گزارش طرح پژوهشی

بررسی و تعیین روش‌های شناخت و اصلاح داده‌های دورافتاده در طرح آمارگیری از کارگاه‌های صنعتی

مجری طرح:

مجتبی گنجعلی

همکاران:

زهرا رضایی قهرودی

تابان باغ‌فلکی

مهناز کاظمی

گلستان احسنی

فرشاد روشن سنگاچین

پژوهشکده‌ی آمار

گروه پژوهشی طرح‌های فنی و روش‌های آماری

تابستان ۱۳۹۲

به نام خداوند جان و خرد

پیش‌گفتار

نقش و اهمیت بخش صنعت در جریان توسعه‌ی اقتصادی کشور، لزوم و اهمیت در اختیار داشتن آمار و اطلاعات درست و بهنگام این بخش در برنامه‌ریزی‌های دقیق را مشخص می‌کند. اطلاعات و داده‌های آمارگیری از کارگاه‌های صنعتی دارای ۱۰ کارکن و بیشتر از غنی‌ترین منابعی است که هر ساله توسط مرکز آمار ایران جمع‌آوری می‌شود. از آنجا که یکی از مشکلات آمارگیری از کارگاه‌های صنعتی دارای ۱۰ کارکن و بیشتر، وجود داده‌های دورافتاده در دادگان این آمارگیری است، بنابراین شناسایی این داده‌ها و اصلاح آن‌ها با استفاده از روش‌های مناسب ضروری است. پژوهشکده‌ی آمار با توجه به رسالت خود در زمینه‌ی اجرای طرح‌های پژوهشی مربوط به ارتقاء کیفیت اطلاعات جمع‌آوری شده، اجرای طرح مذکور را در دستور کار خود قرار داد که اجرای آن از آبان ۱۳۹۰ آغاز شد و گزارش نهایی آن اکنون در دسترس علاقه‌مندان قرار گرفته است. این پژوهش در گروه پژوهشی طرح‌های فنی و روش‌های آماری پژوهشکده‌ی آمار با همکاری جناب آقای دکتر مجتبی گنجعلی عضو هیئت علمی دانشگاه شهید بهشتی به‌عنوان مجری طرح، خانم‌ها دکتر زهرا رضایی قهرودی، تابان باغفلکی، گلستان احسنی و مهناز کاظمی و آقای فرشاد روشن سنگاچین به‌عنوان همکار طرح به انجام رسیده است که بدین‌وسیله از یکایک این عزیزان تشکر و قدردانی می‌شود.

در طول انجام این طرح پژوهشی سرکار خانم نجمه ناظریان با دقت و حوصله‌ی بسیار زیاد زحمت حروف‌چینی و تایپ مستندات طرح را برعهده داشته‌اند که بدین‌وسیله از ایشان تشکر و قدردانی می‌شود.

گروه پژوهشی طرح‌های فنی و روش‌های آماری

پژوهشکده‌ی آمار

پیش‌گفتار مجری

یکی از مشکلات طرح آمارگیری از کارگاه‌های صنعتی وجود داده‌های پرت در پرسشنامه‌ها است. با توجه به این که جامعه‌ی کارگاه‌های صنعتی، جامعه‌ی نامتقارنی است وجود داده‌های پرت در اطلاعات، به خصوص اطلاعات کارگاه‌های بزرگ، می‌تواند موجب بروز اریبی در نتایج شود. راستی‌آزمایی‌های اطلاعات پرسشنامه‌ها با سایر اطلاعات و منابع از جمله صورت‌های مالی و اطلاعات قبلی نقش بسیار مهمی در یافتن خطاهای اندازه‌گیری دارد. ارتقای کیفیت اطلاعات جمع‌آوری شده در آمارگیری‌ها و سرشماری‌ها از هدف‌های اصلی نظام آماری است، از این‌رو بالا بودن دقت اطلاعات، برنامه‌ریزان و سیاستگذاران را در شناخت بخش‌های مختلف اقتصادی کمك می‌کند تا با آگاهی دقیق‌تری نسبت به برنامه‌ریزی اقدام کنند. از آنجا که یکی از مشکلات آمارگیری از کارگاه‌های صنعتی دارای ۱۰ کارکن و بیشتر، وجود داده‌های دورافتاده در دادگان این آمارگیری است، بنابراین شناسایی این داده‌ها و اصلاح آن‌ها با استفاده از روش‌های مناسب ضروری است.

در این طرح پژوهشی پس از بررسی سوابق و تجربه‌ی سایر کشورها در زمینه‌ی روش‌های تشخیص داده‌های دورافتاده و همچنین بررسی روش‌های شناسایی داده‌های دورافتاده آمارگیری کارگاه‌های صنعتی مرکز آمار ایران، به انتخاب روش مناسب تشخیص داده‌های دورافتاده این آمارگیری پرداخته شده است و نحوه‌ی اصلاح آن داده‌ها نیز ارائه شده است.

در این‌جا لازم است از همکاری صمیمانه‌ی جناب آقای زاهدیان رئیس محترم پژوهشکده‌ی آمار و پیشنهادات ارزشمند ایشان در ارائه‌ی روش اصلاح داده‌های پرت شناسایی شده در طرح آمارگیری از کارگاه‌های صنعتی دارای ۱۰ کارکن و بیشتر تشکر و قدردانی نمایم.

از خوانندگان محترم تقاضا می‌شود، نظرات اصلاحی خود در ارتباط با محتوای مجموعه‌ی حاضر را به گروه پژوهشی طرح‌های فنی و روش‌های آماری پژوهشکده‌ی آمار منعکس نمایند.

مجتبی گنجعلی

فهرست

عنوان **شماره‌ی صفحه**

فصل ۱: کلیات و مطالعات تطبیقی ۱

۱-۱- مقدمه ۱

۱-۲- نقاط دورافتاده ۲

۱-۳- طبقه‌بندی روش‌های شناسایی داده‌های پرت ۸

۱-۳-۱- روش‌های تک متغیره ۹

۱-۳-۱-۱- روش‌های یک مرحله‌ای در مقابل روش‌های متوالی ۱۱

۱-۳-۱-۲- روش‌های استوار تک متغیره ۱۲

۱-۳-۱-۳- کنترل فرایند آماری ۱۲

۱-۳-۲- روش‌های چند متغیره ۱۴

۱-۳-۲-۱- اندازه‌های استوار چند متغیره ۱۸

۱-۳-۲-۲- دورافتاده‌های چندمتغیره یا مقادیر کرانگین ۲۱

۱-۳-۲-۳- تشخیص دورافتاده‌های سازوار ۲۲

۱-۳-۳- برگشت از حالت چند متغیره به حالت تک‌متغیره ۲۴

۱-۳- ۴- روش‌های داده‌کاوی برای شناسایی نقاط پرت ۲۵

فصل ۲: تاریخچه‌ی تولید آمار صنعت در مرکز آمار ایران ۲۷

۲-۱- تاریخچه‌ی تولید آمار صنعت ۲۷

۲-۲- روش شناسایی داده‌های پرت در مرکز آمار ایران ۲۹

فصل ۳: مطالعات تطبیقی در خصوص روش‌های شناسایی داده‌های پرت ۳۹

۳-۱- روش‌های شناسایی داده‌های پرت در برخی کشورها و روش‌های اصلاح آن ۳۹

۳-۱-۱ کانادا ۳۹

۳-۱-۱-۱- آمارگیری سالانه‌ی صنعت و برش و تولید چوب ۳۹

۳-۱-۱-۲- آمارگیری ماهانه‌ی صنعت ۴۴

۳-۱-۲- نیوزلند ۴۶

۳-۱-۳- نپال ۴۹

۳-۱-۳-۱-سرشماری و آمارگیری کارگاه‌های صنعتی کشور نپال ۴۹

۳-۱-۴- کشورهای متحده‌ی پادشاهی ۵۱

۳-۱-۴-۱- آمارگیری سالانه‌ی کسب و کار ۵۱

۳-۱-۴-۲- آمارگیری ماهانه‌ی کارگاه‌های کسب و کار ۵۳

۳-۱-۵- آمارگیری سالانه‌ی تولیدات اتحادیه‌ی اروپا ۵۸

۳-۱-۶- ترکیه ۶۰

۳-۱-۶-۱- آمارگیری سالانه‌ی آمارهای بخش خدمات و صنعت ۶۰

۳-۱-۶-۲- آمارگیری فصلی تولیدات رشته فعالیت‌های صنعتی ۶۱

۳-۱-۶-۳- سرشماری عمومی کارگاه‌های صنعتی و کسب و کار ۶۱

۳-۱-۶-۴- آمارهای سالانه‌ی تولیدات صنعتی ۶۲

۳-۱-۷- اداره‌ی آمار اتحادیه‌ی اروپا ۶۳
۳-۱-۷-۱- آمارگیری PRODCOM اتحادیه‌ی اروپا در زمینه‌ی تولید کالاهای صنعتی ۶۳
۳-۱-۸- آمریکا ۶۶
۳-۱-۸-۱- آمارگیری سالانه‌ی صنعت ۶۶
۳-۱-۸-۲- آمارگیری فروش، موجودی انبار و سفارشات کارگاه‌های صنعتی ۶۷
۳-۲- درمان داده‌های پرت در آمارگیری‌های کسب و کار ۶۸
۳-۲-۱- دلایل رخداد داده‌ی پرت ۶۸
۳-۲-۲- روش‌های شناسایی داده‌های پرت ۶۹
۳-۲-۳- روش‌های درمان داده‌های پرت ۷۳
۳-۲-۳-۱- روش اصلاح وزن ۷۴
۳-۲-۳-۲- روش اصلاح مقدار ۷۴
۳-۲-۳-۳- روش اصلاح وزن و مقدار ۷۵
۳-۲-۳-۴- روش کم کردن داده‌ی پرت ۷۶
فصل ۴: معرفی روش‌های تشخیص مقادیر دورافتاده ۷۷
۴-۱- مقدمه ۷۷
۴-۲- تعاریف ۷۸
۴-۲-۱- هم‌وردای آفین ۷۸
۴-۲-۲- M- براوردها ۷۸
۴-۳- روش‌های یک متغیره ۷۹

۴-۳-۱- روش وینزوریدن برای شناسایی و درمان مقادیر دورافتاده ۷۹
۴-۳-۲- شناسایی مقادیر دورافتاده از طریق جستجوی پیشرو ۸۴
۴-۴- روش‌های چند متغیره ۸۶
۴-۴-۱- الگوریتم BACON ۸۶
۴-۴-۲- براوردهای مینیمم حجم بیضی‌وار ۸۷
۴-۴-۳- براوردگر مینیمم دترمینان کواریانس ۸۸
۴-۴-۴- براوردگر متعامد شده ندسیکن و کتترینگ ۹۰
۴-۴-۵- براوردگرهای استیهل- دونوهو استوار ۹۱
۴-۴-۶- براوردگرهای تعمیم‌یافته استیهل- دونوهو ۹۴
فصل ۵: کاربرد روش‌های تشخیص داده‌های پرت در طرح آمارگیری از کارگاه‌های صنعتی ۹۷
۵-۱- شناسایی داده‌های پرت کارگاه‌های صنعتی صنایع مواد غذایی و آشامیدنی ۹۷
۵-۲- بررسی داده‌های کد فعالیت ۲۷۱۰: تولید محصولات اولیه آهن و فولاد ۱۳۴
۵-۲-۱- معرفی داده‌ها و تحلیل توصیفی یا کاوشی آن‌ها ۱۳۴
۵-۲-۲- روش‌های تحلیلی فاصله‌های استوار ۱۴۲
۵-۲-۲-۱- تشخیص داده‌های دورافتاده با استفاده از روش جستجوی پیشروی هادی و سایمونوف ۱۴۲
۵-۲-۲-۲- روش مینیمم دترمینان کواریانس ۱۴۵
۵-۲-۲-۳- روش استیهل- دونوهو ۱۴۷
۵-۲-۲-۴- روش BACON ۱۵۰
۵-۲-۲-۵- روش مینیمم حجم بیضی‌وار ۱۵۱

