مدرسان شریف

فصل اول

«آمـــار توصیفی»

درسنامه (۱): مفاهیم اولیه

معرفی علم آمار: علم آمار به سه شاخه‌ی آمار توصیفی، آمار استنباطی (پارامتری) و آمار ناپارامتری تقسیم می‌شود.

آمار توصیفی (Descriptive Statistics)

این شاخه از علم آمار با کمک ابزارهایی مانند جداول فراوانی، نمودارهای آماری و محاسبه‌ی شاخص‌های آماری به توصیف داده‌ها می‌پردازد. در آمار توصیفی همه‌ی داده‌ها به طریق سرشماری جمع‌آوری می‌شوند و هیچ‌گونه استنباطی بر روی آن‌ها انجام نمی‌شود و فقط به توصیف آن‌ها پرداخته می‌شود.

آمار استنباطی (پارامتری) (Inferential Statistics)

در آمار استنباطی به کمک نمونه‌گیری به تحلیل و استنتاج در مورد پارامترهای کلی جامعه‌ی آماری پرداخته می‌شود. توجه کنید که در آمار استنباطی توزیع یا نحوه‌ی پخش داده‌های آماری مشخص است.

آمار ناپارامتری (آزاد از توزیع)(Non – Parametric Statistics)

آمار ناپارامتری دقیقاً مانند آمار استنباطی عمل می‌کند، با این تفاوت که در آمار ناپارامتری اغلب صفات کیفی هستند و توزیع یا نحوه‌ی پخش داده‌های آماری مشخص نیست.

مثال ۱: مرکز آمار ایران هر چند سال یک بار سرشماری انجام می‌دهد. اطلاعات جمع‌آوری شده با استفاده از کدام شاخه علم آمار مورد بررسی قرار می‌گیرند؟

۱) آمار توصیفی　　۲) آمار استنباطی　　۳) آمار ناپارامتریک　　۴) هیچکدام

☑ **پاسخ:** گزینه «۱» در آمار استنباطی و آمار ناپارامتریک نمونه‌گیری انجام می‌گیرد در حالی‌که در سرشماری، مرکز آمار ایران همه‌ی واحدهای جامعه را مورد بررسی قرار می‌دهد.

جامعه‌ی آماری یا جمعیّت (Population)

مجموعه‌ی تمام افراد یا اشیایی که مطالعات آماری در مورد یک یا چند صفت آن‌ها در یک مکان و زمان معین انجام می‌گیرد، جامعه‌ی آماری یا جمعیت گفته می‌شود. تعداد اعضای جامعه، حجم جامعه نامیده می‌شود که آن را با N نشان می‌دهند.

نمونه (Sample): در بررسی‌های آماری به دلیل هزینه‌ی زیاد، کمبود وقت و در بعضی مواقع غیرممکن بودن انجام کار، زیرمجموعه‌ای از جامعه با قاعده و ضابطه‌ی خاصی انتخاب می‌شود که به آن نمونه‌ی آماری گفته می‌شود. تعداد اعضای نمونه، حجم نمونه نامیده می‌شود و آن را با n نشان می‌دهند.

صفت مشخصه (Attribute): صفتی است که بین همه‌ی عناصر جامعه‌ی آماری مشترک بوده و جداکننده‌ی جامعه‌ی آماری از سایر جوامع است. به‌طورکلی صفات، خود به دو دسته‌ی کمی و کیفی تقسیم می‌شوند.

صفات کمی: صفات کمی صفاتی هستند که می‌توانند به صورت عددی بیان شوند. تعداد دانشجویان یک دانشگاه، درآمد یک خانواده و ... دارای صفت کمی هستند.

صفات کیفی: صفات کیفی صفاتی هستند که نمی‌توانند به صورت عددی بیان شوند. گروه خون، رنگ چشم انسان‌ها و ... دارای صفت کیفی هستند.

مثال ۲: **کدام‌یک از گزینه‌های زیر تعریف صفت مشخصه است؟** (مدیریت و حسابداری ـ سراسری ۹۳)

۱) از فردی به فرد دیگر تغییر می‌کند.
۲) صفت مشترک بین کلیه‌ی افراد جامعه است.
۳) متمایزکننده‌ی عناصر جامعه از یکدیگر است.
۴) عنصر مشترک جوامع آماری مختلف است.

☑ **پاسخ:** گزینه «۲» صفت مشترک بین اعضاء جامعه است.

انواع مقیاس‌های اندازه‌گیری صفات (مقیاس‌های استیونز)

الف ـ کیفی (گروهی)

این مقیاس، خود به دو مقیاس اسمی و رتبه‌ای تقسیم می‌شود. ماهیت عددی ندارند، قابل اندازه‌گیری نیستند و واحد ندارند. وضعیت تأهل، رنگ چشم، جنسیت و مهارت، مثال‌هایی از مقیاس کیفی هستند.

۱ـ مقیاس اسمی (Nominal Scale): زمانی که صفتی دارای حالت‌های مختلف باشد، مفهوم بهتر (بدتر) و بزرگتر (کوچکتر) بودن را نداشته باشد، همچنین صفت، ماهیت عددی نداشته باشد و بتوان آن را در طبقات جداگانه‌ای قرار داد، مقیاس این صفت اسمی است. به‌طور مثال جنسیت افراد را در نظر بگیرید. دو حالت مرد و زن دارد و مرد و زن بودن بر یکدیگر ارجحیت ندارند. همچنین ماهیت آن عددی نیست.

۲ـ مقیاس رتبه‌ای (Rank Scale): زمانی که صفتی مفهوم بهتر (بدتر) و بزرگتر (کوچکتر) بودن را داشته باشد و نتوان آن را اندازه‌گیری کرد، از این مقیاس استفاده می‌شود. مثلاً طبقه‌بندی مردم یک کشور به سه طبقه‌ی پر درآمد، با درآمد متوسط و کم درآمد تقسیم شود.

ب ـ مقیاس کمّی (عددی)

وسیله یا روشی خاص برای اندازه‌گیری یک صفت است که ماهیت و حاصل آن بصورت عددی است، مانند تعداد فرزندان، طول قد افراد، درجه‌ی حرارت .... این مقیاس خود به سه دسته‌ی شمارشی، فاصله‌ای و نسبی (نسبتی) تقسیم می‌شود.

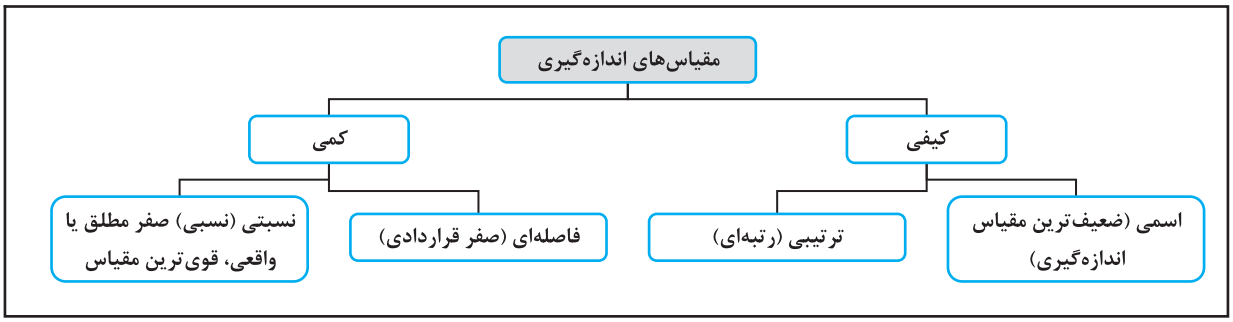
۱ـ مقیاس فاصله‌ای (Interval Scale): زمانی که صفتی دارای ماهیت عددی باشد، اختلاف بین حالتها را بیان کند و مفهوم بهتر (بدتر) و بزرگتر (کوچکتر) بودن را داشته باشد مقیاس آن فاصله‌ای است. اعداد روی دماسنج و یا نمره‌های مربوط به هوش افراد دارای مقیاس فاصله‌ای هستند.

۲ـ مقیاس نسبی (نسبتی) (Ratio Scale): این مقیاس دارای صفر مطلق است و چهار عمل اصلی روی این مقیاس انجام می‌گیرد. طول قد افراد، درآمد و هزینه‌ها همه مثال‌هایی از مقیاس نسبی هستند. در مقیاس نسبتی نسبت حفظ می‌شود، مثلاً اگر برای دو جسم و ویژگی وزن با x_1 و x_2 اندازه‌گیری شود باید $\frac{x_1}{x_2}$ ثابت بماند و به واحد اندازه‌گیری بستگی نداشته باشد.

مثال ۳: **در اندازه گیری داده‌های سطح تحصیلات، دمای اجسام و طول قد دانشجویان از چه مقیاس‌هایی استفاده شده است؟**

۱) ترتیبی ـ فاصله‌ای ـ نسبتی ۲) فاصله‌ای ـ فاصله‌ای ـ نسبتی ۳) ترتیبی ـ ترتیبی ـ اسمی ۴) اسمی ـ فاصله‌ای ـ فاصله‌ای

☑ **پاسخ:** گزینه «۱» در داده‌های سطح تحصیلات ترتیب رعایت می‌شود ـ دمای اجسام دارای مقیاس فاصله‌ای و طول قد دانشجویان دارای مقیاس نسبی است.



