NASTINFO

ویژگی‌های مشترک قوانین تجربی معروف در علم‌سنجی:

نگاهی از زاویه دسته‌بندی داده‌ها براساس توزیع فراوانی

محمد توکلی‌زاده راوری | فرامرز سهیلی

چکیده

هدف: بررسی چهار قانون معروف در حوزه کتاب‌سنجی و علم‌سنجی (لوتکا، برادفورد، زیپف، و پرتو) از زاویه دسته‌بندی داده‌های آنها براساس توزیع فراوانی.

روش/ رویکرد پژوهش: روش سندی- تحلیلی. با بررسی داده‌های مرتبط با قواعد چهارگانه علم‌سنجی به‌دسته‌بندی داده‌ها براساس توزیع فراوانی آنها پرداخته است.

یافته‌ها: شکل اولیه داده‌ها در این چهار توزیع، حاوی رتبه هر موجودیت، فراوانی ویژگی مورد مطالعه، و مقدار فراوانی مطلق آن ویژگی در آن موجودیت است. در هریک از این قوانین، موجودیت‌ها براساس میزان فراوانی در ویژگی مورد نظر به چند دسته تقسیم می‌شوند: پرتو (۲ دسته)، برادفورد (۳ دسته)، لوتکا (به تعداد حداکثر فراوانی)، و زیپف (به تعداد واژگان متن).

نتیجه‌گیری: نتایج مربوط به نحوه دسته‌بندی موجودیت‌ها و تعیین تعداد موجودیت‌های درون هر دسته حاکی از آن است که تفاوت بین فراوانی دسته‌ها تابع یک توزیع مشخص است. قوانین چهارگانه را می‌توان برای دسته‌بندی موجودیت‌های مختلف براساس ویژگی‌های متعدد تعمیم داد.

کلیدواژه‌ها

قانون لوتکا، قانون برادفورد، قانون زیپف، قانون پرتو، توزیع آماری، قانون توانی، علم‌سنجی، قوانین تجربی

ویژگی‌های مشترک قوانین تجربی معروف در علم‌سنجی:
نگاهی از زاویه دسته‌بندی داده‌ها براساس توزیع فراوانی

محمد توکلی‌زاده راوری[1]
فرامرز سهیلی[2]

تاریخ دریافت: ۹۳/۱۱/۱٤
تاریخ پذیرش: ۹٤/۰٤/۳۱

مقدمه

کشف قواعد، توزیعات، یا قوانین از اساسی‌ترین زیربناهای توسعه حوزه‌های سنجشی است (هود و ویلسون[3]، ۲۰۰۱) و بخش بزرگی از مطالعات علم‌سنجی به توزیع فراوانی موجودیت‌های علمی و روابط میان آنها می‌پردازد. موجودیت‌های علمی به متغیرهایی اطلاق می‌شود که به‌گونه‌ای با علم و فناوری در ارتباط هستند. مدارک و منابع علمی و فنی، همراه با اطلاعات کتابشناختی و سایر فراداده‌های مرتبط با آنها را می‌توان موجودیت‌های علمی و فنی نامید. برخی منابع علمی عبارت‌اند از: مقالات نشریات، مقالات همایش‌ها، صفحات وب، و پروانه‌های ثبت اختراع. موجودیت‌های اطلاعاتی مانند نویسندگان، موضوعات، استنادها، تاریخ تولید، و زبان، نمونه‌هایی از اطلاعات کتابشناختی هستند.

یکی از فنونی که می‌توان از طریق آنها به مطالعات مربوط به توزیع فراوانی موجودیت‌های علمی و فنی پرداخت، به‌کارگیری قوانین تجربی معروف در علم‌سنجی و کتاب‌سنجی است. به این قوانین، اصطلاح "توزیع" نیز اطلاق می‌شود. اینکه کدام قوانین تجربی را می‌توان قوانین پایه در این حوزه نامید، بیشتر از همه به زیپف، لوتکا، و برادفورد اشاره می‌شود. به‌طور نمونه، اوکنور و ووس[4] (۱۹۸۱) بر این باورند که توزیع‌های برادفورد، لوتکا، و زیپف به‌عنوان قوانین پایه کتاب‌سنجی شناخته می‌شوند و هریک از اینها به‌صورت تجربی حاصل شده‌اند. این توزیع‌ها مانند بخش‌هایی خاص از یک هذلولی شبیه هم هستند؛ در کنار آن، قانون پرتو نیز آمده است. از این رو، در مقاله حاضر به این چهار قانون پرداخته می‌شود. نام

۱. استادیار گروه علم اطلاعات و دانش‌شناسی، دانشگاه یزد
mravari@yahoo.com
۲. استادیار گروه علم اطلاعات و دانش‌شناسی، دانشگاه پیام نور (نویسنده مسئول)
fsohieli@gmail.com

3. Hood & Wilson
4. O'connor & Voos

این قوانین، برگرفته از نام کاشفان آنها جورج کینگزلی زیپف[1]، آلفرد جیمز لوتکا[2]، ساموئل کلمنت برادفورد[3]، و ویلفردو پرتو[4] است (برادفورد، ۱۹۳٤ و ۱۹۸۵؛ زیپف، ۱۹۳۲؛ لوتکا، ۱۹۲٦؛ و پرتو، ۱۹٦٤).

قوانین تجربی در برابر قوانین طبیعی قرار می‌گیرند و به توصیف الگوهایی می‌پردازند که قاعده‌مند و تکرارپذیرند و همانند قوانین طبیعی دو خاصیت اساسی دارند: نخست) این امکان را به ما می‌دهند که رخدادها را پیش‌بینی کنیم یا عکس‌العمل خود را نسبت به آنها مشخص کنیم؛ دوم) ممکن است منجر به نظریه‌هایی شوند که توضیح می‌دهند چرا یک الگوی خاص اتفاق می‌افتد (دروت[5]، ۱۹۸۱). یکی از کاربردهای آنها در علم‌سنجی، دسته‌بندی موجودیت‌های علمی برای اهداف گوناگون است. این اهداف از دسته‌بندی نشریات برای فراهم‌آوری مجموعه کتابخانه تا کاهش تعداد موجودیت‌ها برای سنجش رابطه میان آنها از طریق فنون ترسیم نقشه علم گسترده است. از این رو، مقاله حاضر با هدف مطالعه ویژگی‌های مشترک قوانین تجربی معروف در علم‌سنجی از زاویه دسته‌بندی موجودیت‌های علمی براساس توزیع فراوانی تدوین شده است.

