الإحصاء الوصفي

سينت فيسر و جون بول جونز الثالث

الفصل 17

ترجمة بتصرف
أ.د. مضر خليل عمر

مقدمة

لا يُعلّمك هذا الفصل والفصل التالي كيفية إجراء التحليلات الإحصائية للبيانات . كما هو الحال في نظم المعلومات الجغرافية والاستشعار عن بُعد ، تُقدّم معظم أقسام الجغرافيا دوراتٍ دراسيةً في الإحصاء تمتدّ لفصل دراسي أو حتى عام دراسي . ومع ذلك ، نأمل أن نُعرّفك على التفكير الإحصائي وأن نُحسّن من قدرتك على فهم النتائج الإحصائية للباحثين الآخرين . تفترض الفصول أن لديك فهمًا للجبر ورموزه (مثل علامة الجمع) ، وأنك تستطيع قراءة الرسوم البيانية الديكارتية ، وأنك تفهم معنى المتغيرات والعلاقات ، وأن لديك فهمًا جيدًا لمستويات القياس (ينظر الفصل 4). **الإحصاء هو رقم يُلخّص متغيرًا أو علاقةً بين متغيرات بطريقةٍ ما . تصف هذه الأرقام توزيع قيم متغير، وطبيعة ودرجة العلاقات بين المتغيرات**. ولأن تكثيف البيانات في رقم أو معادلة واحدة يفقد الكثير من المعلومات ، ينبغي استخدام الجداول والرسوم البيانية لتتمكن من تصور التوزيعات والعلاقات ، خاصةً وأن البرامج الإحصائية قادرة على إنشائها بسهولة . تكمن المهمة الأساسية للإحصاء في كشف العلاقات بين المتغيرات ووصفها.

يُعد وصف المتوسطات والتباين والعلاقات البسيطة - وهي مواضيع هذا الفصل - وسيلة لتحقيق هذه الغاية . غالبًا ما تُوصف العلاقة بأنها الفرق بين متوسطات متغير لمجموعتين فرعيتين من الملاحظات ضمن العينة ؛ على سبيل المثال ، متوسط دخل الرجال والنساء في المهن نفسها . يُعد التباين ذا صلة لأننا نصف العلاقات بأنها تباين مشترك بين المتغيرات ، ولأن طبيعة التباين داخل كل متغير تؤثر على ما إذا كان التباين المشترك بين متغيرين ذا معنى . يمكن أيضًا تصنيف الإحصاءات إلى بارامترية وغير بارامترية (نوركليف 1982). تشير الإحصاءات البارامترية إلى تلك التقنيات التي تنطبق فقط على المتغيرات المترية أو متغيرات الفترات/النسب . تنطبق التقنيات غير البارامترية على المتغيرات الترتيبية والاسمية . تُعد الإحصاءات البارامترية أكثر "قوة" بمعنى أنها تعتمد على معرفة مقدار الفرق بين قيم المتغير .

لا تأخذ الإحصاءات غير البارامترية في الحسبان سوى ترتيب الملاحظات على متغير، أو الفئة التي تندرج فيها الملاحظة . في بعض الأحيان ، تُستخدم الإحصاءات البارامترية على البيانات الترتيبية ، ولكن النتائج تكون غير مرضية. هناك تصنيف آخر للإحصاءات وهو بسيط ومتعدد المتغيرات . تشمل الإحصاءات متعددة المتغيرات التقنيات التي تحاول الكشف عن العلاقات المستقلة الموجودة بين أكثر من متغيرين (جونستون 1978) . تنظر الإحصاءات البسيطة إلى الترابط بين متغيرين فقط في كل مرة ، وتتجاهل تأثير المتغيرات الأخرى على توزيع قيم هذين المتغيرين .