مثال ۴: **کدام‌یک از مقیاس‌ها دارای صفر قرار دادی است؟** (مدیریت و حسابداری ـ سراسری ۹۴ ـ طراحی و برنامه‌ریزی شهری ـ سراسری ۸۹)

۱) نسبی ۲) فاصله‌ای ۳) اسمی ۴) رتبه‌ای

☑ **پاسخ:** گزینه «۲» صفر قراردادی به صورت قرارداد صفر در نظر گرفته شده است و مفهوم آن عدد صفر واقعی نیست.

داده‌های آماری (Statistics data)

در بررسی‌های آماری باید صفت مورد بررسی به صورت اعداد و ارقام نمایش داده شود. اگر صفت مورد بررسی کمی باشد، این عمل به سادگی امکان‌پذیر است ولی اگر داده‌ها کیفی باشند، باید طبق ضوابط خاصی با عدد و رقم نمایش داده شوند. به‌طورکلی داده‌های آماری به دو دسته‌ی گسسته (طبقه‌بندی‌نشده) و پیوسته (طبقه‌بندی شده) تقسیم می‌شوند.

داده‌های گسسته (Discrete data): داده‌هایی هستند که بین هر دو مقدار متوالی a مورد نظر از آن‌ها هیچ عددی نمی‌تواند قرار گیرد. تعداد دانشجویان یک دانشکده، تعداد فرزندان یک خانواده و ... به صورت گسسته بیان می‌شوند.

داده‌های پیوسته (Continious data): داده‌هایی هستند که بین هر دو مقدار مورد نظر از آن‌ها بی‌شمار عدد می‌تواند قرار گیرد. طول قد و یا وزن افراد، طول عمر انسان‌ها و ... همه مثالهایی از داده‌های پیوسته هستند. توجه کنید که ماهیت داده‌های پیوسته به صورت اعشاری است، اگرچه در پاره‌ای از مواقع ممکن است آن‌ها را به دلیل رند کردن یا به طور تصادفی به صورت گسسته نمایش دهند، اما این نمایش مهم نیست و ماهیت اعداد برای ما مهم است. به طور مثال اگر وزن فردی ۷۴ کیلوگرم نمایش داده شود این عدد واقعی نیست چرا که وزن، ماهیتی پیوسته دارد و باید اعشاری باشد لذا ما به ماهیت عدد توجه می‌کنیم نه به نوع نمایش آن.

مثال ۵: کدام‌یک از موارد زیر داده‌های گسسته می‌باشد؟

۱) قد افراد یک کلاس
۲) وزن افراد یک شرکت
۳) زمان رسیدن اتوبوس به یک ایستگاه
۴) تعداد افرادی که به یک فروشگاه وارد می‌شوند.

☑ **پاسخ:** گزینه «۴» بنا به تعریف داده‌های گسسته داریم: داده‌هایی که بین هر دو مقدار متوالی a موردنظر از آن‌ها هیچ عددی نمی‌تواند قرار گیرد بنابراین گزینه‌ی (۴) جواب است.

مراحل یک پژوهش علمی در آمار

در تحقیقات علمی به روش آماری چند مرحله‌ی مهم باید به‌طور کامل انجام شود:

الف) مشخص‌کردن هدف: در این مرحله با کمک روش‌های تحقیق، تلاش برای افزایش اطلاعات از موضوع و بررسی اطلاعات پایدارتر برای رسیدن به هدف انجام می‌گیرد.

ب) جمع‌آوری داده‌ها: در هر پژوهش آماری، تهیه‌ی داده‌های واقعی با توجه به هدفی که از پژوهش داریم، اهمیت اساسی دارد. فرآیند جمع‌آوری داده‌ها ممکن است به روش‌های مختلفی انجام گیرد.

ج) تجزیه و تحلیل داده‌ها: داده‌هایی که به روشهای مناسب گردآوری شده، منبع اساسی برای کسب اطلاعات جدید درباره‌ی پدیده مورد مطالعه هستند در این مرحله داده‌ها را بررسی کرده و معلومات جدیدی به‌دست می‌آوریم که تعیین کننده‌ی نقاط قوّت و ضعف آن‌ها است.

د) بیان یافته‌ها: در این مرحله، تحلیل پایانی بر روی داده‌های آماری که در مرحله‌ی اول تحقیق مشخص شده‌اند، انجام می‌شود و همچنین در این مرحله پاسخ به سؤالات اولیّه امکان پذیر است.

مثال ۶: اولین مرحله در تحقیق علمی کدام است؟ (مدیریت بازرگانی ـ آزاد ۸۸)

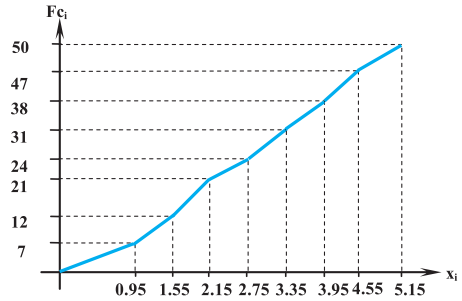
۱) فرضیه سازی
۲) تحلیل یافته‌ها
۳) هدف‌های پژوهش
۴) جمع‌آوری داده‌ها

☑ **پاسخ:** گزینه «۳» با توجه به ترتیب مراحل گفته شده در بالا هدف‌های پژوهش اولین مرحله است.

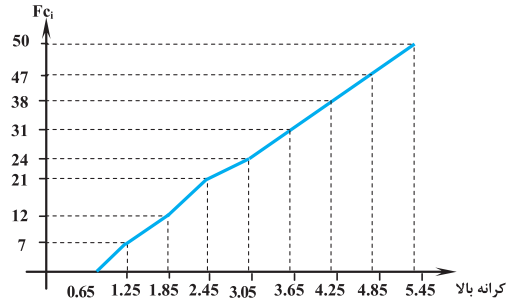
مطالعه توصیفی داده‌ها (آمار توصیفی)

مفهوم و کاربرد نمادها $\sum$

مجموعه‌ی داده‌های آماری را معمولاً با نماد $x_1, x_2, \ldots, x_N$ نشان می‌دهند. نماد $\sum$ برای جمع‌کردن داده‌های آماری بکار می‌رود. علامت $\sum$ برای عدم استفاده مکرر از علامت "+" بکار می‌رود. در نماد $\sum$ یک کران پایین، یک کران بالا و یک متغیر (چند متغیر) در جلوی سیگما نوشته می‌شود. جمله‌ای که بعد از نماد $\sum$ نوشته می‌شود، نشان دهنده‌ی مقادیری است که باید با هم جمع شوند و کران‌های پایین و بالای $\sum$ نشان‌دهنده‌ی ابتدا و انتهای جمع هستند.



«نمودار پلی‌گان فراوانی تجمعی»



«نمودار فراوانی تجمعی»

۲. نمودارهای وصفی

این نوع از نمودارها برای داده‌های کیفی بکار برده می‌شوند. مهم‌ترین این نمودارها عبارتند از:

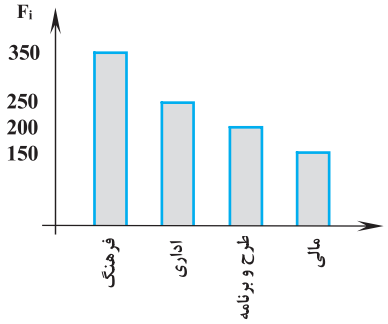
الف) نمودار ستونی (میله‌ای) (Bar-chart)

در این نوع نمودار، معمولاً محور Xها نشان‌دهنده‌ی کیفیت مشاهدات و محور عمودی آن نشان‌دهنده‌ی فراوانی مطلق یا نسبی هر گروه است.

مثال ۱۳۲: **تعداد کارکنان یک سازمان برحسب معاونت به شرح زیر است، نمودار ستونی آن را رسم کنید.**

معاونت	اداری	مالی	طرح و برنامه	فرهنگ
تعداد کارکنان	۲۵۰	۱۵۰	۲۰۰	۳۵۰

☑ **پاسخ:** می‌توانیم به جای مستطیل‌ها از خطوط استفاده کنیم ولی در اینجا مستطیل رسم می‌کنیم که نشان می‌دهد فراوانی‌ها بر روی موضوعات پخش شده است.