هرگاه موجودیت‌ها براساس شباهت‌شان دسته‌بندی یا خوشه‌بندی شوند مشاهده می‌شود که اندازه بسیاری از موجودیت‌هایی که در یک خوشه قرار گرفته‌اند نزدیک به هم است. سرعت ماشین‌ها در بزرگراه‌ها، وزن سیب‌ها، فشار هوا، سطح دریا، و دمای هوای یک شهر نمونه‌ای از آن است. اگرچه در اندازه آنها تفاوت‌هایی وجود دارد، اما توزیع احتمالی آنها از یک مقدار خاص خیلی دور نیست. در حقیقت، این مقدار خاص نماینده بسیاری از مشاهدات است. به‌عنوان نمونه، قد مردان یک کشور حدود ۱۸۰ سانتی‌متر است، به این معنا که قد سایر مردان آن کشور از این مقدار کمی بیشتر یا کمتر است. حتی بزرگ‌ترین انحراف از این مقدار که استثناء و بسیار نادر است، انتظار می‌رود خیلی از آن دور نباشد. اما توزیع همه موجودیت‌ها از این قاعده پیروی نمی‌کند (کلاوزت[6] و همکاران، ۲۰۰۹). از این رو، درک این چهار قانون معروف در حوزه کتاب‌سنجی و علم‌سنجی مستلزم داشتن دانش در زمینه توزیع آماری براساس قواعد خطی، توانی، نمایی، و لگاریتمی است. از این میان، توزیع‌های توانی و نمایی همخوانی بهتری با توزیع‌های تجربی مشاهده‌شده دارند و در بسیاری از حوزه‌ها نتایج مشابهی طی سال‌ها تکرار شده است (میتزن‌ماچر[7]، ۲۰۰٤). البته، توزیع توانی به‌دلیل خواص ریاضی‌اش توجه بیشتری را در طول سال‌ها به‌خود جلب کرده است؛ زیرا این خواص گاهی به نتایج فیزیکی جالبی می‌انجامد. این توجه همچنین به‌دلیل منطبق بودن آن بر بسیاری از موجودیت‌های طبیعی و ساخت دست بشر است. جمعیت شهرها و شدت زلزله‌ها نمونه‌هایی منطبق بر توزیع توانی هستند. برای موجودیت‌هایی که چنین توزیعی

1. George Kingsley Zipf
2. Alfred James Lotka
3. Samuel Clement
4. Bradford
5. Drott Vilfredo Pareto
6. Clauset
7. Mitzenmacher

دارند، میانگین معنادار نیست (کلاوزت و همکاران، ۲۰۰۹).

مفهوم توزیع توانی این است که اشیا و موجودیت‌هایی که در یک ویژگی ارزش و فراوانی بالایی دارند، مثل تعداد نویسندگان پرکار و شهرهای پرجمعیت، کم است (میلوجویچ[1]، ۲۰۱۰). این اتفاق معمولاً زمانی می‌افتد که تعداد موجودیت‌ها خیلی زیاد باشد. بنابراین، "اگر مقدار یا فراوانی موجودیت‌ها در کل کم باشد، توزیع توانی حاصل نمی‌شود" (ون ران[2]، ۲۰۰۱). البته، گاهی به‌جز توزیع توانی ممکن است اصطلاحات دیگری برای آن نیز به‌کار برود. برای نمونه، فرانسچت[3] (۲۰۰۸) بیان می‌دارد توزیع پرتو که توزیع توانی هم نامیده می‌شود، به‌عنوان وسیله‌ای برای مدل‌سازی موجودیت‌ها مورد استفاده قرار می‌گیرد. نمونه‌ای از موجودیت‌هایی که از توزیع پرتو پیروی می‌کنند عبارت‌اند از زیپف و لوتکا، یا به‌گفته میتزن‌ماچر (۲۰۰٤) "از آن به‌عنوان توزیع دنباله بزرگ، پرتو، زیپف و غیره یاد می‌شود".

با توجه به اهمیت قوانین تجربی در حوزه‌های سنجشی علم اطلاعات و دانش‌شناسی، مطالعات چندانی وجود ندارد که به‌صورت یک‌جا به نقاط مشترک و تفاوت این قوانین پرداخته و به کارکرد اصلی آنها (یعنی دسته‌بندی داده‌ها) خارج از پیچیدگی‌های ریاضی و به‌صورت کاربردی توجه کرده باشد. این موضوع، اهمیت مطالعه حاضر را روشن می‌سازد. مسئله مطالعه حاضر، روشن نبودن زوایای مختلف شباهت‌ها و تفاوت‌های قوانین تجربی در حوزه‌های سنجشی علوم اطلاعات و دانش‌شناسی از جنبه دسته‌بندی داده‌ها و براساس توزیع فراوانی آنها با بیانی خارج از فرمول‌های پیچیده ریاضی است.

هدف پژوهش حاضر، مطالعه ویژگی‌های مشترک قوانین تجربی یا توزیع‌های معروف در علم‌سنجی از زاویه دسته‌بندی داده‌ها براساس توزیع فراوانی است. برای رسیدن به این هدف، اهداف ویژه زیر مورد توجه قرار گرفته است:

۱. مطالعه شکل اولیه داده‌ها در توزیع‌های مورد مطالعه؛
۲. مطالعه تعداد دسته‌ها در هریک از توزیع‌های مورد مطالعه؛
۳. بررسی نحوه دسته‌بندی در هریک از توزیع‌های مورد مطالعه؛ و
٤. چگونگی تعمیم‌پذیری توزیع‌های مورد مطالعه.

مروری بر مطالعات گذشته نشان می‌دهد که در بسیاری از مطالعات مربوط به توسعه حوزه‌های سنجشی علوم اطلاعات و دانش‌شناسی (مانند کتاب‌سنجی، علم‌سنجی، و اطلاع‌سنجی) نگاه اصلی به قوانین تجربی این حوزه معطوف است که اهمیت این قوانین در حوزه‌های سنجشی را نشان می‌دهد. به‌عنوان نمونه، هرتزل[4] (۱۹۸۷) از طریق آمارها و کتابشناسی‌ها به مطالعه ریشه‌های توسعه کتاب‌سنجی می‌پردازد. او به قوانین کتاب‌سنجی به‌عنوان اساس توسعه این حوزه توجه کرده است. برادوس[5] در سال ۱۹۸۷ به تاریخ

1. Milojević
2. Van Raan
3. Franceschet
4. Hertzel
5. Broadus

کتاب‌سنجی تا سال ۱۹۶۹ می‌پردازد. سال ۱۹۶۹ با ظهور اصطلاح کتاب‌سنجی توسط پریچارد[1] مصادف است. او در مطالعه تاریخچه کتاب‌سنجی به مطالعات مربوط به سه قانون کتاب‌سنجی، تحلیل استنادی، و استفاده از کتابخانه توجه دارد. ویلسون (۱۹۹۹) در سلسله بررسی‌های سالانه در حوزه علوم و فناوری اطلاعات[2] به موضوع اطلاع‌سنجی می‌پردازد و به تفصیل نشان می‌دهد که بین سه قانون زیپف، لوتکا، و برادفورد ارتباط وجود دارد.

