کنیم به نحوی که فرآیند رخداد تماس‌های تلفنی را فرآیند پواسن می‌گیریم. فرض سوم این که همه منشی‌ها در ساعات کاری سر کارشان حضور دارند. فرض چهارم این که نرخ تعداد تماس‌ها در ساعات مختلف کاری ثابت می‌ماند.

این فرض‌ها اگرچه ممکن است خیلی واقع‌بینانه نباشند ولی برای حل مسئله با رویکرد ریاضی لازم‌اند. البته فرض آخر را می‌توان حذف کرد، یعنی نرخ تعداد تماس را در ساعات مختلف متغیر فرض کرد[۱]، ولی در این صورت حل مسئله پیچیده‌تر می‌شود.

براساس فرض‌های بالا، مدل ریاضی مناسب برای مسئله، توزیع احتمال پواسن است (که در فصل چهارم جلد اول کتاب اشاره شده است). براساس این مدل تعداد خطوط باید به اندازه‌ای باشد که احتمال اشغال بودن همه آنها در یک دقیقه خاص حداکثر ۰/۰۱ باشد یعنی احتمال آن که در یک دقیقه خاص بیش از تعداد خطوط، تماس تلفنی گرفته شود حداکثر ۰/۰۱ باشد. اگر تعداد خطوط لازم را k فرض کنیم باید از نامعادله زیر مقدار k را به دست آوریم.

$$P(X > k) = \sum_{x=k+1}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \leq ۰/۰۱$$

در رابطهٔبالا احتمال هر تعداد تماس در یک دقیقه از مدل احتمال پواسن که به صورت $P(x) = e^{-\lambda}\lambda^x/x!$ است محاسبه شده است و λ در این مدل نرخ تعداد تماس به مؤسسه در طول یک دقیقه است. مقدار λ به عنوان پارامتر این مدل بر ما مجهول است، و نیاز داریم برآورد آن را جایگزین کنیم. برای این منظور در چند زمان مختلف، هر بار یک دقیقه، تعداد تماس‌هایی را که به

۱ - در چنین مواردی، از مدل فرآیند پواسن با نرخ غیرثابت استفاده می‌شود. [۵]

شرکت می‌شود (چه آنهایی که پاسخ داده می‌شوند و چه آنهایی که با خطوط اشغال مواجه می‌شوند) شمارش کرد، و به عنوان نمونه‌ای تصادفی به صورت زیر در نظر می‌گیریم:

$x_1, x_2, x_3, ..., x_n$

در نمادهای بالا x_iها را مشاهدات نمونه تصادفی دردقیقه و n را حجم نمونه می‌نامند. فرض کنید نتایج این مشاهدات در n=۱۰ زمان مختلف چنین به دست آمده است:

۶، ۳، ۴، ۵، ۲، ۷، ۱، ۸، ۷، ۷

در فصل ۸ خواهیم دید که برآورد مناسب برای پارامتر λ در توزیع پواسن $\hat{\lambda} = \bar{x}$ است که عبارت است از:

$$\hat{\lambda} = \bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{۶+۳+۴+۵+۲+۷+۱+۸+۷+۷}{۱۰} = ۵$$

بنابراین در نامعادله بالا به جای λ برآورد آن یعنی ۵ را قرار می‌دهیم:

$$\sum_{x=k+1}^{\infty} \frac{e^{-۵}۵^x}{x!} \leq ۰/۰۱ \Rightarrow ۱ - \sum_{x=0}^{k} \frac{e^{-۵}۵^x}{x!} \leq ۰/۰۱ \Rightarrow \sum_{x=0}^{k} \frac{e^{-۵}۵^x}{x!} \geq ۰/۹۹$$

با استفاده از جدول احتمالات تجمعی پواسن در ضمیمه کتاب، مقدار k برابر ۱۱ به دست می‌آید، بنابراین تعداد منشی‌ها باید به ۱۱ نفر افزایش یابد.

■

در این مسئله ملاحظه می‌کنیم که در ابتدا، مدل به طور کامل مشخص نبود زیرا مقدار پارامتر λ در آن مجهول بود و برای حل مسئله، λ را به کمک نمونه‌ای که از جامعه اختیار کردیم برآورد کردیم. در این مسئله به روشنی ارتباط بین احتمالات و توزیع‌های احتمال را با مباحث آمار استنباطی، یعنی نمونه‌گیری و برآورد، مشاهده کردیم. در این مسئله، توزیع احتمال پواسن را برای مسئله به عنوان مدل اختیار کردیم و برای مشخص شدن کامل مدل، به کمک نمونه‌گیری از جامعه‌ی مورد تحقیق، پارامتر مجهول مدل را برآورد کردیم. به طور کلی توزیع‌های احتمال به عنوان مدل‌هایی برای جوامع آماری اختیار می‌شوند و برای مشخص شدن کامل آن مدل‌ها (که معمولاً شامل پارامترهایی مجهول هستند)، با انتخاب نمونه تصادفی از جامعه، به برآورد پارامتر مجهول آن مدل می‌پردازیم.