۵-۲-۲-۶- روش متعامدسازی ندسیکن و کتترینگ.....۱۵۳

۵-۲-۲-۷- روش رسم مانده در مقابل فاصله (هادی و BACON) ۱۵۵

۵-۲-۲-۸- روش نمودار مانده‌های استاندارد در مقابل فاصله‌های ماهالانوبیس استوار.....۱۵۶

۵-۳- شناسایی داده‌های پرت کارگاه‌های صنعتی بر اساس اطلاعات سری زمانی ۱۵۹

۵-۴- روش اصلاح داده‌های پرت.....۱۶۱

فصل ۶: نتیجه‌گیری.....۱۶۵

پیوست‌ها.....۱۶۷

پیوست آ- کد فعالیت‌های چهار رقمی دارای صورت مالی در بورس.....۱۶۷

پیوست ب- داده‌های پرت شناسایی شده به روش مینیمم دترمینان کواریانس و BACON برای تمام کد فعالیت‌های چهار رقمی کارگاه‌های صنعتی در سال ۱۳۸۹ .....۱۶۹

پیوست پ: واحدهای صنعتی شناسایی شده به‌عنوان داده‌ی پرت بر اساس متغیر ستانده و داده به روش هیدروگلو و برتلت ۲۱۱

پیوست ت- برنامه‌نویسی با نرم‌افزار R.....۲۶۱

مرجع‌ها.....۲۹۳

فهرست جداول

عنوان **شماره‌ی صفحه**

جدول ۳-۱: شناسایی داده‌های پرت..... ۷۰

جدول ۳-۲:: روش‌های درمان داده‌های پرت..... ۷۳

جدول ۳-۳: برخی روش‌های جانهی برای داده‌های دورافتاده..... ۷۵

جدول ۵-۱: مقادیر متغیرها برای ٤ کارگاه با ارزش آب و برق مصرفی صفر (میلیون ریال) ۱۳۹

جدول ۵-۲: مقادیر چهار متغیر مورد بررسی و اطلاعات ارزش داده و تعداد کارکن برای داده‌های دورافتاده شناسایی شده در روش پیشرو برای کد فعالیت ۲۷۱۰ ..... ۱۴۴

جدول ۵-۳: براورد استوار پارامتر مکان با استفاده از روش مینیمم دترمینان کواریانس برای لگاریتم متغیرهای مورد مطالعه‌ی کد فعالیت ۲۷۱۰ ..... ۱۴۵

جدول ۵-۴: واحدهای صنعتی شناسایی شده به‌عنوان داده‌ی پرت با استفاده از روش مینیمم دترمینان کواریانس به همراه برخی متغیرهای اصلی مربوط به کارگاه..... ۱۴۷

جدول ۵-۵: براورد استوار پارامتر مکان با استفاده از روش استیهل- دونوهو برای لگاریتم متغیرهای مورد مطالعه‌ی کد فعالیت ۲۷۱۰ ..... ۱۴۸

جدول ۵-۶:واحدهای صنعتی شناسایی شده به‌عنوان داده‌ی پرت با استفاده از روش استهیل- دونوهو به همراه برخی متغیرهای اصلی مربوط به کارگاه..... ۱۴۹

جدول ۵-۷: براورد استوار پارامتر مکان با استفاده از روش BACON برای لگاریتم متغیرهای مورد مطالعه و برای کد فعالیت ۲۷۱۰ ..... ۱۵۰

جدول ۵-۸: واحدهای صنعتی شناسایی شده به‌عنوان داده‌ی پرت با استفاده از روش BACON به همراه برخی متغیرهای اصلی مربوط به کارگاه..... ۱۵۱

جدول ۵-۹: براورد استوار پارامتر مکان با استفاده از روش مینیمم حجم بیضی‌وار برای لگاریتم متغیرهای مورد مطالعه‌ی کد فعالیت ۲۷۱۰ ..... ۱۵۱

جدول ۵-۱۰: واحدهای صنعتی شناسایی شده به‌عنوان داده‌ی پرت با استفاده از روش مینیمم حجم بیضی‌وار به همراه برخی متغیرهای اصلی مربوط به کارگاه ۱۵۳

جدول ۵-۱۱: براورد استوار پارامتر مکان با استفاده از روش متعامدسازی ندسیکن و کتنرینگ برای لگاریتم متغیرهای مورد مطالعه‌ی کد فعالیت ۲۷۱۰ ..... ۱۵۳

جدول ۵-۱۲: واحدهای صنعتی شناسایی شده به‌عنوان داده‌ی پرت با استفاده از روش متعامدسازی ندسیکن و کتنرینگ به همراه برخی متغیرهای اصلی مربوط به کارگاه ۱۵۵

جدول ۵-۱۳: مقادیر متغیرهای مورد مطالعه در روش نمودار مانده‌های استاندارد در مقابل فاصله‌های ماهالانوبیس استوار برای کد فعالیت ۲۷۱۰ ..... ۱۵۷

جدول ۵-۱۴: مقایسه شماره‌ی داده‌های دورافتاده شناسایی شده توسط روش‌های مختلف برای متغیرهای مورد مطالعه‌ی کد فعالیت ۲۷۱۰ ..... ۱۵۸

جدول ۵-۱۵: تعداد کارگاه‌های جابه‌جا شده از طبقات مختلف تعداد کارکن طی سال‌های ۱۳۸۷ و ۱۳۸۸ ..... ۱۶۰

جدول ۵-۱۶: توزیع داده و ستانده قبل و بعد از تعدیل به تفکیک دهک ۱۶۴

فهرست شکل‌ها

عنوان **شماره‌ی صفحه**

شکل ۱-۱: یک فضای دو بعدی با یک مشاهده‌ی پرت..... ۱۴

شکل ۱-۲: داده‌های تولید شده از توزیع نرمال استاندارد با در نظرگرفتن یک همبستگی از پیش تعیین شده..... ۱۸

شکل ۱-۳: نمودار پراکنش log (Be) و log (Sr) ۲۰

شکل ۱-٤: نمودارهای رسم فاصله‌ی ماهالانوبیس در مقابل چندک توزیع خی‌دو ۲۲

شکل ۱-۵: نحوه شناسایی دورافتاده‌های سازوار..... ۲۴

شکل ۱-۶: نمودار متغیرهای مختلف تکی تحت عناوین As، Cd، Co، ...، Zn ۲۵

شکل ۲-۱: فرآیند بررسی داده‌های کارگاه‌های صنعتی دارای ۱۰ کارکن و بیشتر در دفتر صنعت، معدن و محیط‌زیست تا قبل از سال ۱۳۸۶ ..... ۳۰

شکل ۲-۲: فرآیند بررسی داده‌های کارگاه‌های صنعتی دارای ۱۰ کارکن و بیشتر در دفتر صنعت، معدن و زیربنایی از سال ۱۳۸۷ تا سال ۱۳۸۹ ..... ۳۱

شکل ۲-۳: فرآیند بررسی داده‌های کارگاه‌های صنعتی دارای ۱۰ کارکن و بیشتر در دفتر حساب‌های اقتصادی ۳۳

شکل ۴-۱: وینزوریدن نوع ۱ و نوع ۲..... ۸۱

شکل ۴-۲: نمودار پراکنش و تصویر در داده‌های ۲ متغیره به منظور تشخیص مقادیر دورافتاده..... ۹۴

شکل ۴-۳: نمودار پراکنش و تصویر در داده‌های ۳ متغیره به منظور تشخیص مقادیر دورافتاده..... ۹۴

شکل ۵-۱: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۱۲ ..... ۱۰۱

شکل ۵-۲: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۱۴ ..... ۱۰۲

شکل ۵-۳: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۱۵ ..... ۱۰۳

شکل ۵-۴: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۱۶ ..... ۱۰۴

شکل ۵-۵: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۱۷ ..... ۱۰۵

شکل ۵-۶: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۱۸ ..... ۱۰۶

شکل ۵-۷: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۱۹ ..... ۱۰۷

شکل ۵-۸: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۲۰ ..... ۱۰۸

شکل ۵-۹: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۳۱ ..... ۱۰۹

شکل ۵-۱۰: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۳۲ ..... ۱۱۰

شکل ۵-۱۱: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۳۳ ..... ۱۱۱

شکل ۵-۱۲: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۴۲ ..... ۱۱۲

شکل ۵-۱۳: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۴۳ ..... ۱۱۳

شکل ۵-۱۴: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۴۴ ..... ۱۱۴

شکل ۵-۱۵: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۴۵ ..... ۱۱۵

شکل ۵-۱۶: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۴۶ ..... ۱۱۶

شکل ۵-۱۷: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۴۷ ..... ۱۱۷

شکل ۵-۱۸: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۴۸ ..... ۱۱۸

شکل ۵-۱۹: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۵۱ ..... ۱۱۹

شکل ۵-۲۰: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۵۳ ..... ۱۲۰

شکل ۵-۲۱:نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۵۵ ..... ۱۲۱

شکل ۵-۲۲: نمودار پراکنش و جعبه‌ای متغیرهای نسبت مصرف واسطه به ستانده، مصرف واسطه، ستانده، ارزش تولید و ارزش‌افزوده‌ی کد فعالیت صنعتی ۱۵۵۶ ..... ۱۲۲

شکل ۵-۲۳: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۱۲ ..... ۱۲۳

شکل ۵-۲۴: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۱۴ ..... ۱۲۳

شکل ۵-۲۵: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۱۵ ..... ۱۲۴

شکل ۵-۲۶: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۱۶ ..... ۱۲۴

شکل ۵-۲۷: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۱۷ ..... ۱۲۵

شکل ۵-۲۸: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۱۸ ..... ۱۲۵

شکل ۵-۲۹: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۱۹ ..... ۱۲۶

شکل ۵-۳۰: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۲۰ ..... ۱۲۶

شکل ۵-۳۱: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۳۱ ..... ۱۲۷

شکل ۵-۳۲: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۳۲ ..... ۱۲۷

شکل ۵-۳۳: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۳۳ ..... ۱۲۸

شکل ۵-۳۴: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۴۲ ..... ۱۲۸

شکل ۵-۳۵: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۴۳ ..... ۱۲۹

شکل ۵-۳۶: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۴۴ ..... ۱۲۹

شکل ۵-۳۷: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۴۵ ..... ۱۳۰

شکل ۵-۳۸: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۴۶ ..... ۱۳۰

شکل ۵-۳۹: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۴۷ ..... ۱۳۱

شکل ۵-۴۰: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۴۸ ..... ۱۳۱

شکل ۵-۴۱: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۵۱ ..... ۱۳۲

شکل ۵-۴۲: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۵۳ ..... ۱۳۲

شکل ۵-۴۳: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۵۵ ..... ۱۳۳

شکل ۵-۴۴: نمودار پراکنش تغییرات مقایسه‌ی ۵ متغیر مورد مطالعه طی دو سال ۸۷ و ۸۸، ۸۸ و ۸۹ و اختلاف سال‌های ۸۷ و ۸۸ در مقابل اختلاف سال‌های ۸۸ و ۸۹ برای کد فعالیت ۱۵۵۶ ..... ۱۳۳