پس از معرفی قوانین یادشده توسط کاشفان آنها کارهای بسیاری در زمینه قوانین تجربی صورت گرفته است. از جدیدترین آنها می‌توان به مطالعه اگه[3] (۲۰۰۵) اشاره کرد. وی به مسائل نابرابری یا تمرکز در توابع زیپف و لوتکا توجه دارد و پس از یک نگاه مختصر به ارتباط بین توابع آنها نشان می‌دهد که قانون تمرکز پرایس منطبق بر قانون زیپف است.

مایر[4] (۲۰۱۳) با هدف بهبود فرایند جستجو به توزیع فراوانی در سه ناحیه مختلف برادفورد می‌پردازد. او می‌خواهد بداند که آیا برادفوردسازی[5] می‌تواند جستجو را بهبود ببخشد. به‌عبارتی، او کارآمدی برادفوردسازی را به‌عنوان یک فن کتاب‌سنجی در بازیابی اطلاعات مورد ارزیابی قرار داد. وی در نظر داشت تا با برادفوردسازی، مجموعه مدارکی را که از نمایه‌نامه‌ها و چکیده‌نامه‌های سنتی حاصل شده است دوباره رتبه‌بندی کند و آنها را به مدارک هسته و فرعی تقسیم نماید.

آوس لوس[6] (۲۰۱۴) نشان داد که به هر دانشمند براساس میزان همکاری در نوشتن می‌توان یک رتبه مبنی بر اهمیت اختصاص داد. این رتبه به میزان همکاری او در انتشار مقالات بستگی دارد. در این پژوهش، براساس قانون زیپف- ماندل بروت - پرتو نشان داده شد که رابطه بین رتبه افراد و میزان مشارکت آنها در نوشتن مقاله براساس یک توزیع توانی است.

مطالعه در زمینه قوانین تجربی یادشده، صرفاً محدود به حوزه‌های سنجشی مربوط به علوم اطلاعات و دانش‌شناسی نیست. مثلاً در زمینه زیپف که مبدأ آن زبان‌شناسی است، از همان ابتدا تاکنون در مباحث زبانی به آن توجه شده است. یکی از کارهای دهه ۵۰ میلادی مربوط به میلر و نیومن[7] (۱۹۵۸) است که به توضیح آماری رابطه بین رتبه و فراوانی واژه‌ها در متون انگلیسی توجه کردند. این نوع مطالعات هنوز ادامه دارد. گلبوخ و سیدورف[8] (۲۰۰۱) ضرایب قوانین زیپف و هیپ را مطالعه کردند. در حوزه‌های دیگر نیز این قوانین بسط یافته است، مانند حوزه نظریه اطلاعات (هارموییز و تاپسوی[9]، ۲۰۰۵)، جمعیت شهرها (مالاکارنی[10] و همکاران، ۲۰۰۲)، ترافیک وب (ژوهانسون و سورنتی[11]، ۲۰۰۰)، امور مالی و تجارت (آکستل[12]، ۲۰۰۱)، و بسیاری موضوعات دیگر.

1. Pritchard
2. Annual Review of Information Science and Technology (ARIST)
3. Egghe
4. Mayr
5. Bradfordization
6. Ausloos
7. Miller & Newman
8. Gelbukh & Sidorov
9. Harremoees & Topsoe
10. Malacarne
11. Johansen & Sornette
12. Axtell

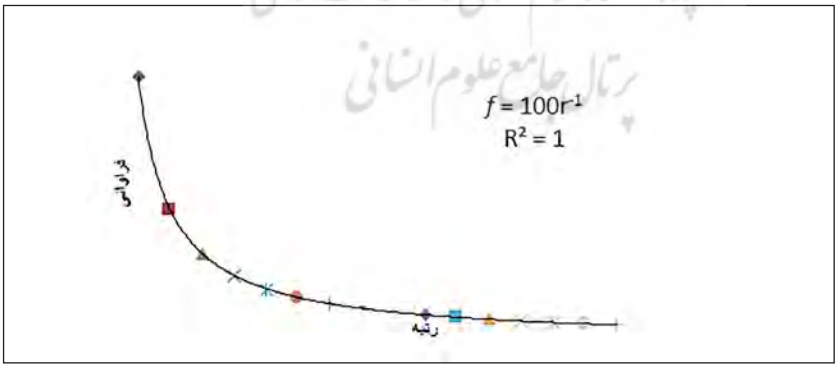
روش‌شناسی

مطالعه حاضر توصیفی و از لحاظ هدف، کاربردی و از جنبه ابزار و منبع داده‌های مورد نیاز از پژوهش‌های سندی- تحلیلی محسوب می‌شود. برای انجام این پژوهش، آثار مرتبط اصلی که توسط کاشفان قوانین مورد نظر نوشته شده، و نیز آثاری که توسط دیگران درباره این قوانین نگاشته شده، و در انتها، پژوهش‌هایی که از این قوانین در آنها بهره گرفته شده است مطالعه و بررسی شد. با توجه به اهداف ویژه، این مطالعات و بررسی‌ها از زاویه شکل اولیه داده‌ها، تعداد دسته‌ها، نحوه دسته‌بندی و توابع و ضرایب حاصل از این دسته‌بندی‌ها، و نحوه تعمیم آنها در هریک از توزیع‌ها انجام شد. علاوه بر آن، منابعی که به توابع رگرسیونی، به‌ویژه توابع نمایی و توانی، از نگاه قوانین تجربی و توزیع فراوانی پرداخته‌اند، بررسی شدند.

یافته‌ها

۱. شکل اولیه داده‌ها

اولین گام عملی در هریک از این توزیعات (زیپف، برادفورد، لوتکا، و پرتو) را می‌توان روی یک محور مختصات دوبعدی نشان داد. در محور افقی، همیشه رتبه و در محور عمودی، فراوانی قرار دارد. اگر به ریشه تاریخی هریک از این قوانین نگاه کنیم موجودیتی که بررسی می‌شود متفاوت است. موجودیت‌های مورد توجه در این چهار قانون به این صورت است: قانون زیپف به واژه‌های یک متن، قانون لوتکا به نویسندگان در یک مجموعه، قانون برادفورد به مدارک مرتبط با یک موضوع در نشریات، و قانون پرتو به افراد جامعه از نظر ثروت و دارایی می‌پردازد. نکته دیگر این است که همیشه موجودیت‌هایی که فراوانی آنها سنجیده شده است از زیاد به کم در یک ستون تنظیم می‌شوند. این فرایند سبب می‌شود که توزیع همیشه شیب منفی داشته باشد.