ب) نمودار دایره‌ای (کلوچه‌ای) (Pie chart)

برای رسم این نمودار، ابتدا دایره‌ای رسم کرده و سپس قطاع‌هایی برحسب درصد یا درجه از دایره برای هر طبقه جدا می‌کنیم. سهم این قطاع‌ها از دایره به صورت زیر محاسبه می‌گردد:

$$\text{سهم برحسب درصد} = \frac{F_i}{n} \times ۱۰۰ \quad , \quad \text{سهم برحسب درجه} = \frac{F_i}{n} \times ۳۶۰$$

مثال ۱۳۳: **کدام نمودار، برای نشان دادن وضعیت توزیع فراوانی داده‌های کیفی، مناسب است؟** (علوم اقتصادی ـ سراسری ۹۷)

۱) هیستوگرام ۲) پراکنش ۳) میله‌ای ۴) چند بر فراوانی انباشته

☑ **پاسخ:** گزینه «۳» به بررسی گزینه‌ها می‌پردازیم. نمودار هیستوگرام برای داده‌های فاصله‌ای و کمی می‌باشد، نمودار پراکنش برای داده‌های کمی و دومتغیره می‌باشد و چند بر فراوانی انباشته از طریق نمودار هیستوگرام به‌وجود می‌آید که اگر بخواهیم دو جامعه را از نظر داده های کمی با هم مقایسه نماییم نمودار مناسبی می‌باشد ولی برای داده‌های کیفی، نمودارهای دایره‌ای، میله‌ای و غیره مناسب می‌باشد.

مثال ۱۳۴: **در یک نمودار دایره‌ای، قطاعی به اندازه ۳۶° جدا شده است. اگر مجموع فراوانی‌های مطلق برابر با ۱۰۰ باشد، فراوانی مطلق طبقه‌ی مورد نظر کدام است؟**

۱) ۱۰ ۲) ۲۰ ۳) ۳۰ ۴) ۴۰

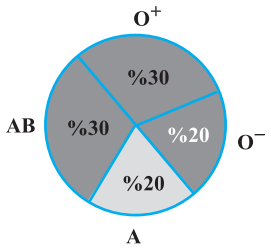
☑ **پاسخ:** گزینه «۱» با توجه به رابطه‌ی $\text{سهم بر حسب درجه} = \frac{F_i}{n} \times ۳۶۰$ داریم:

$$۳۶ = \frac{F_i}{۱۰۰} \times ۳۶۰ \Rightarrow ۳۶۰۰ = ۳۶۰ F_i \Rightarrow F_i = ۱۰$$

مثال ۱۳۵: تعدادی بیمار برحسب گروه خونی به صورت زیر مرتب شده‌اند. نمودار دایره‌ای را رسم کنید.

گروه‌های خونی	A	AB	O^+	O^-
F_i	۱۰۰۰	۱۵۰۰	۱۵۰۰	۱۰۰۰

پاسخ: با توجه به رابطه‌های گفته شده در بالا خواهیم داشت:



$$\text{سهم گروه خونی A} = \frac{F_i}{N}\times 100 = \frac{1000}{5000}\times 100 = 20\%$$

$$\text{سهم گروه خونی AB} = \frac{F_i}{N}\times 100 = \frac{1500}{5000}\times 100 = 30\%$$

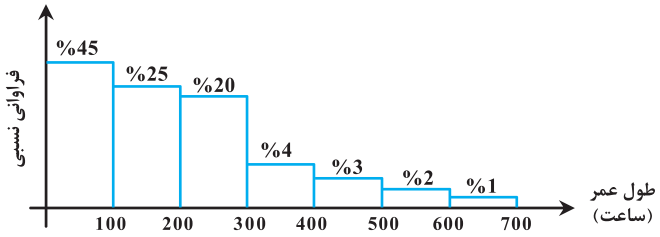
$$\text{سهم گروه خونی } O^+ = \frac{F_i}{N}\times 100 = \frac{1500}{5000}\times 100 = 30\%$$

$$\text{سهم گروه خونی } O^- = \frac{F_i}{N}\times 100 = \frac{1000}{5000}\times 100 = 20\%$$

ج) نمودار پارتو (Pareto chart)

در نمودار پارتو، فراوانی نسبی هر موضوع روی محور عمودی و نوع آن روی محور افقی آورده می‌شود. نمودار پارتو همیشه به ترتیب نزولی فراوانی‌ها رسم می‌شود. پر وقوع‌ترین موضوع در سمت چپ قرار می‌گیرد. توجه کنید که این نمودار سه محور دارد که محور سوم آن نشان دهنده‌ی فراوانی‌های نسبی تجمعی است. این نمودار برای تحلیل موجودی انبار کالا، نواقص سیستمها، توزیع درآمد و توزیع کارمندان سازمانها، کاربرد فراوانی دارد.

مثال ۱۳۶: **توزیع طول عمر ۱۰۰۰ تولید نوعی قطعه ترانزیستوری که در یک کارخانه تولید شده است به شکل زیر است. اگر بدانیم ۲۰% از قطعه‌ها از درجه کیفیت کمتری برخوردار می‌باشند و دارای طول عمر پایین هستند پس این تعداد کمتر از چه مدت عمر می‌کنند؟** (محیط زیست ـ سراسری ۸۷)



۱) ۲۰

۲) ۴۴/۴

۳) ۵۰/۵

۴) ۷۰

پاسخ: گزینه «۲» با کمی دقت از روی نمودار، جدول فراوانی را به‌صورت زیر می‌کشیم:

طول عمر (ساعت)	$c-100$	$100-200$	$200-300$	$300-400$	$400-500$	$500-600$	$600-700$
f_i (فراوانی نسبی)	%۴۵	%۲۵	%۲۰	%۴	%۳	%۲	%۱

در سؤال از ما صدک ۲۰ام طول عمرها خواسته شده است یعنی نقطه‌ای که ۲۰% از قطعات کمتر یا مساوی آن نقطه عمر می‌کنند برای بدست آوردن صدک ۲۰ام به صورت زیر عمل می‌کنیم. توجه کنید اولین طبقه‌ای که فراوانی تجمعی نسبی آن بزرگ‌تر یا مساوی ۰/۲ است طبقه‌ی اول است بنابراین صدک بیستم در طبقه‌ی اول قرار دارد. از رابطه چندک‌ها در داده‌های پیوسته خواهیم داشت:

$$P_{20} = L + \frac{0/2 - fC_{i-1}}{f_i}\times I = 0 + \frac{0/2-0}{0/45}\times 100 = \frac{20}{0/45} = 44/4$$

تحلیل اکتشافی داده‌ها (Exploratory data Analysis)

شناسایی انفرادی داده‌ها را تحلیل اکتشافی داده‌ها می‌گویند. نمودارهایی در مراحل اولیه‌ی داده‌ها به‌عنوان نمودارهای تحلیل اکتشافی رسم می‌شوند در اینجا به دو نمونه از آن‌ها اشاره می‌کنیم.

۱. نمودار شاخه و برگ (ساقه و برگ)، (Stem and leaf display)

فرض کنید $x_1, x_2, \ldots, x_n$ مشاهدات باشند و هر مشاهده دست کم دورقمی باشد. برای تهیه‌ی نمودار شاخه و برگ، ارقام را به دو بخش تقسیم می‌کنیم: شاخه شامل یک یا چند رقم اولیه و برگ آن‌ها شامل رقم‌های باقیمانده است.

مثال ۱۳۷: برای داده‌های روبرو، نمودار شاخه و برگ را رسم کنید.

۰,۱۷,۱۳,۱۳,۱۸,۱۰,۱۹,۵,۷,۴,۱,۲۱,۲۵,۲۴,۲۹,۳۵

پاسخ: رقم‌های یکان برگ و رقم‌های دهگان به عنوان شاخه هستند. بنابراین شکل به صورت روبرو خواهد شد:

شاخه	برگ
۰	۵,۴,۷,۰,۱
۱	۷,۳,۳,۸,۹,۰
۲	۱,۵,۴,۹
۳	۵

مثال ۱۳۸: داده‌های آماری به صورت نمودار شاخه و برگ زیر داده شده است. با حذف ۲۵ درصدی پایین و ۲۵ درصدی بالا، تفاضل میانه از میانگین پیراسته کدام است؟ (مدیریت و حسابداری ـ سراسری ۹۳)

شاخه	برگ					
۷	۱	۱	۲	۴	۷	۹
۸	۰	۰	۳	۵	۶	۷
۹	۲	۴	۵	۵	۷	۸

۱) ۰/۴ ۲) ۰/۲ ۳) ۰/۳ ۴) ۰/۱

پاسخ: گزینه «۳» داده‌ها را از نمودار بیرون می‌کشیم، اکنون می‌توانیم برای راحتی از تک‌تک داده‌ها ۸۰ واحد کم کنیم تا مقیاس کوچک شود.