التوزيعات التكرارية

يمكن استخدام جداول ورسوم بيانية للتوزيعات التكرارية لمعرفة كيفية توزيع قيم متغير ما بسرعة. يجب تجميع قيم المتغير للقيام بذلك ، مع أن وحدة القياس في حالة المتغير الاسمي هي مجموعة أو فئة . يشير التكرار إلى عدد المشاهدات في مجموعة . في الجدول 17.1، النوع الوظيفي هو متغير تم توليده من خلال

تحليل هيكل التوظيف في المدن الأمريكية عام 1950 باستخدام تقنية تُسمى تحليل المجموعات . تُجمّع هذه التقنية المدن وفقًا لأوجه التشابه في هيكل التوظيف فيها . العمود المسمى "f" (التكرار) هو عدد المدن في كل فئة من فئات المتغير . التكرار النسبي (f%) هو النسبة المئوية أو النسبة المئوية لإجمالي عدد المشاهدات في كل مجموعة . (النسبة هي رقم معبر عنه كنسبة مئوية من الرقم 1؛ 13.10% معبر عنها كنسبة مئوية هي 0.1310).

Table 17.1 Functional type of US SMSAs, 1950

Functional type	*f*	*% f*
Diversified	19	13.10
Diversified/manufacturing	44	30.34
Federal government	1	0.69
Manufacturing	12	8.28
Mining	3	2.07
Other government	7	4.83
Public admin./regional centers	6	4.14
Transportation/wholesale/finance	25	17.24
Service centers	7	4.83
Transportation	1	0.69
Professional service	20	13.79
Number of Cases/Sum	145	100.00

Notes: SMSAs are Standard Metropolitan Statistical Areas. This designation was later replaced by MSAs, Metropolitan Statistical Areas.
Source: data from Subramaniam-Bryson (1991)

يُفضل عرض كل من التكرارات والتكرارات النسبية في هذا النوع من الجداول لتسهيل المقارنة مع أنواع مماثلة من الجداول . على سبيل المثال ، يمكن مقارنة هذا الجدول بجداول لعقود أخرى يختلف فيها عدد المدن المصنفة عن عدد المدن في عام 1950. تُنظَّم الجداول التي تُظهر توزيعات التكرار للمتغيرات الترتيبية بنفس طريقة تنظيم جداول المتغيرات الاسمية ، باستثناء أن فئات المتغير الترتيبي تُرتَّب حسب رتبها ، من الأدنى إلى الأعلى أو من الأعلى إلى الأدنى . في حالة المتغيرات الاسمية ، يُمكن ترتيب الفئات بأي ترتيب (ينظر الفصل 4).

إذا كان المتغير الترتيبي لا يحتوي على العديد من الملاحظات المرتبطة في كل رتبة ، فيجب إعداد جدول المتغير كما هو الحال بالنسبة لمتغير النسبة . يجب تجميع الرتب . لإنشاء جداول للمتغيرات التي تُقاس على مقياس فاصل أو نسبة ، يجب تجميع قيم المتغيرات أو تصنيفها بحيث تقع عدة ملاحظات في مجموعة واحدة . وينطبق هذا أيضًا على المتغير الترتيبي الذي تكون فيه معظم الملاحظات ذات رتبة فريدة . يوضح الجدول 17.2 التوزيع التكراري لمتغير"VTO" ، وهو عدد الأشخاص الذين صوتوا في الانتخابات الرئاسية

لعام 1980 كنسبة مئوية من إجمالي السكان الذين تبلغ أعمارهم 18 عامًا فأكثر في جميع مقاطعات تكساس عام 1980. يوجد 254 مقاطعة ، ولذلك تم تجميع المتغير في 14 فئة (مجموعات) متنافية .