نمودار ۱. نمونه‌ای از یک توزیع توانی

در این چهار توزیع، اگر سه ستون داده داشته باشیم، ستون اول حاوی رتبه هر موجودیت براساس فراوانی ویژگی مورد مطالعه است. در این ستون، رتبه نشریه براساس مقالات مرتبط (در قانون برادفورد)، رتبه نویسنده براساس میزان مشارکت در نوشتن مقاله (در قانون لوتکا)، رتبه واژه براساس تکرار در متن (در قانون زیپف)، و رتبه فرد براساس میزان دارایی (در قانون پرتو) آورده می‌شود. ستون دوم، حاوی مقدار فراوانی مطلق ویژگی مورد نظر در موجودیت مربوط است. ستون سوم، فراوانی را به‌صورت تراکمی نشان می‌دهد. بنابراین، شباهت توزیع‌های لوتکا، برادفورد، زیپف، و پرتو عجیب نیست. این توزیع‌ها برپایه رتبه و فراوانی (یا رابطه رتبه- مقدار) قرار دارند که موجودیت‌ها پس از طبقه‌بندی رتبه‌بندی می‌شوند (اوکنور و ووس[1]، ۱۹۸۱).

جدول ۱. شکل اولیه داده‌ها براساس توزیع آماری

رتبه موجودیت (محور افقی)	فراوانی مطلق ویژگی (محور عمودی)	فراوانی تراکمی ویژگی
۱	۹۳	۹۳
۲	۸۶	۱۷۹
۳	۵۶	۲۳۵
۴	۴۸	۲۸۳
۵	۴۶	۳۲۹
۶	۳۵	۳۶۴

سنجیدن رابطه بین ستون اول (رتبه) و دوم (فراوانی) در بسیاری از موارد به یک توزیع توانی با توان منفی منجر خواهد شد که تابع آن از فرمول زیر پیروی می‌کند:

$$f = cr^{p}$$

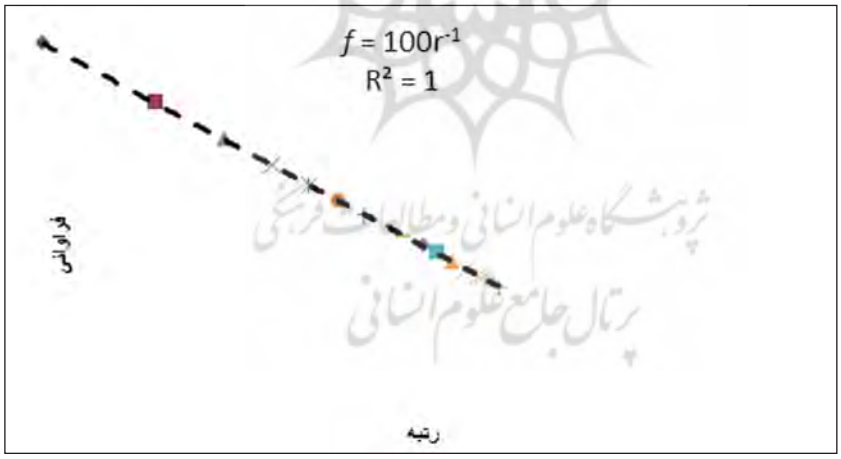
فرمول ۱. تابع توزیع توانی

f مقدار فراوانی موجودیت در ویژگی مورد نظر (مقدار فراوانی قابل انتظار) است؛ c یک عدد ثابت است و نشان می‌دهد که اگر رتبه، صفر باشد مقدار فراوانی چقدر است؛ r رتبه هر موجودیت از نظر فراوانی ویژگی مورد نظر را نشان می‌دهد؛ و p عددی منفی و توان توزیع است. به این ترتیب، براساس این تابع می‌توان پیش‌بینی کرد که فراوانی ویژگی مورد نظر برای هر رتبه چقدر است. براساس خاصیتی که توزیع توانی منفی دارد، با تنزل رتبه یک

1. O'connor & Voos

موجودیت، فراوانی ویژگی مورد نظر در آن متناسب با رتبه کاهش می‌یابد.

در این نوع توزیع، اگر بخواهیم تصویر نمودار را به‌جای خط منحنی به‌صورت خط صاف نمایش دهیم باید هر دو محور افقی و عمودی را به‌صورت لگاریتمی درجه‌بندی کنیم که این کار توسط نرم‌افزارهایی مانند اکسل به‌سادگی میسر است؛ از این رو، به این نوع توزیع، Log–log نیز گفته می‌شود. نیومن[1] (۲۰۰۵) بیان می‌دارد: "برای اینکه شکل توزیع فراوانی را نشان دهیم، بهتر است آن را روی نمودار قرار دهیم و محور عمودی و افقی را به‌صورت لگاریتمی در بیاوریم. وقتی چنین کاری انجام شد، رابطه به‌صورت یک خط مستقیم نشان داده می‌شود. این کار باعث می‌شود که بتوان مشاهده کرد کدام بخش از داده‌ها بر این توزیع منطبق نیست. مثلاً ممکن است در بعضی موارد، دنباله نمودار بر آن خط صاف منطبق نباشد. در این حالت، یکی از گزینه‌ها می‌تواند کنار گذاشتن داده‌های انتهایی باشد. اما مشکل اینجاست که گاهی اطلاعات سودمندی وجود دارد که ممکن است از دست برود. به‌علاوه، بسیاری از داده‌ها در دنباله نمودار دارای توزیع توانی هستند. این خطر وجود دارد که ما بچه را همراه با آب تشت بیرون بیندازیم". در توزیع‌های نمایی اگر محور افقی و در توزیع‌های لگاریتمی اگر محور عمودی به‌صورت لگاریتمی درجه‌بندی شوند، چنین خط صافی حاصل می‌شود. از این رو، به توزیع‌های نمایی Log-normal و به توزیع لگاریتمی Normal-log نیز اطلاق می‌شود.



نمودار ۲. نمایش یک توزیع توانی به‌صورت Log–log

در توزیع برادفورد، ستون اول حاوی رتبه هر نشریه از نظر فراوانی مقالات مرتبط با یک موضوع خاص و ستون دیگر، فراوانی مطلق مقالات مرتبط به موضوع مورد نظر در هر نشریه

1. Newman

است. بنابراین، مقادیر مرتبط با هریک از متغیرهای فرمول ۱ چنین است: f فراوانی مقالات مرتبط با یک موضوع خاص در یک نشریه است، c یک عدد ثابت که معمولاً مقدار آن نزدیک به فراوانی مقالات آن نشریه‌ای است که از همه نشریات دیگر مقالات مرتبط بیشتری داشته، r همان رتبه نشریه از نظر فروانی مقالات مرتبط، و p عددی منفی و توان توزیع است.

در توزیع پرتو، ستون اول رتبه هر شخص از نظر دارایی و ثروت، و ستون دیگر میزان ثروت وی است. بر این اساس، منطبق با فرمول ۱، f نشانگر میزان ثروت هر فرد، c یک عدد ثابت و نزدیک به دارایی ثروتمندترین فرد، r رتبه وی، و p توان توزیع است.