۷۱,۷۱,۷۲,۷۴,۷۷,۷۹,۸۰,۸۰,۸۳,۸۵,۸۶,۸۷,۹۲,۹۴,۹۵,۹۵,۹۷,۹۸

$x_i - ۸۰ : -۹,-۹,-۸,-۶,-۳,-۱,۰,۰,۳,۵,۶,۷,۱۲,۱۴,۱۵,۱۵,۱۷,۱۸$

تعداد داده‌ها ۱۸ تا می‌باشد، بنابراین $n = ۱۸$ زوج است. پس میانه $Me = \frac{x_۹ + x_{۱۰}}{۲}$ است که برابر است با: $Me = \frac{۳+۵}{۲} = ۴$. اکنون $\frac{۱}{۴}$ داده‌ها از بالا و پایین حذف می‌کنیم (میانگین‌گیری پیراسته)

$$[np] = [۱۸ \times \frac{۱}{۴}] = [۴/۵] = ۴$$

پس ۴ تا از بالا و پایین حذف کرده میانگین‌گیری می‌کنیم:

$x_i : -۳,-۱,۰,۰,۳,۵,۶,۷,۱۲,۱۴$

$$\mu = \frac{\sum_{i=۱}^{N} x_i}{N} = \frac{-۳+-۱+۰+۰+۳+۵+۶+۷+۱۲+۱۴}{۱۰} = \frac{۴۳}{۱۰} = ۴/۳ \Rightarrow \text{مقدار خواسته شده} = \mu - Me = ۴/۳ - ۴ = \underline{۰/۳}$$

۲. نمودار جعبه‌ای (Box Plot)

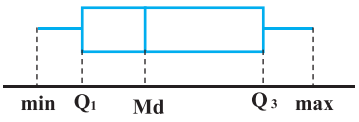
مراحل رسم این نمودار به شرح زیر است:

۱ـ داده‌های خام را به صورت غیر نزولی مرتب می‌کنیم، سپس یک خط افقی را چنان مدرج می‌کنیم که بتوان همه‌ی داده‌ها را روی آن نشان داد.

۲ـ میانه و چارکهای اول و سوم داده‌ها را محاسبه می‌کنیم.

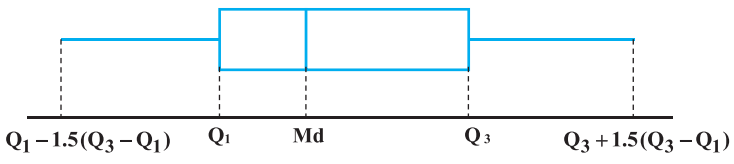
۳ـ بالای خط مدرج شده مستطیلی رسم می‌کنیم که طول آن برابر با $Q_۳ - Q_۱$ بوده و از نقطه $Q_۱$ شروع و به $Q_۳$ ختم شود، عرض مستطیل به اندازه‌ی معقول در نظر گرفته می‌شود. این مستطیل را جعبه یا باکس می‌نامیم.

۴ـ اندازه میانه را به صورت خطی به موازات عرض مستطیل رسم نموده و مستطیل را به وسیله یک خط منقطع به موازات خط مدرج شده به دو قسمت تقسیم می‌کنیم (شکل روبرو را مشاهده کنید).



۵ـ مرزهای داخلی و خارجی داده‌ها به صورت زیر تعیین می‌شوند:

$$\text{مرز داخلی پایین} = Q_۱ - ۱/۵(Q_۳ - Q_۱)$$
$$\text{مرز داخلی بالا} = Q_۳ + ۱/۵(Q_۳ - Q_۱)$$
$$\text{مرز خارجی پایین} = Q_۱ - ۳(Q_۳ - Q_۱)$$
$$\text{مرز خارجی بالا} = Q_۳ + ۳(Q_۳ - Q_۱)$$



۶ـ هر داده‌ای که خارج از مرزهای داخلی قرار گرفته باشد را یک داده‌ی پرت می‌نامند و چنانچه بین مرزهای داخلی و خارجی قرار گیرد، آن را داده‌ی پرت ضعیف نامیده و با علامت o نشان می‌دهیم و چنانچه خارج از مرزهای خارجی قرار گیرد آن را داده‌ی پرت قوی نامیده و با علامت • نشان می‌دهیم.

با استفاده از نمودار جعبه‌ای می‌توان اطلاعات زیر را در مورد داده‌ها کسب نمود:

۱ـ اگر میانه نزدیک وسط مستطیل (جعبه) باشد، توزیع داده‌ها تقریباً متقارن است.

۲ـ اگر میانه در طرف چپ وسط مستطیل باشد، توزیع چوله به راست و اگر میانه در طرف راست وسط مستطیل قرار گیرد، توزیع چوله به چپ است.

۳ـ در مقایسه‌ی نمودار جعبه‌ای دو مجموعه از داده‌ها می‌توان پراکندگی آن‌ها را با توجه به طول مستطیل‌های نمودار با یکدیگر مقایسه نمود. مستطیلی که طول بزرگتری دارد، داده‌های آن دارای پراکندگی بیشتری است.

مثال ۱۳۹: داده‌های آماری به صورت شاخه و برگ زیر، نشان داده شده است. در نمودار جعبه‌ای این داده‌ها، واریانس داده‌های داخل جعبه، تقریباً کدام است؟

(مدیریت و حسابداری ـ سراسری ۹۴)

شاخه	برگ						
۴	۱	۱	۲	۴	۶	۸	
۵	۰	۲	۲	۳	۵	۶	
۶	۰	۰	۱	۴	۵	۷	۹

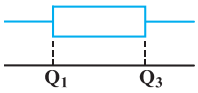
۱) ۱۵/۱

۲) ۱۴/۹

۳) ۱۵/۳

۴) ۱۶/۲

☑ **پاسخ:** گزینه «۳» نمودار جعبه‌ای به فرم زیر است:



در واقع سؤال واریانس بین چارک اول و سوم را خواسته است. (داخل جعبه)

از روی نمودار شاخه و برگ داده شده ابتدا چارک اول و سوم را به دست می‌آوریم. برای چارک اول، میانه کل داده‌ها را به دست آورده و مجدداً میانه را در ۵۰٪ اول داده‌ها مشخص می‌کنیم و در ۵۰٪ دوم باز هم میانه را بدست می‌آوریم که همان چارک سوم است.

$$۴۱,۴۱,۴۲,۴۴,\underset{Q_1}{۴۶},۴۸,۵۰,۵۲,۵۲,\underset{Me}{۵۳},۵۵,۵۶,۶۰,۶۰,\underset{Q_3}{۶۱},۶۴,۶۵,۶۷,۶۹$$

اکنون داده‌ها بین چارک اول و سوم عبارتند از:

$$۴۸,۵۰,۵۲,۵۲,۵۳,۵۵,۵۶,۶۰,۶۰$$

می‌دانیم که اضافه یا کم کردن تأثیری در واریانس ندارد. بنابراین می‌توانیم با کم کردن عددی دلخواه مقیاس داده‌ها را کوچک کنیم. از هر کدام از داده‌ها ۵۳ واحد کم می‌کنیم تا داده‌های جدید با مقیاس کوچک‌تر و بهتر حاصل شود.

$$-۵,-۳,-۱,-۱,۰,۲,۳,۷,۷ \Rightarrow \sigma^۲ = \frac{\sum x_i^۲}{N} - \mu^۲$$

$$\sigma_x^۲ = \frac{۲۵+۹+۱+۱+۴+۹+۴۹+۴۹}{۹} - (\frac{۹}{۹})^۲ = ۱۶/۳۳ - ۱ = ۱۵/۳۳$$

مثال ۱۴۰: کدام‌یک از نمودارهای ذیل برای نمایش متغیرهای کیفی کاربرد دارد؟

(مدیریت ـ دکتری ۹۵)

۱) شاخه و برگ　　۲) هیستوگرام　　۳) جعبه‌ای　　۴) پارتو

☑ **پاسخ:** گزینه «۴» برای رسم داده‌های کمّی (فاصله‌ای و نسبی) از نمودارهای ۱ـ هیستوگرام ۲ـ چندضلعی ۳ـ شاخ و برگی ۴ـ جعبه‌ای استفاده می‌شود و برای رسم داده‌های کیفی (اسمی و رتبه‌ای) از نمودارهای ۱ـ ستونی (میله‌ای) و ۲ـ دایره‌ای استفاده می‌شود.

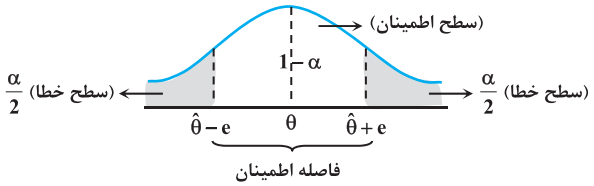
درسنامه (۲): برآورد فاصله‌ای

در بخش قبل میانگین نمونه $(\bar{X})$ یک برآوردگر خوب برای میانگین جامعه (μ) و همچنین واریانس نمونه (S^2) و نسبت نمونه $(\bar{P})$ برآوردگرهای خوبی برای واریانس و نسبت کل جامعه بودند. اما میانگین، واریانس و نسبت نمونه از نمونه‌ای به نمونه‌ی دیگر تغییر می‌کند، لذا نمی‌توان دقیقاً ادعا کرد که p, σ^2, μ دقیقاً برابر با $\bar{P}, S^2, \bar{X}$ است. بنابراین برای آنکه برای تمام نمونه‌ها بتوانیم برآورد دقیقی از پارامترهای جامعه داشته باشیم، یک برآورد فاصله‌ای به نام فاصله‌ی اطمینان تعریف می‌کنیم. که این فاصله به صورت $(\hat{\theta}-e, \hat{\theta}+e)$ است و به e خطای برآورد گفته می‌شود. در برآورد فاصله‌ای به این علت که یک فاصله با درجه‌ی خطا یا اطمینان برای پارامتر به دست می‌آوریم درجه‌ی صحت برآورد مشخص می‌شود.