Table 17.2 Voter turnout, Texas counties, 1980 presidential election

VTO	*f*	*% f*	*F*	*% F*
22.5–27.5	2	0.79	2	0.79
27.5–32.5	3	1.18	5	1.97
32.5–37.5	4	1.57	9	3.54
37.5–42.5	14	5.51	23	9.06
42.5–47.5	55	21.65	78	30.71
47.5–52.5	85	33.46	163	64.17
52.5–57.5	31	12.20	194	76.38
57.5–62.5	28	11.02	222	87.40
62.5–67.5	22	8.66	244	96.06
67.5–72.5	6	2.36	250	98.43
72.5–77.5	0	0.0	250	98.43
77.5–82.5	2	0.79	252	99.21
82.5–87.5	1	0.39	253	99.61
87.5–92.5	1	0.39	254	100.00
Number of cases/sum	254	99.97		

Source: Data from Tebben (1990)

مع انخفاض عدد الملاحظات في الدراسة ، يجب أن يكون عدد الفئات أصغر. يبلغ عرض الفئة المختار لـ .5% VTO بشكل عام ، يجب أن تكون الفئات بنفس العرض ؛ وإلا ، فسيكون توزيع قيم المتغير مشوهًا . ومع ذلك ، هناك حالات قد ترغب فيها في الحصول على فئات بعرض مختلف (على سبيل المثال، تجميع السكان بهذه الطريقة : الأطفال ، 0-17 عامًا ؛ البالغون ، 18-64 عامًا ؛ كبار السن ، 65 عامًا فأكثر). يُطلق على العمود المسمى "f" **اسم التوزيع التكراري** . يوضح كيفية توزيع قيم .VTO هل هناك تركيز للقيم ؟ حول أي قيمة تتركز؟ هل توزيع القيم متماثل ؟ أي ، هل هناك أعداد متساوية تقريبًا من القيم العالية جدًا والمنخفضة جدًا ؟

العمود المسمى "f%" هو التوزيع التكراري النسبي . وهو التكرار كنسبة مئوية من إجمالي عدد المشاهدات (n) . يشير الرمز "F" إلى التوزيع التكراري التراكمي . وهو إجمالي عدد المشاهدات التي تحتوي على قيم أصغر من الحد الأعلى للفئة . يُوفر التكرار التراكمي النسبة المئوية للحد الأعلى للفئة . النسبة المئوية k لمتغير ما هي قيمة المتغير بحيث تكون قيم k% على الأقل من المقاطعات أصغر منها أو مساوية لها . 67.5 هي النسبة المئوية 96.06 لـVTO ، وبالتالي فإن 3.94% فقط من الحالات لها قيم أكبر من 67.5 . تنطبق النسب المئوية أيضًا على المتغيرات الترتيبية ، ولكن ليس لها معنى في حالة المتغيرات الاسمية فقط.

الرسوم البيانية

على الرغم من أنه قد يكون بإمكانك تخيل توزيع VTO من الجدول 17.2 ، إلا أن الرسوم البيانية أفضل من الجداول في توفير انطباع بصري فوري عن توزيع المتغير. يمكن تمثيل التوزيع التكراري للمتغير الاسمي من خلال مخطط دائري (رسم بياني دائري) أو مخطط شريطي . في المخطط الشريطي ، يكون التكرار أو النسبة المئوية على المحور الصادي ، بينما تكون فئات المتغير على المحور السيني . لكل فئة

شريط يتناسب طوله مع تكرار الفئة . الأشرطة متساوية العرض وليست متجاورة لعدم وجود استمرارية في القيم بين فئات المتغير . المدرج التكراري هو مخطط شريطي يُستخدم لتصوير توزيع متغير مُقاس على مستوى الفاصل أو النسبة . الأشرطة متجاورة لإظهار استمرارية قيم المتغير. يمثل كل شريط تكرار فئة من متغير مُجمّع . الأشرطة متساوية العرض لأن جميع الفئات لها نفس العرض.