در توزیع لوتکا، ستون اول، رتبه هر نویسنده از نظر تعداد مقالات و ستون دوم، فراوانی مطلق تعداد مقالات وی است. بنابراین، در فرمول ۱، f فراوانی مقالات هر فرد، c یک عدد ثابت نزدیک به فراوانی مقالات پرکارترین نویسنده، r رتبه نویسنده، و p توان توزیع است.

در توزیع زیپف، ستون اول، رتبه هر واژه از لحاظ فراوانی و ستون دیگر، فراوانی مطلق آن واژه در متن است. در این حالت، f فراوانی واژه، r رتبه آن واژه از نظر فراوانی، c یک عدد ثابت نزدیک به پربسامدترین واژه در متن، و p توان توزیع است.

۲. تعداد دسته‌ها

کاری که عملاً برپایه این قوانین صورت می‌گیرد، طبقه‌بندی یا دسته‌بندی موجودیت‌ها براساس توزیع فراوانی آنهاست. در هریک از این قوانین، موجودیت‌ها براساس میزان فراوانی در ویژگی مورد نظر به چند دسته تقسیم می‌شوند: پرتو (۲ دسته)، برادفورد (۳ دسته)، لوتکا (به تعداد حداکثر فراوانی یا همان c در فرمول ۱)، و زیپف (به تعداد واژگان متن).

علاوه بر تعداد دسته‌ها، تفاوت دیگر این قوانین در نحوه دسته‌بندی موجودیت‌هاست که هریک از آنها روش خاصی را به‌کار می‌برند. اما در هر صورت، مبنای این تقسیم‌بندی‌ها، فراوانی است. به این صورت که باید حاصل جمع ستون فراوانی مطلق بر یک عدد تقسیم یا در یک عدد ضرب شود تا تعداد اعضای هر دسته به‌دست آید.

در توزیع پرتو، تعداد افراد دسته اول برابر با آن ردیفی است که مقدار فراوانی تراکمی آن برابر با حاصل جمع فراوانی مطلق، ضرب در عدد ۰/۸ باشد، یعنی تعداد افراد دسته اول برابر است با شماره ردیفی که مقدار فراوانی تراکمی آن برابر با x باشد:

$$X = \text{حاصل جمع ستون فراوانی} \times ۰/۸$$

فرمول ۲. تعیین تعداد اعضای دسته اول در توزیع پرتو

تعداد موجودیت‌های دسته دیگر برابر است با تعداد کل ردیف‌ها، منهای شماره ردیفی که مقدار فراوانی تراکمی آن برابر با x است. مثلاً اگر تعداد کل ردیف‌ها ۱۰۰ و شماره ردیفی که فراوانی تراکمی آن x است، برابر با ۲۰ باشد، تعداد افراد دسته دوم ۸۰ نفر می‌شود (۸۰ = ۲۰ – ۱۰۰). به‌عبارتی، از نظر ثروت و دارایی ۲۰ نفر اول در دسته اول و ۸۰ نفر بعدی در دسته دوم قرار می‌گیرند. البته، این توزیع در حالت‌های دیگر مانند ۳۰ به ۷۰ (۷۰ = ۳۰ – ۱۰۰) یا ۴۰ به ۶۰ (۶۰ = ۴۰ – ۱۰۰) و غیره هم می‌تواند نمود پیدا کند. به هر حال، مجموع باید برابر ۱۰۰ باشد. در هریک از این حالت‌ها برای محاسبه x باید به‌جای ۰/۸ به‌ترتیب، عددهای ۰/۷، ۰/۶ و غیره را قرار داد.

در توزیع برادفورد، تعداد نشریات دسته اول برابر است با شماره ردیفی که فراوانی تراکمی آن تقریباً به اندازه حاصل جمع ستون فراوانی مطلق، ضرب در یک‌سوم باشد. یعنی تعداد نشریات دسته اول برابر است با شماره ردیفی که مقدار فراوانی تراکمی آن برابر با x باشد:

$$X = (۱ \div ۳) \times \text{حاصل جمع ستون فراوانی}$$

فرمول ۳. تعیین تعداد اعضای دسته اول در توزیع برادفورد

اگر شماره ردیفی که فراوانی تراکمی آن دو برابر x است از شماره ردیفی که فراوانی تراکمی آن برابر با x است کم شود، تعداد نشریات دسته دوم به‌دست می‌آید. تعداد نشریات دسته سوم برابر است با تعداد کل ردیف‌ها، منهای شماره ردیفی که مقدار فراوانی تراکمی آن برابر با x۲ است.

در توزیع لوتکا، تعداد دسته‌ها برابر با بزرگ‌ترین عدد موجود در ستون فراوانی مطلق است. هر دسته حاوی نویسندگانی است که فراوانی مقالاتشان شبیه هم است (فقط اولین نویسنده در نظر گرفته می‌شود). به‌طور مثال، اگر بزرگ‌ترین عدد در ستون فراوانی مطلق ۲۰ و فراوانی مقالات فقط یک نفر ۲۰ باشد، تعداد افراد دسته اول یک نفر خواهد بود. قاعدتاً باید با افزایش شماره دسته، تعداد افراد درون آن نیز به‌ترتیب افزایش یابد.

در توزیع زیپف، عملاً تعداد دسته‌ها برابر با تعداد ردیف‌ها یا همان واژه‌های به‌کاررفته در متن است. از این رو، می‌توان گفت که واژه‌ها عملاً دسته‌بندی شده‌اند و نیاز به عمل خاصی برای دسته‌بندی آنها نیست.

۳. نحوه دسته‌بندی

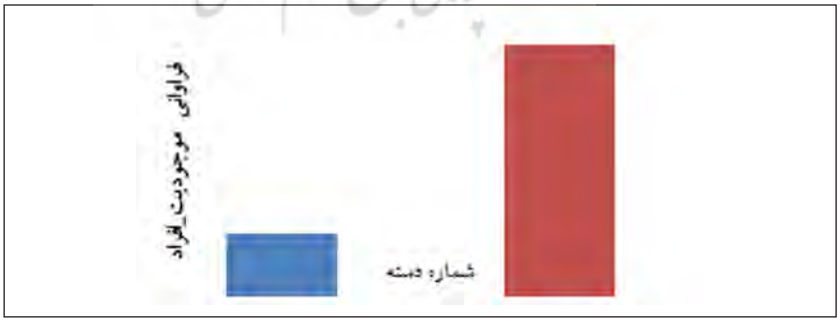
در عمل، آنچه تحت عنوان قوانین چهارگانه یاد می‌شود نحوه دسته‌بندی موجودیت‌ها

و تعیین تعداد موجودیت‌های درون هر دسته است. به‌عبارتی، در هر قانون گفته می‌شود که فراوانی موجودیت‌های هر دسته با دسته‌های قبل از خود چقدر تفاوت دارد. یکی از کارکردهای این قوانین این است که نشان می‌دهند تفاوت بین فراوانی دسته‌ها تابع یک توزیع مشخص است، یعنی اگر شماره دسته را در محور افقی و فراوانی موجودیت‌های هر دسته را در محور عمودی قرار بدهیم، یک توزیع ریاضی به‌دست خواهد آمد.