برآورد فاصله‌ای (فاصله اطمینان) یک فاصله است که بیان می‌دارد که با درصد اطمینان $(1-\alpha)100$ پارامتر جامعه (θ) در فاصله‌ای به صورت $(\hat{\theta}-e, \hat{\theta}+e)$ قرار می‌گیرد، مثلاً اگر برای میانگین جامعه یک فاصله‌ی اطمینان ۹۵٪ درصد به صورت $(-2,4)$ بدست آید، این به مفهوم آن است که با اطمینان ۹۵ درصد میانگین جامعه در فاصله $(-2,4)$ قرار دارد. توجه کنید که درصد اطمینان برای ساختن فواصل اطمینان معمولاً ۹۰٪ و ۹۵٪ و یا ۹۹٪ در نظر گرفته می‌شود که به آن‌ها سطوح اطمینان گفته می‌شود. در اینصورت سطح خطاها به صورت ۱۰٪ و ۵٪ و ۱٪ خواهد بود. رابطه‌ی احتمالی برای فاصله‌ی اطمینان به صورت زیر است:

$$P(\hat{\theta}-e<\theta<\hat{\theta}+e)=1-\alpha$$

به شکل مقابل توجه کنید:

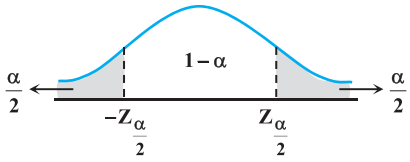


این شکل بیان می‌دارد که با احتمال $(1-\alpha)$ اطمینان داریم که فاصله‌ی $(\hat{\theta}-e, \hat{\theta}+e)$ پارامتر θ را در بر می‌گیرد و با احتمال $(\frac{\alpha}{2}+\frac{\alpha}{2}=\alpha)$ احتمال می‌دهیم که پارامتر (θ) در فاصله‌ی $(\hat{\theta}-e, \hat{\theta}+e)$ قرار نگیرد و همچنین می‌خواهیم به اندازه‌ی مقدار e، برآورد θ یعنی $\hat{\theta}$ از پارامتر θ تفاوت داشته باشد. بنابراین $1-\alpha$ سطح اطمینان و α سطح خطا و خطای برآورد $e=|\hat{\theta}-\theta|$ است. اکنون در اینجا به معرفی فواصل اطمینان برای پارامترهای جامعه μ، σ^2 و p در شرایط مختلف می‌پردازیم.

مثال ۴۳: هر چقدر سطح خطای α کاهش یابد کدام گزینه صحیح است؟(بقیه عوامل ثابت)

۱) سطح اطمینان $1-\alpha$ افزایش می‌یابد.
۲) خطای برآورد افزایش می‌یابد.
۳) ضریب اطمینان افزایش می‌یابد.
۴) هر سه گزینه

پاسخ: گزینه «۴» به شکل مقابل دقت کنید:



هر چقدر سطوح هاشور خورده کمتر شود، سطح هاشور نخورده افزایش می‌یابد.

و مقادیر $Z_{\frac{\alpha}{2}}$ افزایش می‌یابد و خطای برآورد (e) نیز افزایش می‌یابد. به مثال‌های عددی زیر توجه کنید:

(به مقادیر Z_α یا $Z_{\frac{\alpha}{2}}$ ضریب اطمینان گفته می‌شود).

$$\begin{cases}\alpha=0/05 \Rightarrow 1-\alpha=1-0/05=0/95\\ \frac{\alpha}{2}=0/025 \Rightarrow Z_{0/025}=1/96\end{cases} \Rightarrow \text{خطای برآورد میانگین: } e=Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}$$

$$\begin{cases}\alpha=0/01 \Rightarrow 1-\alpha=1-0/01=0/99\\ \frac{\alpha}{2}=0/005 \Rightarrow Z_{0/005}=2/58\end{cases} \Rightarrow \text{خطای برآورد میانگین: } e=Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}$$

پس اگر $\alpha=0/05$ به $\alpha=0/01$ کاهش یابد، هر سه گزینه‌ی دیگر افزایش می‌یابند.

مثال ۴۴: فرض کنید X دارای تابع چگالی $f_x(x)=\frac{1}{\theta}e^{-\frac{1}{\theta}x}$, $x>0$ است. مقدار c چقدر باشد تا فاصله‌ی $(0,cX)$ یک فاصله اطمینان ۹۰٪ برای θ در نظر گرفته شود؟

(علوم اقتصادی ـ دکتری ۹۴)

۱) $Ln(9)$
۲) $Ln(\frac{10}{9})$
۳) $-\frac{1}{Ln(0/9)}$
۴) $-Ln(0/9)$

☑ **پاسخ:** گزینه «۳» مفهوم فاصله اطمینان عبارت است از: $P(a<\theta<b)=1-\alpha \xrightarrow{\text{طبق گفته مسئله}} P(0<\theta<CX)=0/90$

برای محاسبه احتمال در یک بازه روی آن بازه از تابع چگالی احتمال انتگرال‌گیری می‌کنیم:

$$\rho(x>\frac{\theta}{C})=0/90 \Rightarrow \int_{\frac{\theta}{C}}^{\infty}\frac{1}{\theta}e^{-\frac{1}{\theta}x}dx=0/90 \Rightarrow -e^{-\frac{x}{\theta}}\Big|_{\frac{\theta}{C}}^{\infty}=0/9 \Rightarrow e^{-\frac{1}{C}}=0/9$$

$$\Rightarrow -\frac{1}{C}=\ln 0/9 \Rightarrow C=\frac{-1}{\ln 0/9}$$

مثال ۴۵: **فرض کنید X طول عمر یک لامپ باشد که دارای توزیع یکنواخت روی فاصله $(0,\theta)$ می‌باشد و θ پارامتر نامعلوم است. مقدار c چقدر باشد تا فاصله تصادفی $(\frac{c}{2X},\frac{c}{X})$ مقدار $\frac{1}{\theta}$ را با احتمال ۰/۹ شامل شود؟** (علوم اقتصادی ـ سراسری ۹۶)

۱) ۰/۹ ۲) ۰/۱ ۳) ۱/۸ ۴) ۱

☑ **پاسخ:** گزینه «۳» طبق تعریف فاصله اطمینان $p(a<x<b)=1-\alpha$ می‌باشد. رابطه داده شده را به این فرم تبدیل می‌کنیم:

$$p(\frac{c}{2x}<\frac{1}{\theta}<\frac{c}{x})=0/9 \Rightarrow p(\frac{c\theta}{2}<x<c\theta)=0/9$$

متغیر تصادفی x دارای توزیع یکنواخت بر روی فاصله $(0,\theta)$ است، بنابراین تابع چگالی آن $0<x<\theta$ و $f(x)=\frac{1}{\theta}$ می‌باشد. برای محاسبه احتمال روی بازه خواسته شده از تابع چگالی انتگرال‌گیری می‌کنیم:

$$\int_{\frac{c\theta}{2}}^{c\theta}\frac{1}{\theta}dx=0/9 \Rightarrow \frac{x}{\theta}\Big|_{\frac{c\theta}{2}}^{c\theta}=0/9 \Rightarrow (\frac{c\cancel{\theta}}{\cancel{\theta}}-\frac{c\cancel{\theta}}{2\cancel{\theta}})=0/9 \Rightarrow c-\frac{c}{2}=0/9 \Rightarrow \frac{1}{2}c=0/9 \Rightarrow c=1/8$$

توجه: در مسائل آمار استنباطی معمولاً ۳ سطح خطا و ۳ سطح اطمینان بیشتر مورد نظر است. جدول زیر این سطوح را مشخص کرده و ضرایب اطمینان آن‌ها را با توجه به جدول نرمال استاندارد مشخص کرده است. به دانشجویان توصیه می‌شود این جدول را به‌خاطر بسپارند:

ضریب اطمینان	سطح خطا	سطح اطمینان
$Z_{\frac{\alpha}{2}}=1/645$; $Z_{\alpha}=1/28$	$\alpha=0/1$; $\frac{\alpha}{2}=\frac{0/1}{2}=0/05$	$(1-\alpha)=0/90$
افزایش ↓ $Z_{\frac{\alpha}{2}}=1/96$; $Z_{\alpha}=1/645$	کاهش ↓ $\alpha=0/05$; $\frac{\alpha}{2}=\frac{0/05}{2}=0/025$	افزایش ↓ $(1-\alpha)=0/95$
$Z_{\frac{\alpha}{2}}=2/58$; $Z_{\alpha}=2/32$	$\alpha=0/01$; $\frac{\alpha}{2}=\frac{0/01}{2}=0/005$	$(1-\alpha)=0/99$

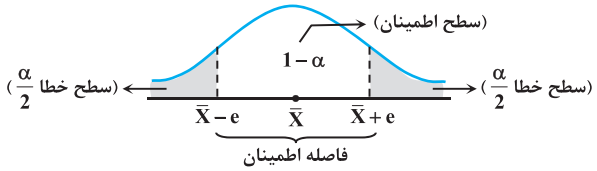
نکته ۱: هرگاه مسأله مقدار α را مشخص نکرد پیش‌فرض بر آن است که $\alpha=0/05$ است.