شکل ۵-۴۵: بافت‌نگار متغیرهای مورد بررسی بدون در نظر گرفتن وزن کارگاه‌ها ۱۳۴

شکل ۵-۴۶: بافت‌نگار لگاریتم متغیرهای مورد بررسی (به استثنای I/O) بدون در نظر گرفتن وزن کارگاه‌ها ۱۳۵

شکل ۵-۴۷: بافت‌نگار متغیرهای مورد بررسی با در نظر گرفتن وزن کارگاه‌ها ۱۳۶

شکل ۵-۴۸: بافت‌نگار لگاریتم متغیرهای مورد بررسی (به استثنای I/O) با درنظر گرفتن وزن کارگاه‌ها ۱۳۶

شکل ۵-۴۹: نمودار جعبه‌ای متغیرهای مورد بررسی بدون در نظر گرفتن وزن کارگاه‌ها ۱۳۷

شکل ۵-۵۰: نمودار جعبه‌ای لگاریتم متغیرهای مورد بررسی (به استثنای متغیر I/O) بدون در نظر گرفتن وزن کارگاه‌ها ۱۳۷

شکل ۵-۵۱: نمودار جعبه‌ای وزنی متغیرهای مورد بررسی ۱۳۸

شکل ۵-۵۲: نمودار جعبه‌ای وزنی لگاریتم متغیرهای مورد بررسی (به استثنای I/O) ۱۳۸

شکل ۵-۵۳: نمودارهای بافت نگار متغیر ارزش آب، برق و سوخت مصرفی و لگاریتم آن با و بدون در نظر گرفتن وزن‌ها ۱۳۹

شکل ۵-۵۴: نمودار جعبه‌ای ارزش آب، برق و سوخت مصرفی و لگاریتم آن با و بدون در نظر گرفتن وزن‌ها ۱۴۰

شکل ۵-۵۵: نمودار همبستگی بین لگاریتم متغیرها و متغیر داده/ستانده (I/O) ۱۴۰

شکل ۵-۵۶: نمودار پراکنش متغیرهای مورد مطالعه با هیستوگرام‌های حاشیه‌ای آن‌ها بر روی قطر اصلی برای کد فعالیت ۲۷۱۰ ..... ۱۴۱

شکل ۵-۵۷: نمودار پراکنش لگاریتم متغیرهای مورد مطالعه و متغیر I/O با هیستوگرام‌های حاشیه‌ای آن‌ها بر روی قطر اصلی برای کد فعالیت ۲۷۱۰ ..... ۱۴۲

شکل ۵-۵۸: نمودارهای بافت‌نگار و جعبه‌ای متغیر ارزش مواد اولیه با و بدون در نظر گرفتن وزن‌ها ۱۴۳

شکل ۵-۵۹: نمودارهای بافت‌نگار و جعبه‌ای متغیر لگاریتم متغیر ارزش مواد اولیه با و بدون در نظر گرفتن وزن‌ها ۱۴۳

شکل ۵-۶۰: آماره di در روش پیشرو برای کد فعالیت ۲۷۱۰ ..... ۱۴۵

شکل ۵-۶۱: نمودار پراکنش متغیرها به همراه بیضی‌وار تحمل کلاسیک و استوار به روش مینیمم دترمینان کواریانس و هیستوگرام متغیرها و منحنی توزیع آن بر روی قطر اصلی برای کد فعالیت ۲۷۱۰ ..... ۱۴۶

شکل ۵-۶۲: نمودار فاصله- فاصله با استفاده از روش مینیمم دترمینان کواریانس برای کد فعالیت ۲۷۱۰ ..... ۱۴۶

شکل ۵-۶۳: نمودار پراکنش متغیرها به همراه بیضی‌وار تحمل کلاسیک و استوار به روش استیهل-دونوهو و هیستوگرام متغیرها و منحنی توزیع آن بر روی قطر اصلی برای کد فعالیت ۲۷۱۰ ..... ۱۴۸

شکل ۵-۶۴: نمودار فاصله- فاصله با استفاده از روش استیهل- دونوهو برای کد فعالیت ۲۷۱۰ ..... ۱۴۹

شکل ۵-۶۵: مقادیر دورافتاده شناسایی شده به روش BACON برای کد فعالیت ۲۷۱۰ ..... ۱۵۰

شکل ۵-۶۶: نمودار فاصله- فاصله با استفاده از روش مینیمم حجم بیضی‌وار برای کد فعالیت ۲۷۱۰ ..... ۱۵۲

شکل ۵-۶۷: نمودار پراکنش متغیرها به همراه بیضی‌وار تحمل کلاسیک و استوار به روش مینیمم حجم بیضی‌وار و هیستوگرام متغیرها و منحنی توزیع آن بر روی قطر اصلی برای کد فعالیت ۲۷۱۰ ..... ۱۵۲

شکل ۵-۶۸: نمودار فاصله- فاصله با استفاده از روش متعامدسازی ندسیکن و کتترینگ برای کد فعالیت ۲۷۱۰ .....۱۵۴

شکل ۵-۶۹: نمودار پراکنش متغیرها به همراه بیضی‌وار تحمل کلاسیک و استوار به متعامدسازی ندسیکن و کتترینگ و هیستوگرام متغیرها و منحنی توزیع آن بر روی قطر اصلی برای کد فعالیت ۲۷۱۰ .....۱۵۴

شکل ۵-۷۰: نمودار فاصله‌های به دست آمده از روش BACON در مقابل مانده‌های حاصل از روش پیشرو هادی۱۵۵

شکل ۵-۷۱: نمودار مانده‌های استاندارد در مقابل فاصله‌های ماهالانوبیس استوار برای کد فعالیت ۲۷۱۰ .....۱۵۶

شکل ۵-۷۲: نمودار نرخ داده به ستانده اظهار شده به تفکیک دهک‌ها ۱۶۳

شکل ۵-۷۳: نمودار نرخ داده به ستانده بعد از تعدیل به تفکیک دهک‌ها ۱۶۴

فصل ۱

کلیات و مطالعات تطبیقی

۱-۱-مقدمه

نقش و اهمیت بخش صنعت در جریان توسعه‌ی اقتصادی کشور، لزوم اهمیت در اختیار داشتن آمار و اطلاعات دقیق و بهنگام این بخش در برنامه‌ریزی‌های دقیق را مشخص می‌کند. اطلاعات آمارگیری از کارگاه‌های صنعتی ۱۰ نفر کارکن و بیشتر از غنی‌ترین منابعی است که هر ساله توسط مرکز آمار ایران جمع‌آوری می‌شود.

هدف از اجرای این طرح، جمع‌آوری اطلاعات مورد نیاز به‌منظور تهیه‌ی زمینه‌ی اطلاعاتی مناسب از ویژگی‌های کارگاه‌های صنعتی دارای ده نفر کارکن و بیشتر، جهت برنامه‌ریزی‌های توسعه صنعتی کشور، اتخاذ سیاست‌های اقتصادی و ارزیابی نتایج حاصل از اجرای برنامه‌های توسعه صنعتی کشور، اعمال سیاست‌های اقتصادی و انجام تحقیقات صنعتی است.

نتایج این طرح، کاربردهای فراوانی در محاسبات و براورد پارامترهای مختلف اقتصادی در بخش صنعت دارد که از جمله می‌توان محاسبه‌ی ارزش داده‌ها و ستانده‌ها و اجزاء آن در فعالیت‌های صنعتی، محاسبه‌ی ارزش‌افزوده‌ی فعالیت‌های صنعتی و نیز ارزش‌افزوده‌ی کارگاه‌های صنعتی، محاسبه‌ی شاخص‌ها و نماگرهای صنعتی از جمله ارزش‌افزوده‌ی سرانه، ستانده‌ی سرانه، مزد و حقوق سرانه، درصد ارزش مواد اولیه خارجی به مواد اولیه مصرفی، بهره‌وری کار، بهره‌وری انرژی را نام برد. همچنین نتایج تفصیلی این طرح اهداف تهیه جدول داده- ستانده اقتصاد کشور در بخش صنعت را نیز تأمین می‌کند.

ارتقاء کیفیت اطلاعات جمع‌آوری شده در آمارگیری‌ها و سرشماری‌ها از هدف‌های اصلی نظام آماری است. بالا بودن دقت اطلاعات، برنامه‌ریزان و سیاست‌گذاران را در شناخت بخش‌های مختلف اقتصادی کمک می‌کند تا با آگاهی دقیق‌تری نسبت به برنامه‌ریزی اقدام کنند. از آنجا که یکی از مشکلات طرح آمارگیری از کارگاه‌های صنعتی ۱۰ نفر کارکن و بیشتر، وجود داده‌های پرت (دورافتاده) در دادگان این طرح است، بنابراین شناسایی این داده‌ها و اصلاح آن‌ها با استفاده از روش‌های مناسب ضروری است. در این طرح

پژوهشی پس از بررسی سوابق و تجربه‌ی سایر کشورها در زمینه‌ی روش‌های تشخیص داده‌های پرت نقاط دورافتاده و همچنین بررسی روش‌های شناسایی داده‌های پرت این طرح در مرکز آمار ایران، به انتخاب روش مناسب تشخیص داده‌های پرت در طرح آمارگیری از کارگاه‌های صنعتی پرداخته خواهد شد.

۱-۲- نقاط دورافتاده

نقاط دورافتاده در داده‌های واقعی بسیار رخ می‌دهند و اغلب از کنار این نقاط با بی‌توجهی گذر می‌شود. داده‌های پرت به دلایل متعددی از جمله خطای انسانی، خطای ابزار، تغییرات طبیعی در جامعه و تغییرات در رفتار سیستم یا خطا در سیستم به وجود می‌آید. نحوه‌ی برخورد سیستم تشخیص داده‌های پرت بستگی به حوزه‌ی کاربردی آن دارد. اگر داده‌ی پرت معرف یک خطای تایپی توسط متصدی ورود اطلاعات باشد، در این صورت متصدی ورود اطلاعات می‌تواند به‌سادگی خطا را رفع کند و داده‌ی پرت به یک رکورد عادی برگردانده شود. داده‌ی پرتی که به دلیل خطای ابزار رخ داده باشد، می‌تواند به‌سادگی حذف شود. به‌عنوان مثال، یک بررسی از ویژگی‌های جمعیت انسانی ممکن است شامل ناهنجاری‌هایی از جمله وجود تعداد انگشت‌شماری از افراد بلند قد باشد. در این حالت، ناهنجاری به وجود آمده کاملاً طبیعی است اگرچه برای اطمینان از عدم وجود خطا باید داده‌ها را طبقه‌بندی کرد. در واقع لازم است از یک الگوریتم طبقه‌بندی استفاده شود که نسبت به داده‌های پرت استوار است تا داده‌هایی را که به‌طور طبیعی به‌عنوان نقاط پرت شناسایی کرده است، مدل‌بندی کند. بنابراین وجود یک داده‌ی پرت در هر سیستمی باید بلافاصله شناسایی شود و هشدار مناسبی به مدیر سیستم برای وجود چنین مشکلی ارائه شود.