در توزیع پرتو، فراوانی تعداد موجودیت‌های دسته دوم، چهار برابر فراوانی دسته اول است. به‌عبارت دیگر، فراونی ویژگی‌های دسته دوم یک‌چهارم دسته اول است، یعنی ۲۰ درصد مردم که در دسته اول قرار می‌گیرند ۸۰ درصد ثروت را در اختیار دارند و ۸۰ درصد مردم که در دسته دوم قرار می‌گیرند، ۲۰ درصد ثروت را در اختیار دارند.

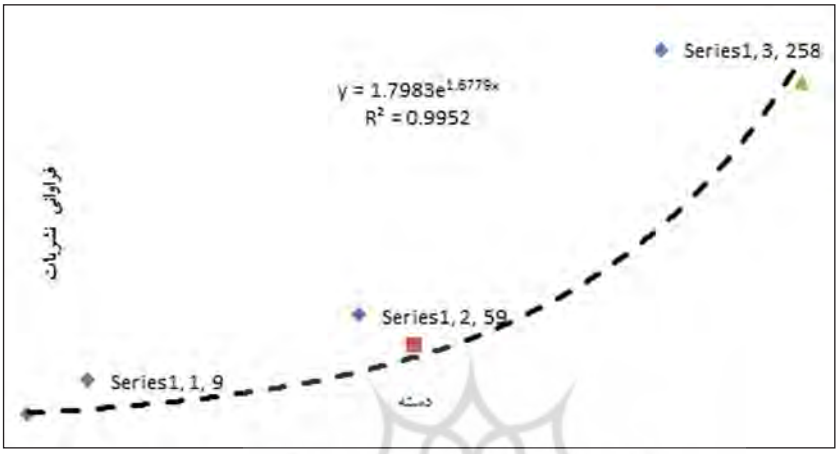
نتیجه این دسته‌بندی را می‌توان به دو شکل نشان داد: الف) فراوانی ویژگی که در نظریه پرتو همان ثروت است؛ و ب) فراوانی موجودیت که در نظریه پرتو همان افراد جامعه است. اگر به‌شکل اول نمایش داده شود، مقدار دسته اول ۸۰ درصد و مقدار دسته دوم ۲۰ درصد کل فراوانی است (نمودار ۳). اما در شکل دوم، مقدار دسته اول ۲۰ درصد و دسته دوم ۸۰ درصد کل فراوانی را دربرمی‌گیرد (نمودار ٤).

نمودار ۳. فراوانی ویژگی‌ها در قانون پرتو



نمودار ٤. فراوانی موجودیت‌ها در قانون پرتو

در توزیع برادفورد، اگر رابطه بین شماره دسته و فراوانی نشریات درون هر دسته محاسبه شود، یک توزیع نمایی حاصل می‌شود. براساس خاصیت توزیع نمایی برای به‌دست آوردن فراوانی نشریات هر دسته باید فراوانی دسته قبل از خود را در عددی ثابت ضرب کرد.



نمودار ۵. توزیع موجودیت‌ها براساس قانون برادفورد

در فرمولی که از تابع نمودار ۵ به‌دست آمده است، y برابر با تعداد نشریات، e همان ۲/۷۱۸۲۸۱ است که لگاریتم طبیعی آن ۱ می‌شود و x شماره دسته است. اگر e را به توانی که بالای آن است برسانیم، عددی حاصل می‌شود که نشان می‌دهد فراوانی هر دسته چند برابر دسته قبل از خود است. این فرایند به توضیح ریاضی لیمکوهلر[1] در سال ۱۹۶۷ شباهت دارد. قبل از آن، برادفورد این مسئله را به‌صورت $n:nc:nc^2$ بیان کرده بود. در این حالت، n فراوانی نشریات دسته اول و c برابر با e به توان عدد بالای آن است که در مورد نمودار ۵ این توان برابر با ۱/۶۷۷۹ است. از این رو، در پژوهش برادفورد برای موضوع ژئوفیزیک، ضریب c برابر با ۲/۷۱۸۲۸۱ به توان ۱/۶۷۷۹ است که عددی مساوی با ۵/۷۲ می‌شود.

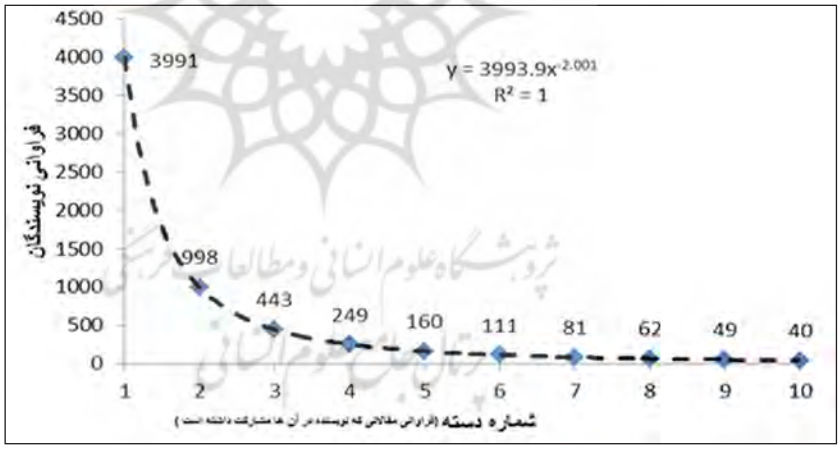
در توزیع لوتکا، شماره‌گذاری دسته‌ها از پایین به بالاست، یعنی دسته آخری شماره ۱، ماقبل آخر ۲ و به‌همین منوال، دسته ابتدایی آخرین شماره را می‌گیرد. ستون اول، همان فراوانی مطلق ویژگی در جدول اولیه (جدول ۱) و ستون دوم، تعداد موجودیت‌هایی است که آن میزان فراوانی را به‌خود اختصاص داده‌اند.

1. Leimkuhler

جدول ۲. توزیع فراوانی موجودیت‌ها در قانون لوتکا

فراوانی مطلق ویژگی (شماره دسته)	فراوانی موجودیت‌های دارای آن فراوانی
۱۰	۴۰
۹	۴۹
۸	۶۲
۷	۸۱
۶	۱۱۱
۵	۱۶۰
۴	۲۴۹
۳	۴۴۳
۲	۹۹۸
۱	۳۹۹۱

اگر رابطه بین دو ستون در جدول ۲ را بسنجیم، یک توزیع توانی حاصل می‌شود که توان آن تقریباً برابر با ۲- است.