يوضح الشكل 17.1 مدرجًا تكراريًا لبيانات إقبال الناخبين في الجدول 17.2 ، باستثناء أن عرض الفئات قد خُفِّض إلى 2.5% فقط . مساحة الشريط في المدرج التكراري تتناسب طرديًا مع تكرار الفئة . أي أن مساحة الهيستوغرام بين أي قيمتين لـ VTO تُحسب كنسبة مئوية من المساحة الكلية للهيستوغرام . وهذا يوفر النسبة المئوية لجميع الملاحظات التي لها قيم ضمن تلك الفئة . هذا صحيح فقط إذا كانت جميع عروض الفئات متساوية . يمكن أيضًا تمثيل التوزيع التكراري كرسم بياني خطي يُسمى مضلعًا تكراريًا . يوضح الشكل 17.2 مضلعًا تكراريًا لبيانات إقبال الناخبين باستخدام الفئات نفسها كما في الجدول 17.2.

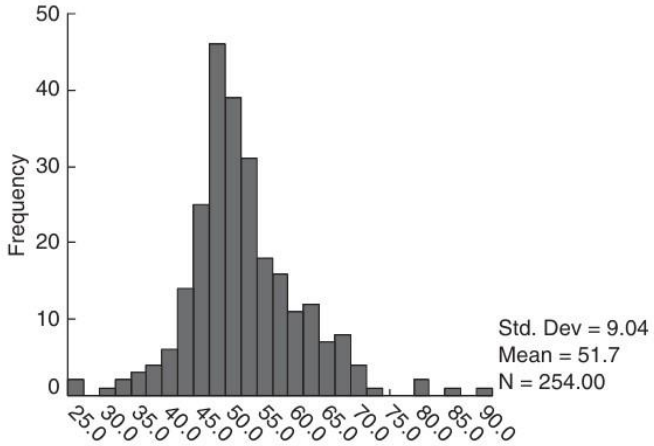

Figure 17.1 Histogram for voter turnout (VTO) in Texas counties

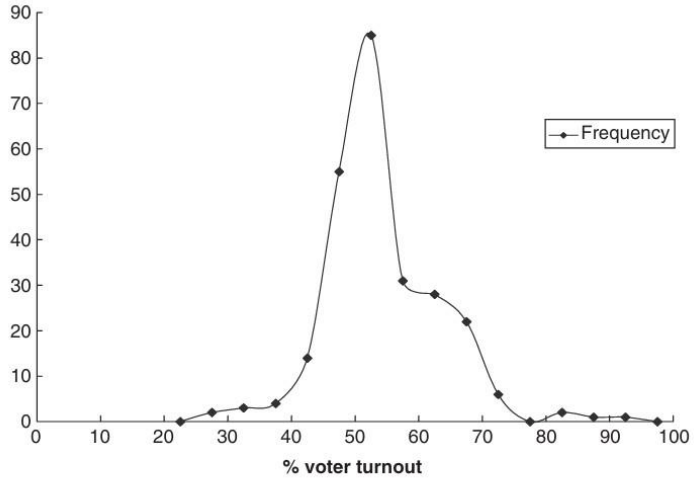

Figure 17.2 Frequency polygon for voter turnout in Texas counties

تُرسم قيمة التكرار للفئة مقابل نقطة منتصفها ، ثم تُربط الرسوم البيانية بخط . يعادل رسم مضلع تكراري وصل نقاط منتصف قمم أعمدة الرسم البياني ، بشرط إضافة فئات فارغة عند طرفي التوزيع . يُرسم التردد الصفري في كل فئة من هذه الفئات بحيث يُنزل المضلع إلى المحور السيني ويُحاط . هذا يجعل مساحة المضلع متساوية . إلى مساحة الهستوغرام ، والمساحة أسفل المنحنى بين قيمتين للمتغير، كنسبة مئوية من إجمالي المساحة أسفل المنحنى ، تساوي التكرار النسبي للملاحظات بين هاتين القيمتين المتغيرتين . تُسمى الرسوم البيانية الخطية للتوزيعات التكرارية التراكمية F % في الجدول 17.2 بالمنحنيات الغاوسية . تميل هذه الرسوم البيانية من أسفل اليسار إلى أعلى اليمين عادةً على شكل حرف S ، وتُظهر العدد الإجمالي أو النسبة المئوية للملاحظات التي تكون أقل من أو تساوي قيمة معينة.