در بسیاری از فعالیت‌های مربوط به تحلیل داده، اطلاعات از طریق نمونه‌گیری یا ثبت به دست می‌آید. یکی از اولین گام‌ها برای دستیابی به یک تحلیل منسجم، شناسایی مشاهدات پرت است. اگرچه نقاط پرت یا دورافتاده اغلب به‌عنوان خطا یا نوفه در نظر گرفته می‌شود، اما ممکن است دارای اطلاعات مهمی باشند. از آن‌جا که داده‌های پرت ممکن است منجر به بدمشخص‌سازی مدل، براوردهای اریب پارامترها و نتایج نادرستی شود، بنابراین شناسایی آن‌ها قبل از مدل‌بندی و تحلیل، ضروری است.

تعریف دقیق از نقاط پرت اغلب به فرضیاتِ پنهانِ مربوط به ساختار داده‌ها و روش‌های تشخیص و شناسایی به‌کار رفته وابسته است. با این حال برخی از تعاریف به اندازه‌ی کافی عمومی است تا بتوان آن‌ها را برای انواع مختلف داده‌ها و روش‌ها استفاده کرد. هاوکینز (۱۹۸۰) نقطه‌ی پرت را به‌عنوان مشاهده‌ای در نظر می‌گیرد که از بقیه‌ی مشاهدات به‌قدری دور است که گویی این داده از طریق مکانیسم دیگری ایجاد شده است. بارنت و لویز (۱۹۹۴) یک داده‌ی پرت را داده‌ای می‌دانند که به‌طور مشخص از بقیه‌ی داده‌های نمونه دور است و جانسن (۱۹۹۲) یک داده‌ی پرت را مشاهده‌ای می‌داند که با بقیه‌ی داده‌ها ناسازگار است.

نکته قابل توجه این است که نقاط پرت نه تنها در متغیر پاسخ می‌تواند رخ دهد بلکه در متغیرهای تبیینی مدل نیز می‌توانند عامل دورافتادگی رخ دهد که به آن‌ها نقاط نافذ[1] گفته می‌شود. هر دوی این‌گونه نقاطِ دورافتاده و نافذ می‌تواند موجب عدم کارایی یک تحلیل کم‌ترین توان‌های دوم معمولی گردند. داده‌های پرت ممکن است به‌صورت تصادفی در هر داده‌ای رخ دهند اما این داده‌ها اغلب یا نشان‌دهنده‌ی خطای اندازه‌گیری هستند یا نشان‌دهنده‌ی آن هستند که جامعه دارای توزیع دم‌سنگین است. در حالت اول اغلب به دنبال حذف این داده‌ها یا استفاده از آماره‌هایی هستیم که نسبت به این نوع داده‌های پرت استوار باشند. در حالت دوم، توزیع داده‌ها نشان‌دهنده‌ی چولگی شدید است.

خطای مشاهداتی یا خطای اندازه‌گیری[2] اختلاف بین مقدار اندازه‌گیری‌شده‌ی کمیت مورد بررسی و مقدار واقعی و نامعلوم کمیت مورد بررسی است که معمولاً از چهار منبع اولیه ناشی می‌شود.

۱- وسیله‌ی گردآوری اطلاعات و کیفیت آن‌ها (طراحی پرسشنامه و ادبیات آن)

۲- روش گردآوری داده‌ها (مصاحبه‌ی رودررو، مصاحبه‌ی تلفنی، پستی و ...)

۳- آمارگیر یا پرسشگر

۴- پاسخگو

خطای مشاهداتی اغلب اشاره به خطاهای پاسخ و برخی انواع دیگر خطاهای غیر نمونه‌گیری دارد. در آمارگیری‌ها، این خطاها می‌تواند اشتباهات در جمع‌آوری داده‌ها، شامل ثبت‌های نادرست یک پاسخ و یا ثبت‌های درست پاسخ نادرست پاسخگوها باشد.

چمبرز (۱۹۸۶) سه نوع مجزا از مقادیر دورافتاده ارائه می‌دهد: نقاط دورافتاده‌ی نمایانگر[3] که مقادیر واقعی هستند که کاملاً منحصر به فرد در جامعه در نظر گرفته نمی‌شوند، نقاط دورافتاده‌ی غیرنمایانگر[4] که کاملاً منحصر به فرد در جامعه در نظر گرفته می‌شوند (داده‌ی واقعی که خیلی با دیگر مقادیر تفاوت داشته باشد) و خطاهای اندازه‌گیری ناخالص[5] که مشاهدات دورافتاده‌ای هستند که مقادیر واقعی نیستند. برای مثال در مطالعه‌ای از یک صنعت خاص که در آن بیشتر شرکت‌ها کمتر از ۱ درصد از سهم بازار را دارد، یکی از پنج شرکت با ۵ -۱۰ درصد سهم بازار ممکن است به‌عنوان یک نقطه‌ی دورافتاده نمایانگر و یک شرکت با ۵۰ درصد سهم بازار یک دورافتاده غیرنمایانگر تلقی شود. همچنین اگر تعداد اشتغال گزارش شده در یک شرکت به جای ۲۰۰۰ به اشتباه ۲۰۰۰۰ گزارش شود، به‌عنوان یک خطای اندازه‌گیری ناخالص در نظر گرفته می‌شود. لازم به ذکر است که به‌دلیل اینکه تصویر واقعی‌ای از جامعه مشخص نیست، تفکیک اینکه یک مشاهده دورافتاده نمایانگر است یا غیر نمایانگر، کار ساده‌ای نیست.

تحقیقات درباره نقاط دورافتاده عمدتاً بر نقاط دورافتاده‌ی نمایانگر و غیرنمایانگر تمرکز دارند و خطاهای اندازه‌گیری ناخالص را در برنمی‌گیرند. دلیل اصلی این کار این است که خطاهای اندازه‌گیری ناخالص

---

[1] Influential Observations
[2] Measurement Error
[3] Representative outliers
[4] Non-representative outliers
[5] Gross measurement errors

باید در مرحله ویرایش داده‌ها[1]، شناسایی و تصحیح شوند و بنابراین هر مقدار بزرگی که پس از ویرایش داده‌ها مانده باشد به‌عنوان مشاهدات واقعی تلقی شوند. بنابراین در ادامه نحوه‌ی برخورد با داده‌های پرتی بیان می‌شود که به‌عنوان مقادیر بزرگ واقعی و صحیح وجود دارند. برای بحث بیشتر در مورد خطاهای اندازه‌گیری ناخالص و روش‌های ویرایش این مشاهدات به اندرسن و همکاران (۲۰۰۳)، چمبرز و رن (۲۰۰۴)، لتوچ و برتلت (۱۹۹۲)، لورنس و مکدیوید (۱۹۹۴) و لیتل و اسمیت (۱۹۸۷) مراجعه شود.

مقادیر دورافتاده از دو جهت مورد توجه هستند. اول اینکه با توجه به حضور داده‌های دورافتاده‌ی نمایانگر و غیرنمایانگر، برخی تحلیل‌گرها علاقه‌ی زیادی به شناسایی واحدهایی دارند که منحصر به فرد هستند و این نوع داده‌ها مانع استفاده از برخی روش‌های تحلیل آماری می‌شوند. دوم اینکه شمول یا حذف این واحدها منجر به تغییرات اساسی در مقادیر براورد شده می‌گردد. از آنجا که حضور یا عدم حضور اینگونه داده‌های دورافتاده به خصوص در تحلیل داده‌های مربوط به آمارگیری‌های کارگاهی (به دلایلی از قبیل وجود داده‌های دورافتاده‌ای که دارای مقدار بزرگی در بین پاسخگوها باشد، وجود داده‌های دورافتاده که دارای وزن بزرگی در بین پاسخگوها باشد و یا حذف و خارج شدن داده‌های دورافتاده از بین پاسخگوها، به دلیل انتخاب نمونه یا بی‌پاسخی) می‌تواند منجر به تغییرات اساسی در مقادیر براوردشده و ایجاد چالش‌های اساسی در این نوع آمارگیری‌ها شود، لذا در روش‌های وزنی، این داده‌ها وزن زیادی را به خود نسبت می‌دهند. بنابراین توجه به مبحث نقاط دورافتاده در آمارگیری‌ها، غالبا در چارچوب توابع تاثیر موزون توسعه پیدا کرده است. برای تعاریف ریاضی پایه‌ای و بحث در توابع تاثیر در آمارگیری‌ها به اسمیت (۱۹۸۷) و مراجع اشاره شده توسط وی مراجعه شود.

بیکر (۲۰۰۰) اشاره می‌کند که شناسایی مقادیر دورافتاده گاهی اوقات به معنی حذف مقادیر دورافتاده به منظور اجتناب از اختلال یا تحلیل‌های بیشتر است. با این‌وجود، مقادیر دورافتاده ممکن است در جای خود مشاهدات جالب توجهی باشند، زیرا آن‌ها دیدگاه‌هایی را در ساختار داده‌ها یا پیشامدهای خاص در حین نمونه‌گیری نشان می‌دهند.

یکی از مشکلات طرح آمارگیری از کارگاه‌های صنعتی وجود داده‌های پرت در پرسشنامه‌ها است. با توجه به این که جامعه‌ی کارگاه‌های صنعتی، جامعه‌ی نامتقارنی است، وجود داده‌های پرت در اطلاعات، به خصوص کارگاه‌های بزرگ، می‌تواند موجب بروز اریبی در نتایج شود. راستی‌آزمایی‌های اطلاعات پرسشنامه‌ها با سایر اطلاعات و منابع، از جمله صورت‌های مالی، بورس و اطلاعات قبلی نقش بسیار مهمی در یافتن خطاهای اندازه‌گیری و شناسایی داده‌های پرت نمایانگر و غیرنمایانگر دارد. بنابراین از آنجا که برای تعدادی از رشته فعالیت‌های صنعتی، اطلاعات صورت‌های مالی آن‌ها وجود دارد و یا در بورس ثبت شده‌اند، پیشنهاد می‌شود که این کارگاه‌ها از طرح آمارگیری از کارگاه‌های صنعتی ۱۰ کارکن و بیشتر حذف شوند و یا اطلاعات این کارگاه‌ها با صورت‌های مالی و اطلاعات موجود در بورس مقایسه شوند.

[1] Data-editing stage

در داده‌های طرح آمارگیری از کارگاه‌های صنعتی ۱۰ کارکن و بیشتر، با توجه به نوع فعالیت کارگاه، کارگاه‌هایی وجود دارند که ارزش‌افزوده، ارزش داده، ستانده و بقیه‌ی مؤلفه‌های اصلی در صنعت آن‌ها بسیار بزرگ است و همچنین در این طرح ممکن است بسیاری از کارگاه‌ها ارقام واقعی ارزش داده، ستانده و ... را ارائه ندهند، بنابراین به نظر می‌رسد که هر دو نوع داده‌ی پرت (نمایانگر و غیرنمایانگر) در این طرح وجود دارد که باید به روش‌های آماری صحیح به شناسایی و اصلاح آن‌ها پرداخته شود.