نمودار ٦. تابع توزیع قانون لوتکا

همان‌گونه که نمودار ٦ نشان می‌دهد، تابع توزیع لوتکا یک تابع توانی است که در آن y فراوانی نویسندگان و x شماره دسته است. شماره هر دسته نشانگر تعداد مدارکی است که هریک از نویسندگان آن دسته نوشته‌اند. به‌طور مثال، ۳۹۹۱ نویسنده از مجموع کل

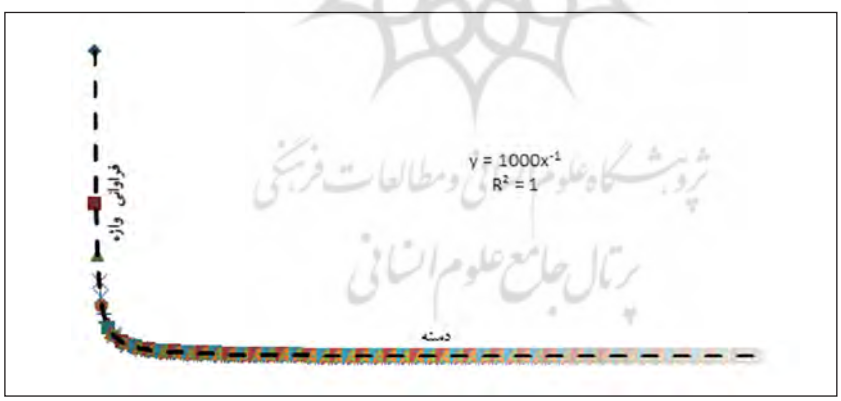
نویسندگان فقط در نوشتن یک مدرک مشارکت داشته‌اند. به‌عبارتی، فراوانی نویسندگان هر دسته برابر با شماره آن دسته به توان ۲- است. اگر بخواهیم این توزیع را به‌زبان برادفورد بیان کنیم، به این صورت خواهد شد:

$$n1^{-2}: n2^{-2}: n3^{-2}: n4^{-2}: n5^{-2}: n6^{-2}: n7^{-2}: n8^{-2}: n9^{-2}: n10^{-2}$$

فرمول ٤. توزیع لوتکا طبق حالت بیان‌شده در توزیع برادفورد

برای تعمیم این نکته می‌توان گفت که فراوانی نویسندگان هر دسته در توزیع لوتکا برابر با nc^p است که در آن n تعداد افراد دسته اول، c شماره دسته، و p توان ثابت است که در این توزیع برابر با ۲- می‌شود.

در توزیع زیپف، عملاً برای دسته‌بندی کاری انجام نمی‌شود و اساس، همان جدول اولیه (مانند جدول ۲) است. به‌همین دلیل، توزیعات سه‌گانه بالا را عملاً شکل تغییریافته‌ای از زیپف می‌دانند، یعنی در همه آنها باید ابتدا مانند زیپف هر موجودیت را از نظر فراوانی از زیاد به کم تنظیم کرد. در این صورت، توزیعی شبیه توزیع توانی حاصل خواهد شد. از آنجا که توزیع زیپف ابتدا برای توزیع فراوانی واژگان یک متن در زبان انگلیسی ارائه شد، آن را می‌توان توزیع فراوانی واژه‌های یک متن دانست. توان این توزیع برای زبان انگلیسی ۱- برآورد شده است.



نمودار ۷. توزیع فراوانی واژه‌های یک متن براساس توزیع زیپف

در نمودار ۷ تعداد واژه‌ها ٤٥٩ مورد فرض شده است که همان تعداد واژگان تشکیل‌دهنده یک متن فرضی است. در ادبیات مربوط به توزیع زیپف به‌جای اصطلاح دسته،

از واژه "رتبه" استفاده می‌شود. در تابع این نمودار y فراوانی واژه و x رتبه آن واژه از نظر فراوانی است. از این رو، می‌توان توزیع زیپف را به‌صورت nc^{-1} بیان کرد. n فراوانی واژه‌ای است که دارای بیشترین فراوانی و c رتبه واژه است. با توجه به اینکه توان ۱- است، می‌توان به‌طور ساده‌تر گفت که توزیع زیپف برابر با n تقسیم بر c است؛ زیرا می‌توان توزیع توانی منفی را به این صورت محاسبه کرد:

$$f = \frac{n}{c^p}$$

در این فرمول، منظور از p همان توان تابع است که چون در مخرج قرار گرفته، علامت آن مثبت شده است.

٤. تعمیم‌پذیری

علم اطلاعات پر از توزیع‌هایی است که "توانی" خوانده می‌شوند، مانند بهره‌وری پدیدآورندگان، دریافت استناد توسط مقالات، توزیع ادبیات علمی، و همکاری در نویسندگی. این موارد جزئی از توزیع‌های تجربی هستند که توزیع توانی دارند (میلوجویچ[1]، ۲۰۱۰). مدت زیادی است که توزیع‌های توانی و نمایی در حوزه‌های دیگر مانند اقتصاد، زیست‌شناسی، و علوم رایانه راه یافته‌اند (میتزن‌ماچر[2]، ۲۰۰٤). بر این اساس، یکی دیگر از کارکردهای قوانین چهارگانه در این است که می‌توان آنها را به موجودیت‌ها و ویژگی‌های مختلف تعمیم داد. گرچه این قوانین بر اثر مطالعه فراوانی موجودیت‌های خاصی از نظر یک ویژگی کشف شده‌اند، این توانایی را دارند تا موجودیت‌های دیگر را براساس توزیع فراوانی در یک ویژگی خاص دسته‌بندی کنند. موجودیتی که پرتو نخستین بار مورد توجه قرار داد، افراد جامعه در ایتالیا بود که از نظر ویژگی ثروت به‌دسته‌بندی آنها پرداخت. برادفورد نشریات را از نظر تعداد مقالات مرتبط به یک موضوع بررسی و آنها را به سه دسته تقسیم کرد. لوتکا نویسندگان را از نظر ویژگی تعداد مقاله و مدارکی که نوشته‌اند دسته‌بندی کرد. زیپف به واژه‌های یک متن از نظر دفعات تکرار در متن پرداخت.

با توجه به ویژگی تعمیم‌پذیری، بسیاری از موجودیت‌های دیگر را می‌توان براساس توزیع پرتو دسته‌بندی کرد. به‌طور مثال، مجموعه کتاب‌های یک کتابخانه را از نظر میزان امانت می‌توان به دو دسته تقسیم کرد: دسته اول متشکل از ۲۰ درصد کتاب‌هاست که توسط ۸۰ درصد مراجعان به امانت رفته‌اند و دسته دوم را ۸۰ درصد بقیه کتاب‌ها را تشکیل می‌دهند که توسط ۲۰ درصد اعضا به امانت رفته‌اند. به‌همین ترتیب، می‌توان وبگاه‌ها را از لحاظ تعداد بازدید و مقالات یک حوزه را از نظر استناد شدن به دو دسته تقسیم کرد. نمونه‌های بیشتری نیز می‌توانیم در اطراف خودمان ببینیم که این توزیع را می‌توان به آنها تعمیم داد. بسیاری از

1. Milojević
2. Mitzenmacher

موجودیت‌های دیگر را می‌توان براساس توزیع‌های دیگر دسته‌بندی کرد. گام اول در این مطالعات، انتخاب یک موجودیت برای مطالعه و تعیین فراوانی یک ویژگی در آنهاست. سپس می‌توان همانند آنچه کاشفان این قوانین انجام داده‌اند و شرح آن در بالا آمد، عمل کرد.