فاصله اطمینان یا برآورد فاصله‌ای برای میانگین جامعه μ

یک فاصله‌ی اطمینان برای میانگین، به حالت‌های زیر بوجود می‌آید:

۱. جامعه نرمال و واریانس جامعه σ^2 معلوم

در جامعه‌ای نرمال با میانگین μ و واریانس معلوم σ^2، نمونه‌ای تصادفی به حجم n اختیار می‌کنیم. اگر میانگین این نمونه n تایی $\overline{X}$ باشد یک فاصله‌ی اطمینان $100(1-\alpha)$ درصد برای میانگین جامعه به‌صورت زیر است:



$Z_{\frac{\alpha}{2}}$: از روی جدول نرمال استاندارد بدست می‌آید.

$$\bar{x}-\frac{\sigma}{\sqrt{n}}Z_{\frac{\alpha}{2}}<\mu<\bar{x}+\frac{\sigma}{\sqrt{n}}Z_{\frac{\alpha}{2}}$$

مثال ۴۶: از جامعه آماری در نمونه‌ای به حجم ۴۰۰ مقدار میانگین ۵۲ و انحراف معیار ۱۶ محاسبه شده است. با احتمال ۹۵ درصد میانگین این جامعه در کدام بازه است؟ (مدیریت و حسابداری ـ سراسری ۹۲)

۱) (۵۰/۴۳۲,۵۳/۵۶۸)　۲) (۵۰/۳۲۴,۵۳/۶۷۶)　۳) (۵۰/۷۱۶,۵۳/۲۸۴)　۴) (۵۱/۰۲۴,۵۲/۹۷۶)

☑ **پاسخ:** گزینه «۱» از رابطه فاصله اطمینان برای میانگین استفاده می‌کنیم: $\bar{x}-\frac{\sigma}{\sqrt{n}}.Z_{\frac{\alpha}{2}}<\mu<\bar{x}+\frac{\sigma}{\sqrt{n}}.Z_{\frac{\alpha}{2}}$

در این مسأله $n=400$ ، $\bar{X}=52$ و $\sigma=16$ و $\frac{\alpha}{2}=\%25\rightarrow\alpha=\%5\rightarrow1-\alpha=\%95$ مقادیر را در فرمول بالا جایگذاری می‌کنیم:

$$52-\frac{16}{\sqrt{400}}\times2<\mu<52+\frac{16}{\sqrt{400}}\times2\Rightarrow52-\frac{32}{20}<\mu<52+\frac{32}{20}\Rightarrow50/432<\mu<53/568$$

مثال ۴۷: از توزیع نرمال یک نمونه به حجم ۱۶ اختیار شده است. می‌دانیم واریانس جامعه ۶۴ می‌باشد. اگر میانگین نمونه ۱۰ باشد، با اطمینان ۹۵٪ میانگین جامعه در چه فاصله‌ای قرار دارد؟ $\int_{-4}^{-2}f(y)dy\approx0/025$ (محیط زیست ـ سراسری ۸۵)

۱) [۰,۶]　۲) [۶,۸]　۳) [۶,۱۴]　۴) [۸,۱۴]

☑ **پاسخ:** گزینه «۳» توزیع جامعه نرمال و واریانس جامعه معلوم است. بنابراین مقدار نمونه (n) هر چه باشد توزیع $\bar{X}$ نرمال خواهد بود و فاصله اطمینان μ بنابر مورد (الف) عبارت است از: $\mu:\bar{x}\pm Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\Rightarrow\mu:10\pm2\frac{\sqrt{64}}{\sqrt{16}}=10\pm2\times\frac{8}{4}=10\pm4=(6,14)$

$\bar{x}=10\ ,\ \sigma^2=64\ ,\ n=16\ ,\ \alpha=0/05\rightarrow Z_{\frac{\alpha}{2}}=Z_{0/025}\simeq2$

مثال ۴۸: در سطح معنی‌داری ۵ درصد، برای m مشاهده از جامعه‌ای نرمال با واریانس ۴، حدود اطمینان میانگین جامعه (۵ و ۳) به دست آمده است. اگر اندازه نمونه را ۴ برابر کنیم، عرض فاصله اطمینان جدید با همان سطح اطمینان قبلی برابر خواهد بود با: (حسابداری و مدیریت ـ دکتری ۹۳)

۱) ۳　۲) ۲　۳) ۱　۴) ۴

☑ **پاسخ:** گزینه «۳» در یک توزیع نرمال با واریانس معلوم فاصله اطمینان به صورت: $(\bar{x}-\frac{\sigma}{\sqrt{m}}.Z_{\frac{\alpha}{2}}\ \text{و}\ \bar{x}+\frac{\sigma}{\sqrt{m}}.Z_{\frac{\alpha}{2}})$

می‌باشد که عرض این فاصله اطمینان به صورت: $(\bar{x}+\frac{\sigma}{\sqrt{m}}Z_{\frac{\alpha}{2}})-(\bar{x}-\frac{\sigma}{\sqrt{m}}.Z_{\frac{\alpha}{2}})=\frac{2\sigma}{\sqrt{m}}Z_{\frac{\alpha}{2}}$

می‌باشد اکنون m تبدیل به $4m$ شود عرض به صورت روبرو تبدیل می‌شود: $\frac{2\sigma}{\sqrt{4m}}.Z_{\frac{\alpha}{2}}\Rightarrow\frac{\not{2}\sigma}{\not{2}\sqrt{m}}.Z\frac{\alpha}{2}$

بنابراین عرض فاصله اطمینان جدید نصف خواهد شد. اما توجه کنید که عرض فاصله اطمینان قبلی $5-3=2$ بوده است بنابراین عرض جدید نصف ۲ یعنی ۱ است.

تذکر۱: در فاصله‌ی اطمینان بالا به مقدار $e=Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}$ خطای نمونه‌گیری یا خطای حدی یا خطای برآورد گفته می‌شود. همچنین به راحتی می‌توانیم با محاسبه‌ی تفاضل حد بالا و پایین فاصله‌ی اطمینان، طول فاصله را به دست آوریم.

طول فاصله اطمینان = حد بالا − حد پایین $=(\bar{x}+Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}})-(\bar{x}-Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}})=2Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}=2e$

بنابراین طول فاصله‌ی اطمینان برابر با $2e$ است.

مثال ۵۳: تابع مصرف از نمونه ای با حجم N=۶۶ به صورت زیر برآورد گردیده است:

$\hat{C}_t = ۸۵ + ۰/۸Y_{dt}$ اگر ضریب همبستگی بین درآمد قابل تصرف (Y_{dt}) و مصرف (C_t) برابر ۰/۶ برآورد شده باشد، مقدار آماره آزمون معنی دار بودن میل نهایی به مصرف کدام یک از گزینه‌های زیر است؟ (علوم اقتصادی ـ آزاد ۸۸)

۱) $-۰/۶$ ۲) ۱۴ ۳) ۱/۴ ۴) +۶

☑ **پاسخ:** گزینه «۴» مقدار آماره آزمون برای معنی دار بودن رگرسیون عبارت است از: $t = \frac{r\sqrt{n-۲}}{\sqrt{۱-r^۲}}$ بنابراین:

$$t = \frac{۰/۶\sqrt{۶۶-۲}}{\sqrt{۱-(۰/۶)^۲}} = ۶$$

مثال ۵۴: در صورتی‌که در یک رگرسیون با دو متغیر مستقل که دارای ۲۳ مورد است $(N = ۲۳)$، همبستگی متغیرهای مستقل با متغیر وابسته (۰/۸) باشد، مقدار عددی F برابر است با: (طراحی و برنامه‌ریزی شهری و منطقه‌ای ـ سراسری ۸۳)

۱) ۱/۷۸۲ ۲) ۸/۷۸ ۳) ۱۲/۷۸ ۴) ۱۷/۸

☑ **پاسخ:** گزینه «۴» آماره آزمون F با استفاده از ضریب تعیین عبارت است از:

$$F_{k-۱,n-k} = \frac{\frac{R^۲}{k-۱}}{\frac{۱-R^۲}{n-k}}$$

$R^۲$: ضریب تعیین متغیرهای مستقل با متغیر وابسته است. k: تعداد متغیرهای مستقل و وابسته.

که در این‌جا: $k = ۳$ است، دو متغیر مستقل و یک متغیر وابسته، مقادیر را جایگذاری می‌کنیم:

$$F_{۲,۲۰} = \frac{\frac{۰/۶۴}{۳-۱}}{\frac{۱-۰/۶۴}{۲۳-۳}} = \frac{۰/۳۲}{۰/۰۱۸} = ۱۷/۷۸ \simeq ۱۷/۸ \quad , \quad R^۲ = (۰/۸)^۲ = ۰/۶۴$$

۳ـ آزمون صفر بودن زیر مجموعه‌ای از ضرایب رگرسیون (آزمون F جزئی):

فرض کنید $R_k^۲$ ضریب همبستگی چندگانه باشد وقتی همه‌ی k متغیر در مدل رگرسیونی باشند و $R_m^۲$ ضریب همبستگی چندگانه باشد وقتی m متغیر مشخص داخل مدل باشد پس فرضیه‌ی صفر بیان می‌کند که $k-m$ متغیر تعیین شده، دارای ضریب رگرسیونی صفر است.