روش‌های مختلفی برای شناسایی داده‌های پرت وجود دارد. یکی از روش‌های شناسایی داده‌های پرت در طرح آمارگیری از کارگاه‌های صنعتی، روشی است که توسط بانک مرکزی استفاده می‌شود. در این روش برای بررسی و محاسبه‌ی ارزش افزوده از نسبت داده به ستانده که اطلاع مهمی است که برای هر کارگاه وجود دارد و معمولاً این نسبت در کارگاه‌های بسیار بزرگ در طول زمان تغییر نمی‌کند، استفاده می‌کنند. با توجه به اینکه معمولاً ارزش تولید کارگاه در سال مورد بررسی وجود دارد، حاصل ضرب این عدد در نسبت داده به ستانده، میزان ارزش داده‌ی کارگاه در سال مورد بررسی را نشان می‌دهد. تفاضل ارزش تولید کارگاه از ارزش داده‌ی کارگاه محاسبه شده به‌عنوان معیاری از ارزش افزوده‌ی کارگاه در سال مورد بررسی می‌باشد. با استفاده از این روش می‌توان داده‌های پرت را شناسایی کرد و خطاهای اندازه‌گیری مقدار ارزش افزوده را تا حدودی به خصوص برای آن دسته از کارگاه‌هایی که اطلاع آن‌ها در منابع ثبتی وجود دارد، تعدیل کرد.

بررسی نمودار پراکنش، متداولترین روش تشخیص مقادیر دورافتاده است (مینیوم و همکاران، ۱۹۹۹؛ پایل، ۱۹۹۹). یک روش ساده‌ی دیگر استفاده از نمودار جعبه‌ای است، به‌طوری‌که مشاهدات خارج ۱/۵± برابر دامنه میان‌چارکی را دورافتاده خفیف[1] و مشاهدات خارج ۳± برابر دامنه میان‌چارکی را دورافتاده کرانگین[2] نامیده‌اند. لازم به ذکر است که این روش تنها برای متغیرهای پیوسته با توزیع احتمال یک‌متغیره کاربرد دارد. همچنین، هیدایرگلو و بردیلت (۱۹۸۶) روشی را بر مبنای تابعی از چارک مشاهدات برای حذف داده‌های دورافتاده پیشنهاد کرده‌اند.

دو روش برای تشخیص مقادیر دورافتاده روش‌های تک‌متغیره و چندمتغیره است که در روش تک‌متغیره، هر متغیر به‌طور فردی آزمون می‌شود ولی در روش‌های چندمتغیره وابستگی بین متغیرها در داده‌های یکسان مورد بررسی قرار می‌گیرد. بر طبق روش پایل (۱۹۹۹) یک مشاهده دورافتاده است، اگر از مقادیر دیگر دور باشد.

مسئله‌ی نقاط دورافتاده همواره در برازش یک مدل مورد توجه قرار می‌گیرند و معرفی آن‌ها منوط به داشتن معیاری برای میزان دوری از اکثریت داده‌هاست که بر طبق مانده‌های مرتبط با مدل استفاده شده تعریف می‌گردند. یکی از آشناترین این مدل‌ها، مدل رگرسیونی ساده می‌باشد که شامل برازش مدل‌هایی به یک پاسخ پیوسته است که امید داریم به مقادیر چندین متغیر کمکی وابسته باشد. همچنین در رگرسیون، فرض می‌کنیم که متغیر پاسخ به همراه خطا مشاهده گردیده در حالی که متغیرهای کمکی مقادیر معلومی بدون خطا

---

[1] Mild outliers
[2] Extreme outliers

هستند (که البته در عمل فرض واقع بینانه‌ای نمی‌باشد) و رابطه‌ی میان این دو مجموعه از متغیرها با استفاده از مجموعه‌ای از پارامترهای نامعلوم که توسط روش کم‌ترین توان‌های دوم برآورد می‌شوند، وجود دارد.

در رگرسیون خطی، متغیر پاسخ به‌صورت خطی به بردار متغیرهای کمکی توسط پارامترهای نامعلوم وابسته است. دو کلاس کلی‌تر از مدل‌ها شامل مدل‌هایی است که غیر خطی در پارامتر هستند و مدل‌های تعمیم‌یافته که در آن‌ها، احتیاجی به پیوسته بودن پاسخ نبوده و وابستگی میان میانگین متغیر پاسخ و متغیرهای مستقل توسط یک تابع ربط برقرار می‌شود. توجه ما در این‌جا بیشتر به مدل‌هایی است که به‌صورت رگرسیون چندگانه باشند یعنی پاسخ به چندین متغیر کمکی وابسته باشد.

برونلی (۱۹۶۰)، رگرسیون چندگانه را توصیف کرده و به برآورد پارامترها و آزمون فرض پارامترهای مدل پرداخته است، اما از روش‌های نموداری هم‌چون رسم داده‌ها، یا مانده‌های مدل برازش شده برای تشخیص نقاط دورافتاده استفاده نکرده است. از زمان انتشار کتاب برونلی، تفاوت‌های تحسین برانگیزی در تحلیل‌های آماری رخ داده که با وجود نرم‌افزارهای قوی گرافیکی آماری به راحتی می‌توان خصوصیات مدل‌های مختلف را با رسم نمودارهای گوناگون بررسی کرد. یکی از کاربردهای این نمودارها که به نمودارهای تشخیصی[1] معروف هستند، شامل تشخیص پرت بودن یک یا چند مشاهده می‌باشد. داده‌هایی که دورافتاده هستند، به حالت‌های مختلفی می‌توانند رخ دهند که در زیر چندین حالت آورده شده است:

۱- خطاهایی که ممکن است در متغیر پاسخ یا متغیرهای کمکی و یا هر دو وجود داشته باشند که ممکن است این‌گونه خطاها ناشی از اشتباه در اندازه‌گیری‌ها یا خطاهای ایجاد شده در ورود داده‌ها باشند.

۲- خطای ناشی از در نظر نگرفتن شکل صحیح رابطه ریاضی مدل.

۳- در نظر نگرفتن تبدیل مقیاس مناسب مانند لگاریتم پاسخ یا دیگر تبدیل‌های مورد نیاز (باکس و کاکس، ۱۹۶۴).

۴- قسمت سیستماتیک مدل و مقیاس آن ممکن است درست بوده اما توزیع خطاهای پاسخ متفاوت از نرمال باشند، برای مثال دم‌هایی کلفت‌تر از توزیع نرمال داشته باشد، که در نتیجه استفاده از روش ماکسیمم درستنمایی با فرض نرمال بودن خطاها، برازش مدل را محدود می‌سازد (دراپر و اسمیت، ۱۹۹۸).

پس از شناسایی داده‌های پرت که در بخش بعدی با جزییات بیشتری به آن پرداخته خواهد شد، نوبت به معرفی روش‌ها و نحوه‌ی برخورد با داده‌های پرت است. دو روش مفید به منظور مقابله با نقاط دورافتاده و نافذ وجود دارد، که یکی ابزارهای تشخیصی در رگرسیون (کوک، ۱۹۷۷ و ۱۹۷۹؛ بلسلی و همکاران، ۱۹۸۰) و دیگری رگرسیون استوار می‌باشد (هابر، ۱۹۷۳ و ۱۹۸۱؛ راسو، ۱۹۸۴؛ راسو و یوهای، ۱۹۸۴). این دو روش دارای هدفی یکسان اما عملکردی در خلاف جهت یکدیگر هستند. در ساختار رگرسیون تشخیصی، ابتدا یک رگرسیون برازش داده می‌شود و سپس تحقیق برای تشخیص نقاط دور افتاده‌ی بالقوه

[1]Diagnostic

انجام می‌گیرد تا مدل دوباره با استفاده از داده‌های خوب، برازش داده شود، در حالی که در رگرسیون استوار، مدلی که مناسب اکثریت داده‌ها باشد برازش داده می‌شود و سپس مشاهداتی که دارای مانده‌های بزرگ از این مدل استوار باشند به‌عنوان نقاط دورافتاده شناسایی می‌گردند.

درکسن (۱۹۹۶) استراتژی‌هایی را برای مقابله با مقادیر دورافتاده و گم شده در داده‌های صنعتی ارایه داد و کارایی روش‌های آمار- مبنا و استوار- مبنای میانگین، انحراف استاندارد و ضریب همبستگی را در جداسازی مقادیر دورافتاده از جامعه اصلی مقایسه کرد. علاوه بر این، روش ماکسیمم درستنمایی را برای چگونگی مقابله با مقادیر گمشده در داده‌های چندمتغیره صنعتی ارائه داد.

تشخیص و شناسایی مقادیر دورافتاده و براوردهای استوار رابطه‌ی بسیار نزدیکی به هم دارند (همپل و همکاران، ۱۹۸۶؛ هوبرت و همکاران، ۲۰۰۸) به‌طوری‌که دو موضوع زیر اساساً یکسانند:

۱. براورد استوار: پیدا کردن براوردی که تأثیری از مقادیر دورافتاده در نمونه نمی‌گیرد.

۲. تشخیص مقادیر دورافتاده: پیدا کردن همه دورافتاده‌ها، که می‌توانند براورد را اریب کنند.

به عبارت دیگر براساس براورد استوار امکان شناسایی مقادیر دورافتاده با استفاده از مانده‌های استوار یا فاصله‌ها فراهم می‌شود، در حالی‌که اگر مقادیر دورافتاده را بشناسیم می‌توانیم آن‌ها را حذف کنیم، یا وزن کمتری به آن‌ها اختصاص دهیم و آنگاه از روش‌های کلاسیک براوردیابی استفاده کنیم.

روش‌های مختلفی برای تشخیص مقادیر دورافتاده پیشنهاد شده است. کاتنی و همکاران (۲۰۰۸) مروری بر روش‌های تشخیص مقادیر دورافتاده برای کاربردهای صنعتی ارائه داده‌اند. سِین و گیتل (۲۰۰۶) به برخی کاربردهای داده‌کاوی در تشخیص مقادیر دورافتاده اشاره کرده‌اند. روش‌های استوار مقادیر دورافتاده مانند پیرایش و وینزوریدن[1] روش‌های مشهوری هستند (چمبرز و کوکیک، ۱۹۹۳). ایده اصلی وینزوریدن این است که اگر مشاهده‌ای از یک نقطه قطع k، تجاوز کند، مشاهده به وسیله مقدار k جایگزین خواهد شد. چمبرز و همکاران (۲۰۰۰) دو نوع وینزوردین تعریف می‌کنند. تحت وینزوردین نوع I، مشاهدات اصلی، Y_i ، در نظر گرفته می‌شوند و یک براوردگر از مجموع کل در جامعه

$$\hat{T}_w = \sum_{i \in s} w_i Y_t^*$$

است که در آن

$$Y_i^* = \begin{cases} c & Y_i > k \\ Y_i & o.w \end{cases}$$

و w_i وزن متناظر با مشاهده‌ی iام است.

برای وینزوریدن نوع II، براورد مجموع جامعه با وینزوردین نوع I مساوی است، اما در این حالت

[1] Winsorization

$$Y_i^* = \begin{cases} \frac{Y_i}{w_i} + k\left(w_i - ۱\right)/w_i & Y_i > k \\ Y_i & o.w \end{cases}$$

برای کسب اطلاعات بیشتر در خصوص روش‌های وینزوریدن به فرانسیکو و فولر (۱۹۹۱) و فصل ۴ این طرح پژوهشی مراجعه شود.

مواجهه با مقادیر دورافتاده، به‌ویژه در تحقیقات اقتصادی و پدیده‌های مالی به وفور در نمونه‌ها رخ می‌دهند. روش‌های طرح- مبنا برای مقابله با مقادیر دورافتاده به وسیله کیش (۱۹۶۵) ارایه شده است.