نتیجه‌گیری

در این پژوهش، تصویری کلی از قوانین تجربی معروف در علم‌سنجی برمبنای قاعده توانی ارائه شد. نتایج پژوهش نشان می‌دهد که شکل اولیه داده‌ها در این چهار توزیع، حاوی رتبه هر موجودیت، فراوانی ویژگی مورد مطالعه، و مقدار فراوانی مطلق آن ویژگی در آن موجودیت است. نتایج مرتبط با نحوه دسته‌بندی موجودیت‌ها و تعیین تعداد موجودیت‌های درون هر دسته حاکی از آن است که تفاوت بین فراوانی دسته‌ها تابع یک توزیع مشخص است. ویژگی تعمیم‌پذیری این قوانین، راه را برای مطالعات بیشتر به‌منظور دسته‌بندی موجودیت‌ها براساس توزیع فراوانی باز می‌کند. این نوع مطالعات، عملاً با تطبیق توزیع فراوانی یک موجودیت براساس یک ویژگی خاص می‌کوشند تا امکان دسته‌بندی موجودیت‌های مورد مطالعه را با رویکردهای به‌کاررفته در این قوانین بررسی کنند. به‌عبارتی، می‌خواهند دریابند که آیا توزیع موجودیت مورد مطالعه بر یافته‌های به‌کاررفته در قوانین تجربی کتاب‌سنجی منطبق است یا خیر، و یا چقدر نزدیک است.

در فرایند تطبیق باید به چند نکته توجه کرد: اول اینکه ممکن است تعداد دسته‌ها براساس یکی از این قوانین باشد؛ اما فاصله بین فراوانی دسته‌ها از توزیع مورد نظر تبعیت نکند. مثلاً ممکن است بتوان نویسندگان را از نظر ویژگی فراوانی دریافت استناد به سه دسته تقسیم کرد که فراوانی تعداد استنادهای دریافتی هر دسته، همانند توزیع برادفورد، یک‌سوم کل استنادهای دریافتی باشد؛ اما توزیع تعداد نویسندگان در این دسته‌ها به‌صورت نمایی نباشد و از توزیع توانی پیروی کند. حتی ممکن است بتوان تعداد دسته‌ها را در توزیع برادفورد افزایش داد. در مواردی حتی ممکن است بتوان دسته‌بندی را براساس بیش از یک توزیع انجام داد، مثلاً یک توزیع همزمان بر دو توزیع برادفورد و زیپف منطبق باشد. این حالت از آنجا پیش می‌آید که در اساس، توزیع فراوانی ویژگی‌های یک موجودیت تا حدودی از یک توزیع توانی پیروی می‌کند. نتایج حاصل همگام با نتایج ویلسون (۱۹۹۹) در ARIST است که به موضوع اطلاع‌سنجی پرداخته و نشان داده است که بین سه قانون زیپف، لوتکا، و برادفورد ارتباط وجود دارد.

مآخذ

Ausloos, M. (2014). Zipf–Mandelbrot–Pareto model for co-authorship popularity. *Scientometrics*, *101* (3), 1565-1586.

Axtell, R. L. (2001). Zipf distribution of US firm sizes. *Science*, 293 (5536), 1818-1820.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering: an Illustrated Weekly Journal (London)*, *137* (3550), 85–86.

Bradford, S. C. (1985) Sources of information on specific subjects. *Journal of Information Science, 10* (4), 173–180

Broadus, R. N. (1987). Toward a definition of 'bibliometrics'. *Scientometrics*, 12 (5-6), 373–379.

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM Review*, *51* (4), 661-703.

Drott, M. C. (1981). Bradford's Law: Theory, empiricism and the gaps between. *Library Trends*, *30* (1), 41-52.

Franceschet, M. (2008). Frozen footprints. *arXiv preprint arXiv:0811.4603*. Retrieved Sep. 05, 2015, from http://arxiv.org/abs/0811.4603

Gelbukh, A., & Sidorov, G. (2001). Zipf and Heaps Laws' coefficients depend on language. *Computational Linguistics and Intelligent Text Processing*. Berlin Heidelberg: Springer.

Harremoës, P., & Topsoe, F. (2005). Zipf's law, hyperbolic distributions and entropy loss. *Electronic Notes in Discrete Mathematics*, *21* (1), 315-318.

Hertzel, D. H. (1987). History of the development of ideas in bibliometrics. Kent, A. (Ed.), *Encyclopedia of library and information sciences* (Vol. 42, pp. 144–219). New York: Marcel Dekker.

Hood, W. W., & Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, *52* (2), 291-314.

Leimkuhler, F. F. (1967). The Bradford distribution. *Journal of documentation*, *23* (3), 197-207.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16* (12), 317–323.

Malacarne, L. C., Mendes, R. S., & Lenzi, E. K. (2002). Q-exponential distribution in urban agglomeration. *Physical Review*, *65* (1), 17-26.

Mayr, P. (2013). Relevance distributions across Bradford Zones: Can Bradfordizing improve search? Retrieved Nov. 14, 2015, from http://arxiv.org/abs/1305.0357

MilojeviĆ, S. (2010). Power law distributions in information science: Making the case for logarithmic binning. *Journal of the American Society for Information Science and Technology*, *61* (12), 2417-2425.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, *1* (2), 226-251.

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46* (5), 323-351.

O'Connor, D. O., & Voos, H. (1981). Empirical laws, theory construction and bibliometrics. *Library Trends*, *30* (1), 9-20.

Pareto, V. (1964). *Cours d'Économie Politique:* Nouvelle édition par G.-H. Bousquet et G. Busino. Geneve: Librairie Droz.

Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation*, *25* (4), 348-349.

Van Raan, A. F. (2001). Two-step competition process leads to quasi power-law income distributions: Application to scientific publication and citation distributions. *Physica A: Statistical Mechanics and its Applications*, *298* (3), 530-536.

Wilson, C. S. (1999). *Informetrics. Annual Review of Information Science and Technology (ARIST), 34* (1), 107-247.

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge (Mass): Harvard University Press.

استناد به این مقاله:

توکلی‌زاده‌راوری، محمد؛ سهیلی، فرامرز (۱۳۹۵). ویژگی‌های مشترک قوانین تجربی معروف در علم‌سنجی: نگاهی از زاویه دسته‌بندی داده‌ها براساس توزیع فراوانی. *مطالعات ملی کتابداری و سازماندهی اطلاعات*، *۲۷* (۱)، ۲۵-۴۲.