آماره آزمون $F = \frac{\frac{R_k^۲ - R_m^۲}{k-m}}{\frac{۱-R_k^۲}{n-k-۱}}$ (ناحیه بحرانی) $F > F_{\alpha,k-k,n-k-۱}$

اگر هریک از آزمون‌های t مربوط به ضرایب انفرادی رگرسیون معنی‌دار باشد ($H_۰$ رد شود) آنگاه F برای آزمون معنی‌داری مدل نیز معنی‌دار است ($H_۰$ رد می‌شود)

شکل ماتریسی رگرسیون خطی چندگانه:

مدل رگرسیون $y = \beta_۰ + \beta_۱x_۱ + \beta_۲x_۲ + ... + e$ را می‌توان به شکل ماتریسی $y = x\beta + e$ نوشت که y بردار $n\times ۱$ از مشاهدات پاسخ، x یک ماتریس $n\times p$ از متغیرهای مستقل، β یک بردار $p\times ۱$ از ضرایب رگرسیونی و e بردار $n\times ۱$ از خطاهای تصادفی است.

برآورد کمترین مربعات بردار β: برآوردگر حداقل مربعات بردار β به شرط **وجود** ماتریس XX' به‌صورت مقابل می‌باشد: $\hat{\beta} = (X'X)^{-۱}X'y$

نکته ۷: می‌توان دید که $\hat{y} = x\tilde{\beta}$ یعنی $\hat{y} = X(X'X)^{-۱}X'y$ که به‌صورت $\hat{y} = Hy$ نمایش داده می‌شود که H ماتریس برازش نامیده می‌شود.

$$e = y - \hat{y} = y - x\hat{\beta} = y - Hy = (I-H)y$$

خواص برآوردگرهای حداقل مربعات:

۱. $\hat{\beta}$ برآورد ناریب بردار β است یعنی $E(\hat{\beta}) = \beta$ ۲. کوواریانس بردار $\hat{\beta}$ یک ماتریس متقارن می‌باشد. ۳. $Cov(\hat{\beta}) = \sigma^۲(X'X)^{-۱}$

معیارهای انتخاب مدل رگرسیونی

پس از برآوردکردن ضرایب رگرسیونی، معیارهایی وجود دارد که باید تست شوند؛ چرا که آنها فرضیات مدل‌های رگرسیونی را آزمایش می‌کنند و نشان می‌دهند که مدلی که به‌دست آمده است مدل مناسبی است و یا خطاهایی در آن‌ها وجود دارد.

۱ـ ضریب تعیین چندگانه: یکی از معیارهای ارزیابی مدل رگرسیون ضریب تعیین چندگانه است که به‌صورت زیر تعریف می‌شود:

$$R^2 = \frac{SSReg}{SS_y} = \frac{SS_y - SSE}{SS_y} = 1 - \frac{SSE}{SS_y}$$

که در اینجا $SSReg$ مجموع مربعات رگرسیونی و SSE مجموع مربعات خطا و SS_y مجموع مربعات انحراف داده‌ها از میانگین نام دارند. R^2 به‌عنوان معیاری برای کاهش در تغییرپذیری Y معرفی می‌شود و $0 \le R^2 \le 1$ ولی R^2 بزرگ لزوماً یک رگرسیون خوب را نتیجه نمی‌دهد، چرا که با اضافه کردن یک متغیر مستقل به مدل، R^2 افزایش می‌یابد. لذا از آماره‌ای به‌نام R^2 اصلاح شده (تصحیح شده) به‌صورت $\bar{R}^2_k = 1 - \frac{n-1}{n-k-1}(1-R^2)$ استفاده می‌کنند.

در یک مدل رگرسیونی مقدار $R^2 = \%95$ شده است این به مفهوم آن است که $\%95$ از تغییرپذیری Y به‌وسیله متغیرهای مستقل که در مدل وجود دارند، توضیح داده شده‌اند.

مثال ۵۵: اگر ضریب همبستگی بین دو متغیر $0/6$ و دو متغیر دیگر $0/3$ باشد، می‌توان گفت همبستگی دو متغیر اول: «چند برابر قوی‌تر» از دو متغیر دوم است؟ (مدیریت و حسابداری ـ دکتری ۹۲)

۱) دو　　۲) چهار　　۳) نه　　۴) سه

☑ **پاسخ:** گزینه «۲» یکی از معیارهای مهم برای مقایسه دو متغیر ضریب تعیین بین آنها است و نباید براساس ضریب همبستگی آنها را مقایسه کرد:

$$\begin{matrix} r_1 = 0/6 \Rightarrow r_1^2 = 0/36 \\ r_2 = 0/3 \Rightarrow r_2^2 = 0/9 \end{matrix} \Rightarrow$$ همبستگی دو متغیر اول ۴ برابر همبستگی دو متغیر دوم است

۲ـ خود همبستگی (آزمون دوربین ـ واتسون): یکی از فرض‌های مهم در رگرسیون آن است که خطاهای مدل ε_iها متغیرهای تصادفی ناهمبسته باشند. در بسیاری از کاربردهای تحلیل رگرسیون با داده‌هایی سروکار داریم که ممکن است این فرض برای آن‌ها برقرار نباشد. برای آزمون کردن این فرضیه از آزمونی به نام آزمون دوربین ـ والتسون استفاده می‌کنیم. دوربین ـ واتسون یک آزمون است که فرضیه‌ها، آماره‌ی آزمون و ناحیه‌ی بحرانی آن به‌صورت زیر است:

$$\begin{cases} H_0 : \rho = 0 & \text{(همبستگی بین خطا وجود ندارد)} \\ H_1 : \rho > 0 & \text{(همبستگی بین خطاها وجود دارد)} \end{cases}$$

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \quad \text{(آماره‌ی آزمون)}$$

مقدار آماره‌ی آزمون D، دارای حداقل مقدار صفر و حداکثر مقدار ۴ است. اگر فرضیه‌ی H_0 درست باشد مقدار مورد انتظار آماره‌ی آزمون برابر با ۲ خواهد بود. ناحیه‌ی بحرانی این آزمون به‌صورت زیر است:

اگر $D > du$ باشد فرض صفر $H_0 : \rho = 0$ رد نمی‌شود.　　اگر $D < dL$ باشد فرض صفر $H_0 : \rho = 0$ رد می‌شود.

اگر $dL < D < du$ باشد، آزمون بی‌نتیجه خواهد بود.

به du و dL مقادیر بحرانی این آزمون گفته می‌شود که از جدولی به نام جدول دوربین واتسون مشخص می‌شوند.

➢ **توجه:** وجود خود همبستگی در خطاها آثار سوء متعددی دارد که در زیر به آن‌ها اشاره خواهیم کرد:

۱) ضرایب رگرسیونی برآورد شده اگر چه هنوز نااریب هستند، اما دیگر برآوردهایی با کمترین واریانس نیستند، بنابراین با تورم واریانس ضرایب، روبرو خواهیم شد.

۲) فواصل اطمینان برای ضرایب کوچکتر، از آن‌چه در واقع باید باشند بزرگ‌تر خواهند شد و همچنین $\hat{\sigma}^2 = S_e^2$ به طور کلی کم برآورد خواهد شد.

۳) به دست آوردن فواصل اطمینان و آزمون فرضیه‌ها بر اساس توزیع‌های t و F دیگر مناسب نیست.

مثال ۵۶: در یک خروجی کامپیوتری مقدار آماره‌ی آزمون دوربین ـ واتسون برابر با ۲/۰۸ به‌دست آمده است اگر مقادیر بحرانی به‌ترتیب ۰/۹۵ و ۱/۵۴ به‌دست آمده باشد کدام گزینه صحیح است؟

۱) فرضیه همبستگی خطاها برابر با صفر رد می‌شود.
۲) فرضیه همبستگی خطاها برابر با صفر رد نمی‌شود.
۳) آزمون بی‌نتیجه است.
۴) نمی‌توان قضاوت کرد.

☑ **پاسخ:** گزینه «۲» از آنجائیکه مقدار آماره‌ی آزمون $D = ۲/۰۸$ از مقدار بحرانی ۱/۵۴ بزرگتر است، لذا فرض $H_۰ : \rho = ۰$ رد نمی‌شود.

مثال ۵۷: برای قضاوت در مورد همبستگی خطاها در یک مدل رگرسیون چند متغیره مقادیر $\sum_{t=۲}^{۲۰}(e_t - e_{t-۱})^۲ = ۳$ و $\sum_{t=۱}^{۲۰} e_t^۲ = ۲$ به‌دست آمده‌اند.