همچنین، خوشه‌بندی یک روش مشهور مورد استفاده برای گروه‌بندی داده‌ها یا شی‌های مشابه است (جین و دوبس، ۱۹۹۸). یک روش مهم برای تحلیل مقادیر دورافتاده برمبنای خوشه‌بندی توسعه پیدا کرده است. در اغلب این روش‌ها، آزمودنی‌های نرمال به خوشه‌های بزرگ و چگال تعلق دارند، درحالی‌که مقادیر دورافتاده، خوشه‌های بسیار کوچک را تشکیل می‌دهند (لوریرو و همکاران، ۲۰۰۴ و نی‌یو و همکاران، ۲۰۰۷).

چمبرز (۱۹۸۶، ۱۹۸۲) روش‌های براورد استوار مقادیر دورافتاده مدل- مبنا را توسعه داده است. همچنین چمبرز و کوکیک (۱۹۹۳)، لی (۱۹۹۱، ۱۹۹۵) و هالیگر (۱۹۹۵) نیز به این مباحث پرداخته‌اند. همچنین در بسیاری از آمارگیری‌های کارگاهی، بعد از شناسایی یک مقدار دورافتاده یا دیگر مشاهدات موثر، غالبا کاهش وزن نمونه مرتبط با آن مشاهده یا مشاهده‌ها، روش استاندارد در تحلیل این داده‌ها است. برای کسب اطلاعات بیشتر و مبحث‌های کلی در روش‌های وزن‌دهی و پیامدهای عملی چنین روش‌هایی به چمبرز (۱۹۹۶) و الیوت و لیتل (۲۰۰۰) مراجعه شود. یک روش عملی برای حل این مشکل در نظر گرفتن وزن مساوی با یک برای این واحدها است.

۱-۳- طبقه‌بندی روش‌های شناسایی داده‌های پرت

همان‌طور که ذکر شد، روش‌های شناسایی داده‌های پرت می‌توانند به دو دسته‌ی روش‌های تک‌متغیره و روش‌های چندمتغیره طبقه‌بندی شوند. یک طبقه‌بندی دیگر از روش‌های شناسایی داده‌های پرت روش‌های پارامتری (آماری) و روش‌های ناپارامتری است (ویلیام و همکاران، ۲۰۰۲).

روش‌های آماری پارامتری یا براساس یک توزیع مشخص برای مشاهدات (هاوکینز، ۱۹۸۰؛ روسیو و لری، ۱۹۸۷؛ بارنت و لویز، ۱۹۹۴) و یا حداقل بر اساس برآوردهای آماری پارامترهای توزیع‌های نامشخص (هادی، ۱۹۹۲؛ کاسینوز و رویز، ۱۹۹۰) عمل می‌کنند. در این روش‌ها، داده‌های پرت به‌عنوان مشاهداتی که از فرضیات مدل دور هستند، شناسایی می‌شوند. این روش‌های شناسایی اغلب برای مجموعه داده‌های با بعد بالا و یا برای مجموعه داده‌هایی که هیچ دانش قبلی از توزیع داده‌ای آن نداریم، روش‌های مناسب نمی‌باشند (پاپادیمیتریو و همکاران، ۲۰۰۲).

در طبقه‌بندی روش‌های ناپارامتری شناسایی داده‌های پرت، روش‌هایی که جدا از روش‌های داده‌کاوی‌اند، روش‌های بر اساس فاصله[1] نامیده می‌شوند که این روش‌ها عمدتاً بر اساس اندازه‌های فاصله‌ی موضعی[2] هستند و قادر به بررسی پایگاه‌های داده‌های بزرگ می‌باشند (کنور و اِن‌جی، ۱۹۹۷؛ کنور و اِن‌جی، ۱۹۹۸؛ فاوسِت و پروؤست، ۱۹۹۷؛ ویلیام و هانگ، ۱۹۹۷؛ هاوکینز و همکاران، ۲۰۰۲؛ ویلیام و همکاران، ۲۰۰۲؛ بِی و اِسکوباکار، ۲۰۰۳). یک دسته‌ی دیگر از روش‌های شناسایی نقاط پرت، روش‌های خوشه‌بندی[3] است که در آن یک خوشه از نمونه‌های کوچک می‌تواند به‌عنوان داده‌های پرت خوشه‌بندی شده در نظر گرفته شود (کاف‌من و روسیو، ۱۹۹۰؛ اِن‌جی و هان، ۱۹۹۴؛ راماسوامی و همکاران، ۲۰۰۰؛ شِخار و چاولا، ۲۰۰۲؛ شِخار و همکاران، ۲۰۰۱؛ آکانا و رُدریج، ۲۰۰۴). روش‌های دیگر شناسایی نقاط پرت شامل روش‌های شناسایی نقاط پرت فضایی است. در این روش‌ها، به دنبال داده‌های فرین یا بی‌ثبات موضعی نسبت به مقادیر مجاور یا همسایه می‌باشیم، اگر چه این مشاهدات ممکن است به‌طور معنی‌دار متفاوت از کل جامعه نباشند (اِسچیف‌من و همکاران، ۱۹۸۱؛ اِن‌جی و هان، ۱۹۹۴؛ شِخار و چاولا، ۲۰۰۲؛ شِخار و همکاران، ۲۰۰۱؛ لو و همکاران، ۲۰۰۳).

اکنون در زیر برخی از روش‌های اشاره شده در بالا بیشتر شرح داده می‌شود. برای کسب اطلاعات بیشتر در مورد دسته‌بندی‌های دیگر روش‌های شناسایی داده‌های پرت به (بارنت و لوییز، ۱۹۹۴؛ پاپادیمیتریو و دیگران، ۲۰۰۲؛ آکانا و رُدریجیوز، ۲۰۰۴؛ هو و سانگ، ۲۰۰۳) مراجعه شود.

۱-۳-۱-روش‌های تک متغیره

یک فرض اصلی در روش‌های آمار- مبنایِ شناسایی داده‌های پرت، ایجاد مدلی است که امکان نمونه‌گیری تصادفی تعداد کمی از مشاهدات از توزیع‌های $G_1,...,G_k$ که متفاوت از توزیع هدف F که دارای توزیع نرمال $N(\mu,\sigma^2)$ است را بوجود آورد (فرگوسن، ۱۹۶۱؛ دیوید، ۱۹۷۹؛ بارنِت و لِویز، ۱۹۹۴؛ گَدِر، ۱۹۸۹؛ دیویس و گَدِر، ۱۹۹۳). مسئله‌ی شناسایی داده‌های پرت به مسئله‌ی مشخص‌سازی آن مشاهداتی که در یک ناحیه‌ی پرت قرار می‌گیرند، تفسیر می‌شود که بر این اساس، تعریف زیر توسط دیویس و گَدِر (۱۹۹۳) ارائه شده است:

برای هر سطح اطمینان α، $0<\alpha<1$، ناحیه‌ی $\alpha-$ پرت توزیع نرمال $N(\mu,\sigma^2)$ به‌صورت زیر تعریف می‌شود:

$$OUT(\alpha,\mu,\sigma^2)=\{x:|x-\mu|>z_{(1-\alpha/2)}\sigma\} \qquad (1)$$

که در آن z_q چندک qام توزیع $N(0,1)$ است. عدد x در صورتی‌که خارج از $|x-\mu|\le z_{(1-\alpha/2)}\sigma$ قرار گیرد، یک $\alpha-$ پرت نسبت به توزیع F است، اگرچه معمولاً توزیع نرمال به‌عنوان توزیع هدف استفاده

[1]Distance- Based Methods
[2]Local Distance Measures
[3]Clustering Methods

می‌شود، اما این تعریف می‌تواند به سادگی به هر توزیع متقارن تک‌مدی با تابع چگالی مثبت نیز تعمیم داده شود. لازم به ذکر است که تعریف نقطه‌ی پرت فوق امکان تشخیص نقاط پرتی که آلوده شده‌اند (از توزیع‌های $G_1,...,G_k$ آمده‌اند) را نمی‌دهد، بلکه امکان تعیین نقاطی که در ناحیه‌ی پرت قرار گرفته باشند را بوجود می‌آورد.

در مجموع روش‌های مختلفی برای بررسی یک متغیره داده‌های پرت (مشاهده پرت از نظر یک متغیر) وجود دارند که می‌توان آن‌ها را به ۲ گروه دامنه و آزمون‌های آماری تقسیم کرد. در روش‌های دامنه توزیع مشاهدات بررسی شده و داده‌های خارج از یک دامنه معین به‌عنوان داده پرت تلقی می‌شوند. مهم‌ترین موضوع در این ارتباط تعیین دامنه یاد شده برای مشخص کردن داده‌های پرت است.

روش سنتی در این مورد، میانگین ($\bar{X}$) به اضافه یا منهای ۳ برابر انحراف معیار (S)[1] می‌باشد (ژانگ و همکاران، ۱۹۹۸؛ هایر و همکاران، ۱۹۹۸؛ چیانگ و همکاران، ۲۰۰۳) و داده‌های بزرگ‌تر از میانگین به اضافه ۳ برابر انحراف معیار و کوچک‌تر از میانگین منهای ۳ برابر انحراف معیار پرت محسوب می‌شوند. چون این روش تحت تأثیر داده‌های پرت است (در محاسبه میانگین و انحراف معیار از تمام داده‌ها از جمله داده‌های پرت استفاده می‌شود)، از این‌رو روش‌های دیگری از جمله میانه به اضافه یا منفی میانه انحراف‌های تمام داده‌ها از میانه[2] و نمودار جعبه‌ای[3] (توکی، ۱۹۷۷؛ ریمن و همکاران، ۲۰۰۵) ارائه شده که تحت تأثیر داده‌ای پرت قرار نمی‌گیرند. در این روش در بخش روش‌های استوار تک متغیره بیشتر توضیح داده خواهد شد. میانه قدرمطلق انحراف‌های تمام داده‌ها از میانه (MAD[4]) از رابطه زیر محاسبه می‌شود:

$$MAD = ۱/٤۸۲ Median(|x_i - x_{median}|) \quad (۲)$$

مقدار ثابت ۱/۴۸۲ برای تبدیل MAD به برآورد نااریبی از انحراف معیار (امید ریاضی انحراف معیار نمونه برابر با انحراف معیار جامعه) داده‌های گوسی (نرمال) است (چیانگ و همکاران، ۲۰۰۳).

نمودار جعبه‌ای نیز از روش‌های دامنه محسوب می‌گردد. این روش نموداری برای نشان دادن موقعیت، پراکنش و چولگی داده‌ها می‌باشد (توکی، ۱۹۷۷) و به وفور برای تشخیص داده‌های پرت استفاده می‌شود (ریمن و همکاران، ۲۰۰۵). این نمودار با استفاده از یک مستطیل (باکس) و دو خط یا میله[5] در دو طرف مستطیل و به وسیله میانه، چارک‌های اول ($Q_۱$) و سوم ($Q_۳$) و کم‌ترین و بیشترین مقادیر رسم می‌شود. طول مستطیل برابر با فاصله چارکی[6] (تفاوت بین چارک سوم و چارک اول یا $IQR = Q_۳ - Q_۱$) است. در یک نوع نمودار جعبه‌ای که از آن برای تشخیص داده‌های پرت استفاده می‌شود، داده‌هایی که کوچک‌تر از $Q_۱ - ۱/۵IQR$ یا بزرگ‌تر از $Q_۳ + ۱/۵IQR$ باشند جزء داده‌های پرت خفیف و داده‌هایی که کوچک‌تر از

---

[1] $\bar{X} \pm ۳S$
[2] Median $\pm$ MAD
[3] Box Plot
[4] Median Absolute Deviation
[5] Whisker
[6] Interquartile Range (IQR)

$Q_1 - 3IQR$ یا بزرگ‌تر از $Q_3 + 3IQR$ باشند جزء داده‌های پرت کرانگین محسوب می‌شوند. در نمودار جعبه‌ای داده‌های پرت خفیف را با علامت (○) و داده‌های پرت کرانگین را با علامت (*) نشان می‌دهند.