مسئله ۳- سازنده نوعی ترانزیستور می‌خواهد محصولات خود را تا مدتی ضمانت کند که محصولاتی که به دلیل نرسیدن به طول عمر ضمانت شده مرجوع می‌شوند حداکثر ۱۰٪ باشد. مدت ضمانت را باید چند اختیار کند؟

پاسخ: طبق معمول مدلی ریاضی برای طول عمر محصولات می‌سازیم. معمولاً توزیع نمایی (که شرح آن در فصل پنجم جلد اول آمد) مدل مناسبی برای طول عمر محصولات الکترونیکی است. بنابراین اگر X را متغیر طول عمر ترانزیستورها بگیریم داریم:

$$f(x)=\frac{1}{\beta}e^{-\frac{x}{\beta}} \quad x>0$$

ملاحظه می‌شود که چون متغیر طول عمر X متغیری پیوسته است، مدل ریاضی آن را نیز یک توزیع احتمال پیوسته اختیار کرده‌ایم؛ اما این مدل به طور کامل مشخص نیست زیرا پارامتر مدل، β، در آن مجهول است، بنابراین ابتدا به کمک نمونه‌ای که از محصولات این شرکت انتخاب کرده و طول عمر آنها را یادداشت کرده‌ایم، β را برآورد می‌کنیم.

۷/۵ ، ۴/۸ ، ۵/۲ ، ۱۰/۲ ، ۱۱/۲ ، ۸/۳ ، ۴/۹ ، ۶/۷ ، ۱۲/۱ ، ۷/۲ : طول عمر محصولات انتخاب شده (برحسب ماه)

در فصل ۸ خواهیم دید که برآورد مناسب برای پارامتر β، در توزیع نمایی $\hat{\beta}=\overline{x}$ است، بنابراین:

$$\hat{\beta}=\overline{x}=\frac{x_1+x_2+\ldots+x_n}{n}=\frac{۷/۲+۱۲/۱+۶/۷+۴/۹+۸/۳+۱۱/۲+۱۰/۲+۵/۲+۴/۸+۷/۵}{۱۰}=۷/۸۲$$

با جایگزینی ۷/۸۲ به جای β، مدل طول عمر به صورت زیر به دست می‌آید:

$$f(x)=\frac{1}{۷/۸۲}e^{-\frac{x}{۷/۸۲}} \quad x>0$$

اکنون برای محاسبه زمان ضمانت t، باید t را چنان به دست آوریم که احتمال طول عمر کمتر از t برابر ۰/۱۰ شود.

$$P(X<t)=۰/۱۰ \Rightarrow \int_{0}^{t}\frac{1}{۷/۸۲}e^{-\frac{x}{۷/۸۲}}dx=۰/۱ \Rightarrow ۱-e^{-\frac{t}{۷/۸۲}}=۰/۱ \Rightarrow e^{-\frac{t}{۷/۸۲}}=۰/۹$$

$$\Rightarrow t=-۷/۸۲\ln ۰/۹=۰/۸۲ \text{ (ماه)} \simeq ۲۵ \text{ (روز)}$$

■

مسئله ۴- یک پزشک محقق با تصور اینکه طول عمر افراد چاق به طور متوسط کمتر از افراد لاغر است نمونه‌ای شامل n=۲۰ نفر را که به مرگ طبیعی مرده‌اند انتخاب کرده و طول عمر و وزن آنها را به صورت زیر ثبت کرده است:

x (وزن)	۹۴	۸۲	۹۶	۷۵	۹۲	۱۰۵	۷۳	۹۵	۷۸	۷۵	۸۹	۱۱۰	۱۰۵	۹۸	۷۷	۹۴	۸۹	۷۸	۸۴	۱۰۳
y (طول عمر)	۶۷	۸۲	۵۷	۸۶	۷۲	۵۹	۷۹	۶۳	۸۳	۹۱	۷۴	۷۲	۶۴	۵۵	۷۳	۵۳	۷۹	۸۵	۷۳	۶۵

او با جمع‌آوری اطلاعات فوق در پی پاسخ به این دو سؤال است. ۱- آیا اساساً بین وزن افراد و طول عمر آنها همبستگی وجود دارد؟ ۲- اگر بین x و y همبستگی وجود دارد رابطهٔ آنها چگونه است؟

پاسخ: برای پاسخ به دو پرسش بالا مدلی خطی بین دو متغیر X (به عنوان متغیر مستقل) و Y (به عنوان متغیر تابع) به صورت $Y = \alpha + \beta X + E$ در نظر می‌گیریم. در این مدل خطی، که به مدل رگرسیونی یک متغیره موسوم است، α و β ضرائب رگرسیونی و E خطای مدل می‌باشد. سئوال اول این است که آیا اساساً همبستگی بین X و Y وجود دارد که پس از آن در پی مدلی باشیم که نوع این رابطه را مشخص کند. در فصل ۶ از جلد اول کتاب شاخصی معرفی شد که به واسطهٔ آن میزان همبستگی دو متغیری که طبق مدل خطی به هم مربوطاند اندازه‌گیری می‌شد. این شاخص را ضریب همبستگی پیرسن[1] (R) نامیدیم و نحوه محاسبه آن برای توزیع دو متغیره X و Y در جامعه در آن جا معرفی شد، اما از آنجا که تمام جامعه در اختیار نیست و تنها نمونه‌ای n=۲۰ نفره در اختیار است، به برآورد R که از فرمول زیر به دست می‌آید بسنده می‌کنیم:

$$\hat{R} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}}}$$

برآورد ضریب همبستگی پیرسن در نمونه $= -0/74$

مقدار به دست آمده برای $\hat{R}$ در نمونه نشان می‌دهد که بین وزن افراد و طول عمر آنها همبستگی خطی معکوس و شدیدی برقرار است. (به خاطر داریم که محدود R از ۱- تا ۱+ بود و علامت R نشان دهنده جهت همبستگی و نزدیکی آن به ۱+ و ۱- نشان دهنده شدت همبستگی بود). بنابراین می‌توان بین X و Y رابطه‌ای خطی قائل شد؛ اما مقدار به دست آمده برای R فقط برآورد R براساس یک نمونه n=۲۰ نفری است و این نمونه کوچک لزوماً نمی‌توانند نماینده کل جامعه باشند، به بیان دیگر باید دید که آیا فاصله $\hat{R}$ از صفر، آنقدر معنی‌دار است که بتوان نتیجه گرفت در کل جامعه نیز بین X و Y همبستگی خطی وجود دارد. برای پاسخ به این سئوال ابتدا آزمون معنی‌داری ضریب همبستگی را انجام می‌دهیم، یعنی آزمون می‌کنیم که آیا می‌توان فرض عدم همبستگی $(H_0 : R = 0)$ را در مقابل وجود همبستگی $(H_1 : R \neq 0)$ رد کرد یا خیر.

1 - Karl Pearson (1857-1936)

جزئیات این بحث در فصل ۹ (آزمون فرضیه) بیان خواهد شد، ولی قبل از آن بدون ذکر جزئیات فنی صرفاً به انجام آزمون می‌پردازیم.

$$\text{آماره آزمون} = T_{۱۸} = \frac{\hat{R}\sqrt{n-۲}}{\sqrt{۱-\hat{R}^۲}} = \frac{-۰/۷۴\sqrt{۱۸}}{\sqrt{۱-(۰/۷۴)^۲}} = -۴/۶۶$$

چون مقدار آماره آزمون کمتر از مقدار بحرانی توزیع استودنت هجده درجه آزادی در سطح خطای $\alpha = ۰/۰۱$، یعنی $-۲/۸۷۸ = -۱/۰۰۵,۱۸$، کوچکتر است، فرض عدم همبستگی $H_۰$ را در سطح خطای $\alpha = ۰/۰۱$ رد می‌کنیم؛ بنابراین با اطمینان ۹۹٪ می‌توان گفت که بین X و Y در جامعه نیز رابطهٔ خطی وجود دارد.

اکنون که از وجود رابطه خطی بین X و Y اطمینان نسبی حاصل کردیم به دنبال مشخص کردن کامل مدل خطی بیان شده می‌رویم. برای مشخص کردن کامل مدل $Y = \alpha + \beta X$ باید ضرائب مدل (ضرائب رگرسیونی) را برآورد کنیم. در فصل ۱۰ (رگرسیون) خواهیم دید که مقادیر α و β براساس داده‌های نمونه به صورت زیر محاسبه می‌شود:

$$\hat{\beta} = \frac{\sum_{i=۱}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{\sum(x_i-\bar{x})^۲}{n}} = -۰/۷۲ \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = ۱۳۶/۲۳$$

بنابراین می‌توان مدل خطی $Y = ۱۳۶/۲۳ - ۰/۷۲x$ را بین وزن و طول عمر آنها در جامعه برقرار دانست.

■

در این مسئله دیده شد که چگونه دو مبحث عمده در آمار استنباطی یعنی برآورد و آزمون فرض برای شناخت مدلی در جامعه به کار گرفته شد.

در چهار مسئله فوق ارتباط میان احتمال و توزیع‌های احتمال گسسته و پیوسته و توام را با مباحث تشکیل دهنده آمار استنباطی یعنی برآورد و آزمون فرض مشاهده کردیم و دیدیم که چگونه از توزیع‌های احتمال به عنوان مدل پیشنهادی برای جوامع استفاده می‌شود و چگونه به کمک مباحث برآورد و آزمون فرض این مدل‌ها را به طور کامل مشخص می‌کنیم. از سوی دیگر دیدیم که چگونه به کمک نمونه‌گیری از یک جامعه می‌توان نتایج به دست آمده از بخشی از جامعه را به کل جامعه تعمیم داد (آن گونه که در مبحث برآورد، نتایج حاصل از محاسبه پارامترها یا شاخص‌هایی در نمونه را به عنوان برآوردی از آن پارامتر یا شاخص در جامعه به کار می‌بریم و آن گونه که در آزمون فرض، فرضیه‌ای درباره‌ی جامعه را، صرفاً براساس اطلاعات نمونه، رد کرده یا می‌پذیریم)، و این در واقع تفسیر واژه‌ی استنباط (inference) است.