مقدار آماره‌ی آزمون مناسب کدام است؟

۱) ۱/۰۸ ۲) ۱/۵ ۳) ۱/۰۱ ۴) ۱

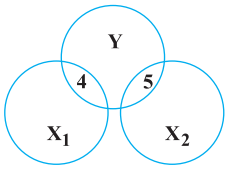
☑ **پاسخ:** گزینه «۲» طبق تعریف آماره‌ی آزمون دوربین ـ واتسون خواهیم داشت:

$$D = \frac{\sum_{t=۲}^{۲۰}(e_t - e_{t-۱})^۲}{\sum_{t=۱}^{۲۰} e_t^۲} = \frac{۳}{۲} = ۱/۵$$

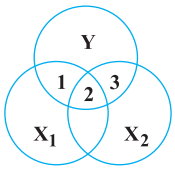
توجه: در حالتی که فرضیه‌ها به‌صورت مقابل باشند $\begin{cases} H_۰ : \rho = ۰ \\ H_۱ : \rho < ۰ \end{cases}$ آماره‌ی آزمون مناسب به‌صورت $D^* = ۴ - D$ می‌باشد و قضاوت و نواحی بحرانی به‌صورت قبل است.

۳ـ هم‌خطی: در اغلب مسائل رگرسیون متغیرهای مستقل با یکدیگر همبسته‌اند در مواردی که این همبستگی متقابل خیلی شدید باشد، گوئیم هم‌خطی وجود دارد که تأثیر بسیار بدی بر برآورد پارامترهای ضرایب رگرسیونی خواهد داشت و باعث تورم واریانس‌های برآوردگرها خواهد شد و روش حداقل مربعات خطا، برآوردهای بسیار ضعیفی از هر یک از پارامترهای مدل به دست خواهد داد.

به شکل‌ها توجه کنید.



عدم وجود هم‌خطی بین $X_۱, X_۲$



وجود هم‌خطی بین $X_۱, X_۲$

در شکل بالا ناحیه‌ی (۲) تغییرات Y را نشان می‌دهد که توسط هر دو متغیر $X_۱$ و $X_۲$ توضیح داده شده است که به دلیل وجود هم‌خطی باعث افزایش تغییرات واقعی Y است و به عبارت دیگر ناحیه (۲) دو بار در محاسبات واریانس Y محاسبه می‌شود و واریانس Y بیش از مقدار واقعی نشان داده خواهد شد.

مثال ۵۸: برای به دست آوردن معادله رگرسیون چند متغیره (چندگانه) خطی حاصل از ۱۰ متغیر مستقل، کدام مورد درست است؟

(مدیریت ـ دکتری ۹۴)

۱) بهترین ترکیب از متغیرهای مستقل، یک ترکیب حداقل سه‌تایی است.
۲) بهترین روش تهیه معادله رگرسیون، روش کل‌گرا (Enter) است.
۳) بهترین ترکیب از متغیرهای مستقل، یک ترکیب ده‌تایی است.
۴) بین متغیرهای مستقل هم‌خطی وجود نداشته باشد.

☑ **پاسخ:** گزینه «۴» یکی از موضوعات بسیار مهم در رگرسیون، هم‌خطی چندگانه است که باید بین متغیرهای مستقل همبستگی وجود داشته باشد؛ در مواردی که بین این متغیرها همبستگی متقابل خیلی شدید باشد، می‌گوییم هم‌خطی وجود دارد که تأثیر بسیار بدی بر برآورد پارامترهای ضرایب رگرسیونی خواهد داشت و باعث تورم واریانس این برآوردگرها خواهد شد.

درسنامه (۲): آنتروپی (Entropy)

کلمه آنتروپی به مفهوم بی‌نظمی است که اولین بار توسط ژول مطرح شد. این مفهوم در آمار توسط شانون در اواسط قرن بیستم معرفی شد. در آمار، آنتروپی به عنوان یک معیار عدم قطعیت به‌کار برده می‌شود. در مسائل مدیریتی واژه‌ی آنتروپی حاکی از تمایل سیستم‌ها به بی‌نظمی است. سیستم‌های بسته به مرور زمان از بین می‌روند و اصطلاحاً آنتروپی آنها مثبت است ولی سیستم‌های باز دارای آنتروپی منفی هستند.

❖ **تعریف:** اگر X یک متغیر تصادفی دلخواه با تابع چگالی احتمال $f(x)$ باشد تابع $H(X)=E(-\log f(X))$ را آنتروپی X یا توزیع X گویند.

(لگاریتم غالباً در پایه‌ی ۲ یا e منظور می‌شود)

لگاریتم بر مبنای ۲ مهم‌ترین معیار می‌باشد که وقتی بر مبنای ۲ است، آنتروپی براساس بیت می‌آید. یعنی وقتی $H(X)=2\,bit$ شود، می‌توانیم با پرسیدن ۲ سؤال که جواب آنها بله یا خیر است به جواب اصلی دست پیدا کنیم.

اگر X یک متغیر تصادفی گسسته با توزیع احتمال $i=1,2,...,n$ $P_i=f(x_i)=P(X=x_i)$ باشد، آنگاه آنتروپی X عبارت است از:

$$H(X)=E(-\log f(X))=-\sum f(x)\log f(x)=-\sum_{i=1}^{n}P_i\log P_i$$

به سادگی مشاهده می‌شود که تابع آنتروپی وابسته به احتمال‌های $P_1,P_2,...,P_n$ است. لذا هر چقدر آنتروپی بیشتر باشد، بی‌نظمی یا عدم قطعیت بیشتر است.

مثال ۵۹: کدام‌یک از عبارت‌های زیر «درجه نامعینی» یک آزمایش α را که k نتیجه ممکن $A_1,A_2,A_3,\cdots,A_k$ هم احتمال دارد. آنتروپی (Entropy) آزمایش α می‌نامند: (مدیریت صنعتی ـ آزاد ۸۸)

۱) $H(\alpha)=\underbrace{-k\log k-k\log k-\cdots-k\log k}_{k}$

۲) $H(\alpha)=\underbrace{k\log k+k\log k+\cdots+k\log k}_{k}$

۳) $H(\alpha)=\underbrace{\frac{1}{k}\log\frac{1}{k}+\frac{1}{k}\log\frac{1}{k}+\cdots+\frac{1}{k}\log\frac{1}{k}}_{k}$

۴) $H(\alpha)=\underbrace{-\frac{1}{k}\log\frac{1}{k}-\frac{1}{k}\log\frac{1}{k}-\cdots-\frac{1}{k}\log\frac{1}{k}}_{k}$

☑ **پاسخ:** گزینه «۴» طبق تعریف آنتروپی که در بالا گفته شد می‌توانیم نتیجه بگیریم که همه A_iها هم شانس هستند یعنی $P(A_i)=\frac{1}{k}$ بنابراین:

$$H(\alpha)=-\frac{1}{k}\log\frac{1}{k}-\frac{1}{k}\log\frac{1}{k}-\cdots-\frac{1}{k}\log\frac{1}{k}\qquad i=1,2,\cdots,k$$

مثال ۶۰: نتیجه آزمایش α توسط جدول زیر بیان می‌شود. آنتروپی این آزمایش کدام است؟ (مدیریت صنعتی ـ آزاد ۸۸)

x	A_1	A_2	A_3	
$P(A_i)$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$=1$

۱) $H(\alpha)=0/4581$　　۲) $H(\alpha)=0/4791$

۳) $H(\alpha)=0/4497$　　۴) $H(\alpha)=0/4381$

☑ **پاسخ:** گزینه «۱» با توجه به رابطه آنتروپی خواهیم داشت:

$$H(\alpha)=-\sum_{i=1}^{n}P_i\log P_i=-P_1\log P_1-P_2\log P_2-P_3\log P_3$$
$$=-0/4\log 0/4-0/4\log 0/4-0/2\log 0/2=(-0/4)(-0/398)-(0/4)(-0/398)-(0/2)(-0/699)=0/4581$$

مثال ۶۱: دو متغیر تصادفی گسسته X و Y با توزیع‌های زیر مفروض است، آنتروپی‌ها X و Y را محاسبه کنید.

$$P_i=P(X=x_i)=\frac{1}{4}\qquad i=1,2,3,4$$

$$q_i=P(Y=x_i)=\begin{cases}\frac{1}{64} & i=1,2,3\\ \frac{61}{64} & i=4\end{cases}$$

☑ **پاسخ:** از رابطه گفته شده بالا استفاده می‌کنیم، توجه کنید که متغیرهای تصادفی X و Y گسسته‌اند.

$$H(X)=-\sum_{i=1}^{4}P_i\log P_i=-\frac{1}{4}\sum_{i=1}^{4}\log_2^{\frac{1}{4}}=-\frac{1}{4}\times\sum_{i=1}^{4}(\log_2 1-\log_2 4)=-\frac{1}{4}\times\sum_{i=1}^{4}-\log_2 4=2$$

$$H(Y)=-\sum_{i=1}^{4}q_i\log_2^{q_i}=-\frac{1}{64}\sum_{i=1}^{3}\log_2^{\frac{1}{64}}-\frac{61}{64}\log_2^{\frac{61}{64}}=0/215$$