از آزمون‌های آماری نظیر آزمون گراب[1] برای تشخیص یک متغیره داده‌های پرت استفاده می‌شود (لالور و ژانگ، ۲۰۰۱). در آزمون گراب فرض بر این است که داده‌ها از توزیع نرمال پیروی می‌کنند. در آزمون یاد شده، در هر مرحله یک داده پرت تشخیص داده می‌شود. در صورتی‌که داده پرتی شناسایی شود، داده یاد شده حذف می‌گردد و آزمون برای بقیه داده‌ها دوباره انجام می‌شود. این‌کار آن قدر ادامه می‌یابد تا هیچ داده پرتی شناسایی نشود. فرض صفر این است که هیچ نوع داده پرتی وجود ندارد و فرض مخالف این است که حداقل یک داده پرت وجود دارد. آماره آزمون گراب (G) از رابطه زیر محاسبه می‌شود:

$$G = \frac{\max|x_i - \bar{x}|}{S} \qquad (۳)$$

که در آن، x_i کوچک‌ترین یا بزرگ‌ترین داده، $\bar{x}$ میانگین داده‌ها و S انحراف معیار داده‌ها می‌باشند. فرض صفر موقعی رد خواهد شد (فرض مخالف یا وجود حداقل یک داده پرت) که

$$G > \frac{(n-1)}{\sqrt{n}} \sqrt{\frac{t^2_{(\frac{\alpha}{2n},n-2)}}{n-2+t^2_{(\frac{\alpha}{2n},n-2)}}} \qquad (۴)$$

که در آن، n اندازه نمونه و $t_{(\frac{\alpha}{2n},n-2)}$ مقدار بحرانی آماره توزیع t استیودنت با درجه آزادی $n-2$ و سطح معنی‌داری $\frac{\alpha}{2n}$، می‌باشد.

۱-۳-۱-۱- روش‌های یک مرحله‌ای در مقابل روش‌های متوالی

دیویس و گَدِر (۱۹۹۳) یک تفاوت مهم بین روش‌های یک مرحله‌ای و روش‌های دنباله‌ای یا متوالی شناسایی داده‌های پرت قائل هستند. در روش‌های تک مرحله‌ای همه‌ی نقاط پرت در یک لحظه شناسایی می‌شوند. در حالی که در روش‌های متوالی، حذف یا اضافه شدن داده‌ها به‌طور متوالی صورت می‌پذیرد. به عبارت دیگر در روش‌های متوالی، در هر مرحله، برای تشخیص نقطه‌ی پرت بودن یا نبودن یک مشاهده، آزمون‌هایی انجام می‌شود.

بر اساس رابطه‌ی (۱)، یک قاعده کلی برای پیدا کردن ناحیه‌ی پرت در حالت یک مرحله‌ای به‌صورت زیر وجود دارد:

$$OUT(\alpha_n, \hat{\mu}_n, \hat{\sigma}_n^2) = \{x : |x - \hat{\mu}_n| > g(n, \alpha_n)\hat{\sigma}_n\} \qquad (۵)$$

که در آن n حجم نمونه، $\hat{\mu}_n$ و $\hat{\sigma}_n$ به ترتیب میانگین و انحراف معیار برآورد شده‌ی توزیع هدف بر اساس نمونه، α_n سطح اطمینان برای آزمون‌های مقایسه‌ای چندگانه و $g(n, \alpha_n)$ حدود نواحی پرت است.

[1] Grubb's Test

عموماً μ_n و σ_n به ترتیب بر اساس میانگین نمونه، $\bar{x}_n$، و انحراف معیار نمونه، S_n، براورد می‌شود. چون این براوردها بسیار تحت تأثیر داده‌های پرت هستند، بنابراین لازم است روش‌های دیگری که نسبت به حضور داده‌های پرت استوار باشد، جایگزین شوند که در این بخش به معرفی این روش‌ها پرداخته خواهد شد.

تصحیح بر اساس مقایسه‌های چندگانه زمانی که آزمون‌های آماری متعددی به‌طور همزمان اجرا شود، مورد استفاده قرار می‌گیرد. ساده‌ترین و محافظه‌کارترین روش، روش تصحیح بونفرونی است که α – مقدار برای کل n مقایسه برابر α است، به گونه‌ای که α – مقدار برای هر مقایسه برابر $\frac{\alpha}{n}$ باشد. یک روش تصحیح ساده‌ی دیگر که مورد استفاده قرار می‌گیرد، استفاده از $\alpha_n = 1-(1-\alpha)^{\frac{1}{n}}$ است. روش‌های تعیین ناحیه‌ی بحرانی $g(n,\alpha_n)$ اغلب توسط روش‌های عددی مانند شبیه‌سازی‌های مونت کارلو برای حجم نمونه‌های مختلف مشخص می‌شود (دیویس و گَدِر، ۱۹۹۳).

۱-۳-۱-۲- روش‌های استوار تک متغیره

میانگین و واریانس نمونه، براوردهای خوبی برای پارامترهای مکان و مقیاس داده‌هایی هستند که با داده‌های پرت آلوده نشده باشند. زمانی که دادگان مورد بررسی با داده‌های پرت آلوده شده باشد، این پارامترها ممکن است به‌طور معنی‌دار عملکرد شناسایی داده‌های پرت را تحت تأثیر قرار دهند. هَمپل (۱۹۷۱، ۱۹۷۴) نقطه‌ی فروریزش را به‌عنوان معیاری برای استواری یک براوردگر در مقابل نقاط پرت معرفی می‌کند. به عبارت دیگر مفهوم نقطه‌ی فروریزش را به‌عنوان معیاری از میزان اغتشاشی که یک براوردگر می‌تواند در مقابل داده‌های پرت تحمل کند و هنوز استوار باقی بماند معرفی کرده است. نقطه‌ی فروریزش، کوچک‌ترین نسبت اغتشاش (نقاط پرت) است که می‌تواند منجر به یک مقدار بزرگ برای براوردگر مورد بررسی شود. بنابراین هر چه نقطه‌ی فروریزش یک براوردگر بزرگ‌تر باشد، آن براوردگر استوارتر است. به‌عنوان مثال میانگین نمونه، دارای نقطه‌ی فروریزش $\frac{1}{n}$ است زیرا یک مشاهده‌ی بزرگ می‌تواند میانگین را بسیار بزرگ کند. از این‌رو، هَمپل، میانه و قدر مطلق انحراف از میانه (MAD) را به‌عنوان براوردگرهای استوار مکان و پراکندگی پیشنهاد کرد. یک روش جدید دیگر توسط توکی (۱۹۷۷) به منظور مقابله با مشکل براوردگرهای استوار پیشنهاد شده است. توکی نمودار جعبه‌ای را به‌عنوان ابزار گرافیکی برای شناسایی داده‌های پرت معرفی کرده است.

۱-۳-۱-۳-کنترل فرایند آماری (SPC)[1]

حوزه‌ی کنترل فرایند آماری به روش‌های تک متغیره‌ی شناسایی داده‌های پرت بسیار مرتبط است. در این روش، حالت‌هایی درنظر گرفته می‌شود که اندازه‌های تک متغیره معرف یک فرایند تصادفی باشد و شناسایی نقاط پرت، نیازمند شناسایی لحظه‌ای و آنی داده‌های پرت است. روش‌های SPC بیش از نیم قرن است که

[1] Statistical Process Control

در حوزه‌ی آمار به شدت مورد استفاده است. بِن‌گال و همکاران (۲۰۰۳) روش‌های *SPC* را بر اساس دو معیار مهم دسته‌بندی کرده‌اند:

- روش‌های مورد استفاده برای داده‌های مستقل در مقابل روش‌های مورد استفاده برای داده‌های وابسته
- روش‌های مدل- ویژه[1] در مقابل روش‌های مدل- عام[2]. روش‌های مدل- ویژه نیازمند فرضیات قبلی روی ویژگی‌های فرایند است که عمدتاً توسط یک توزیع تحلیلی از قبل تعیین شده یا یک عبارت فرم بسته مشخص می‌شود. در روش‌های مدل- عام تلاش برای براورد یک مدل اساسی با داشتن کمترین فرضیات از قبل تعیین شده می‌باشد.

روش‌های سنتی *SPC* مانند شیوهارت[3]، مجموع انباشته (CUSUM)[4] و میانگین متحرک وزنی نمایی (EWMA)[5] ، روش‌های مدل- ویژه برای داده‌های مستقل است. این روش‌ها کاربرد زیادی در صنعت دارند اگرچه فرض‌های استقلال اغلب در عمل برقرار نیست.

عمده‌ترین روش‌های مدل- ویژه برای داده‌های وابسته، بر اساس سری زمانی است که اغلب، اصول اساسی این روش‌ها به‌صورت زیر است: ابتدا یک مدل سری زمانی که بتواند به بهترین روش معرف فرایند خودهمبستگی باشد را پیدا می‌کنیم و سپس از این مدل برای پالایش داده‌ها استفاده می‌کنیم. سپس روش‌های سنتی *SPC* را برای مانده‌ها به کار می‌بریم. در عمل، خانواده‌ی مدل‌های ARIMA[6] برای براورد و پالایش خودهبستگی فرایند انجام می‌شود. تحت فرضیات خاص، مانده‌های مدل ARIMA مستقل و به‌طور تقریبی دارای توزیع نرمال هستند به گونه‌ای که روش‌های سنتی *SPC* می‌تواند برای آن‌ها اعمال شود. از طرف دیگر مدل‌های ARIMA به خصوص روش‌های ساده‌ی آن مانند روش‌های AR می‌توانند به‌طور مؤثر، تغییرات اساسی فرایندهای صنعتی را تشریح کنند (باکس و جنکینز، ۱۹۷۶؛ اَپلِی و شی، ۱۹۹۹).

روش‌های مدل- ویژه برای داده‌های وابسته می‌توانند به دو دسته‌ی روش‌های پارامتر- وابسته[7] که نیازمند براوردهای آشکار برای پارامترهای مدل (آلوان و رُبرت، ۱۹۸۸؛ واردِل و همکاران، ۱۹۹۴؛ لو و رینالد، ۱۹۹۹) و روش‌های فارغ از پارامتر[8] که در آن پارامترهای مدل تنها به‌صورت ضمنی به دست می‌آیند، افراز می‌شوند.

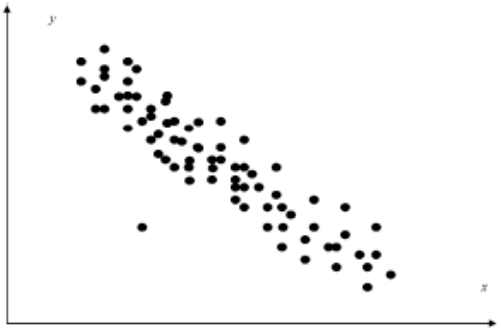
---

[1] Model- specific
[2] Model- generic
[3] Shewhort
[4] Cumulative Sum
[5] Exponential Weighted Moving Average
[6] Auto Regressive Integrated Moving Average
[7] Parameter- dependent
[8] Parameter- free

۱-۳-۲- روش‌های چندمتغیره

تشخیص چند متغیره داده‌های پرت شامل بررسی چند متغیره هر یک از مشاهدات بر اساس ترکیبی از متغیرها است. چون بیشتر تحلیل‌های چند متغیره دارای بیش از دو متغیر است، تعیین داده‌های پرت از نظر مجموعه‌ای ترکیبی از متغیرها نیز ضروری است. به منظور تشخیص چند متغیره داده‌های پرت باید از روش‌ها یا معیارهایی استفاده شود که موقعیت چند بعدی (فضایی) هر یک از مشاهدات را نسبت به یک نقطه مشترک نشان دهند (هایر و همکاران، ۱۹۹۸). روش‌های مختلفی نظیر تجزیه به مؤلفه‌های اصلی (ژانگ و همکاران، ۱۹۹۹، چیانگ و همکاران، ۲۰۰۳؛ لالور و ژانگ، ۲۰۰۱)، رگرسیون چند متغیره (لالور و ژانگ، ۲۰۰۱)، شبکه‌های عصبی (لالور و ژانگ، ۲۰۰۱)، الگوریتم ژنتیک (ویگاند و همکاران، ۲۰۰۹) و روش‌های مبتنی بر فاصله ماهالانوبیس[1] (روسیوو و ون دریسن، ۱۹۹۹؛ فیلزموزر و همکاران، ۲۰۰۵) به این منظور ارائه شده‌اند. از میان روش‌های یاد شده، روش‌های مبتنی بر فاصله ماهالانوبیس معروف‌ترین می‌باشد.

در بسیاری موارد، مشاهدات چندمتغیره زمانی که هر متغیر به‌طور مستقل در نظر گرفته شود، نمی‌تواند نقاط پرت را شناسایی کند و دراین مواقع، شناسایی نقاط پرت تنها زمانی امکان‌پذیر است که تحلیل‌های چند متغیره انجام شود و اثرات متقابل بین متغیرهای مختلف درون گروه داده‌ها با هم مقایسه شوند. یک مثال ساده در شکل ۱-۱ نمایش داده شده است که در آن نشان می‌دهد هر داده دارای دو اندازه در فضای دو بعدی است و مشاهده‌ی چپ پایینی، یک نقطه‌ی پرت چند متغیره است و نه یک نقطه‌ی پرت تک متغیره. زمانی که هر مشاهده را به‌طور مجزا نسبت به پراکندگی مقادیر در امتداد محورهای x و y در نظر می‌گیریم، آن‌چه دیده می‌شود این است که این نقاط به مرکز توزیع تک متغیره نزدیک است. بنابراین هر آزمونی که برای شناسایی نقاط پرت در نظر گرفته می‌شود باید ارتباط بین این دو متغیر را در نظر گیرد که در این صورت غیر عادی بودن و ناهنجاری بین داده‌ها مشخص می‌شود.



شکل ۱-۱: یک فضای دو بعدی با یک مشاهده‌ی پرت

[1] Mahalanobis Distance

تأثیر نقاط پرت در براوردها به دو گونه ممکن است رخ دهد: اثرات درون‌آوری[1] و برون‌بری[2] می‌تواند در مجموعه داده‌های شامل نقاط پرت چندبعدی یا خوشه‌های شامل نقاط پرت رخ دهد. در زیر به بیان تعریفی شهودی که توسط آکوانا و ردریجز (۲۰۰۴) ارائه شده است، می‌پردازیم.

اثر درون‌آوری: این اثر بیان می‌کند که یک نقطه‌ی پرت، نقطه‌ی پرت دوم را پنهان می‌کند. یعنی این‌که نقطه‌ی پرت دوم فقط به خودی خود می‌تواند به‌عنوان یک نقطه‌ی پرت در نظر گرفته شود و نه در حضور نقطه‌ی پرت اول. بنابراین بعد از حذف نقطه‌ی پرت اول، دومین مشاهده به‌عنوان یک نقطه‌ی پرت، پدیدار می‌شود. در واقع درون‌آوری زمانی اتفاق می‌افتد که یک خوشه از مشاهدات پرت، براورد میانگین و کواریانس را به طرف خود اریب کند که نتیجه‌ی آن کوچک شدن فاصله‌ی نقاط پرت از میانگین است.

اثر برون‌آوری: این اثر بیان می‌کند که یک نقطه‌ی پرت، مشاهده‌ی دوم را دورافتاده جلوه می‌دهد به این معنی که مشاهده‌ی دوم تنها در حضور مشاهده‌ی اول به‌عنوان یک نقطه‌ی پرت در نظر گرفته می‌شود. به عبارت دیگر، پس از شناسایی اولین نقطه‌ی پرت، دومین مشاهده به یک مشاهده‌ی غیر نقطه‌ی پرتی تبدیل می‌شود. برون‌بری زمانی رخ می‌دهد که یک گروه از نمونه‌های دورافتاده، براوردهای میانگین و کواریانس را به طرف خودشان بکشانند و از بقیه‌ی نمونه‌های غیردورافتاده دور کنند که نتیجه‌ی آن زیاد شدن فاصله‌ی این نمونه‌ها از میانگین و نقطه‌ی پرت به نظر رسیدن این نمونه‌ها است. یک روش یک مرحله‌ای با درون‌آوری و برون‌بری خفیف و کم در ایگلیوس و مارتینز (۱۹۸۲) وجود دارد.

بنابراین می‌توان گفت که در داده‌های آمارگیری‌های کارگاهی تمایل به استفاده از روش‌های چندمتغیره، به دلیل جمع‌آوری اندازه‌گیری‌های متفاوت از پاسخ‌گویان نمونه وجود دارد. در اکثر این موارد، دو یا چند متغیر اندازه‌گیری‌شده از این داده‌های چندمتغیره دارای توزیع‌های چوله یا دم-کلفت هستند، که در این موارد نیز بحث مقادیر دورافتاده مطرح می‌شود. در این موارد، روش‌های دورافتاده برای یک یا چند متغیر کلیدی اندازه-گیری‌شده اعمال می‌شود که ممکن است از تابعی مشخص مثلاً نسبت دو متغیر اندازه‌گیری شده استفاده شود. استفاده از این روش‌ها دارای مزیت سادگی هستند، اما ممکن است منجر به پیچیدگی‌های متعددی نیز گردند. نکته‌ی اول این است که امکان دارد یک بردار مقدار بزرگی داشته باشد (اندازه‌گیری مقادیر می‌تواند از طریق فاصله ماهالانوبیس یا اندازه‌های چندمتغیره مشابه آن برای توزیع‌های رایج باشد) اما در عین حال هر یک از مؤلفه‌های آن در داخل فاصله‌های تحمل تعیین شده باشند (فرانکلین و همکاران، ۲۰۰۰). نکته‌ی دوم این است که اگر مولفه‌های یک متغیره در برخی شرط‌های تابعی صدق کند، استفاده مستقیم از وینزوردین یا دیگر تغییرات، ممکن است مشکل‌ساز باشد. به همین دلیل برخی نویسندگان از جمله چمبرز (۱۹۹۶) استفاده از روش‌های کاهش وزنی برای مقادیر دورافتاده چندمتغیره با عملیاتی نسبتا ساده‌تر را پیشنهاد کرده‌اند. این روش اگر چه منجر به کاهش وزن مقادیر یک متغیره می‌شود، اما عدم تجاوز از فاصله‌های تحمل را تضمین می‌کند. یک روش دیگر مورد استفاده توسط برخی ادارات، استفاده از ابزار شناسایی داده‌های پرت برای یک

---

[1] Masking
[2] Swamping

اندازه بحرانی است. به‌عنوان مثال در آمارگیری بررسی بیماری‌های شغلی و صدمات (SOII)[1] که توسط اداره‌ی آمار کار آمریکا (BLS)[2] انجام می‌شود، یک روش ساده از تقسیم کردن تعداد صدمات بر تعداد ساعات کار گزارش‌شده برای هر کارگاه مورد بررسی، به‌عنوان اندازه بحرانی درنظر گرفته می‌شود.

یک تقسیم‌بندی دیگر برای روش‌های شناسایی نقاط پرت چندمتغیره بر اساس پارامترهای براورد شده‌ی توزیع و روش‌های داده‌کاوی که فارغ از پارامتر هستند، می‌باشد. روش‌های آماری مورد استفاده برای شناسایی چند متغیره‌ی نقاط پرت، اغلب به مشاهداتی اشاره دارد که نسبتاً از مرکز توزیع داده‌ها دور باشند. اندازه‌های فاصله‌ای متعددی برای چنین مطالعاتی مورد استفاده قرار می‌گیرد. فاصله‌ی ماهالانوبیس یک معیار معروف است که به پارامترهای براورد شده‌ی توزیع چند متغیره وابسته است.

فاصله ماهالانوبیس به‌عنوان معیاری از موقعیت چند بعدی هر یک از مشاهدات نسبت به مرکز ثقل کل مشاهدات عمل می‌کند. به عبارت دیگر فاصله یاد شده، معیاری از فاصله هر یک از مشاهدات در فضای چند بعدی از مرکز میانگین تمام مشاهدات است. برتری عمده فاصله ماهالانوبیس نسبت به سایر فاصله‌ها، در نظر گرفته شدن ماتریس کواریانس در آن است (فلیزموزر و همکاران، ۲۰۰۵)، چون شکل و اندازه داده‌های چند متغیره به وسیله ماتریس کواریانس تعیین می‌شود.

با معلوم بودن n مشاهده از مجموعه داده‌های p بعدی، فاصله‌ی ماهالانوبیس برای نقطه‌ی i ام از داده‌های چند متغیره، $i = ۱,...,n$ ، با M_i نمایش داده می‌شود و به‌صورت زیر تعریف می‌شود:

$$M_i = \left((x_i - t)^T C^{-۱}(x_i - t)\right)^{\frac{۱}{۲}}$$

که در آن $t = \bar{x}_n$ به‌عنوان برآورد میانگین نمونه و $C = V$ به‌عنوان برآورد ماتریس کواریانس به‌صورت زیر تعریف می‌شود:

$$V = \frac{۱}{n-۱}\sum_{i=۱}^{n}(x_i - \bar{x}_n)(x_i - \bar{x}_n)^T$$

در نتیجه، مشاهدات با فاصله‌ی ماهالانوبیس بزرگ نشان دهنده‌ی نقاط پرت است. لازم به ذکر است که اثرات درون‌آوری و برون‌آوری، نقش مهمی در کفایت و شایستگی فاصله‌ی ماهالانوبیس به‌عنوان معیاری برای شناسایی داده‌های پرت دارند. به این معنا که اثرات درون‌آوری ممکن است فاصله‌ی ماهالانوبیس یک داده‌ی پرت را کم کند. به‌عنوان مثال، زمانی که یک خوشه‌ی کوچک از نقاط پرت داشته باشیم، این مشاهدات، $\bar{x}_n$ و V را به سمت خودشان نزدیک می‌کنند. از طرف دیگر اثر برون‌بری ممکن است فاصله‌ی ماهالانوبیس مشاهدات غیر دورافتاده را افزایش دهد. به‌عنوان مثال زمانی که یک خوشه‌ی کوچک از نقاط پرت، $\bar{x}_n$ را به طرف خود بکشد منجر می‌شود که V به درستی برآورد نشود.

---

[1] Survey of Occupational Illnesses and Injuries (SOII)
[2] Bureau of Labor Statistics (BLS)