شكل مضلع التكرار

تميل التوزيعات التكرارية إلى أن يكون لها تركيز قيم قريب من متوسط المتغير. وهذا يميل إلى تكوين شكل الحدبة للتوزيع كما هو موضح في الشكل 17.2. يُوصف شكل التوزيع بالمقارنة مع التوزيع الطبيعي ، والذي يُسمى أيضًا منحنى الجرس أو التوزيع الغاوسي . هذا التوزيع متماثل ، أي أن الجانب الأيسر من الرسم البياني هو صورة طبق الأصل للجانب الأيمن . تختلف قيم المتغير التي تقل عن "المتوسط" بنفس طريقة اختلاف القيم التي تزيد عنه . التوزيع المنتظم متماثل أيضًا . وهو توزيع تحدث فيه جميع قيم المتغير بتردد متساوٍ، ويكون الرسم البياني خطًا مستقيمًا أفقيًا .

إذا تم رمي حجر نرد واحد بشكل متكرر، وسُجلت القيم التي تم رميها ، فيجب أن تكون ترددات كل رقم متساوية تقريبًا . أي أن الأرقام التي تم رميها سيكون لها توزيع منتظم . هذا ما يعتقده معظم الناس على أنه عشوائي ؛ أي فرصة متساوية للوقوع . في عينة عشوائية من الملاحظات ، يكون لكل ملاحظة فرصة متساوية للاختيار، ولكن احتمال الحصول على قيم معينة لمتغير تم قياسها على تلك الملاحظات يعتمد على شكل التوزيع التكراري لذلك المتغير . يعني مصطلح الانحراف أن توزيع القيم ليس متماثلًا. الانحراف الموجب أو الأيمن يعني أن معظم الملاحظات تتجمع حول قيمة أصغر من متوسط جميع القيم . هناك العديد من القيم الصغيرة ، وبعضها كبير جدًا . مضلع التردد له ذيل إلى اليمين . الانحراف السالب أو الأيسر هو عكس الانحراف الأيمن . هناك مجموعة من القيم الكبيرة مع بعض القيم الصغيرة جدًا . الانحراف الأيمن أكثر شيوعًا في متغيرات العلوم الاجتماعية من الانحراف الأيسر.

يشير التفرطح إلى درجة تسطح التوزيع التكراري . يتميز التوزيع اللبتوكورتي بدرجة عالية من تجميع الملاحظات حول قيمة مركزية مع درجة كبيرة من التباين بين عدد قليل من القيم الصغيرة و/أو عدد قليل من القيم الكبيرة . يشير التوزيع المسطح إلى درجة منخفضة من تجميع القيم . التوزيع المنتظم مسطح جدًا . التوزيع الطبيعي متوسط التفرطح . هو المعيار الذي تُعد به التوزيعات الأخرى توزيعات ليبتوكورتيكية أو توزيعات بلاتيكورتيكية . التوزيع الموضح في الشكل 17.2 ليس موحدًا ولا شديد الانحراف . يميل إلى شكل طبيعي ، ولكنه ليس متماثلًا . إنه منحرف قليلًا إلى اليمين وليبتوكورتيكي.

إحصاءات لوصف التوزيعات التكرارية

الميل المركزي

تلخص المتوسطات الموقع المركزي للتوزيع التكراري . وتُسمى مقاييس النزعة المركزية لأنها تميل إلى التواجد بالقرب من مركز توزيع المتغير. وتميل إلى أن تكون القيمة التي تتجمع حولها معظم الملاحظات . ولذلك ، يُشار إلى المتوسطات أيضًا باسم "المعيار" . المنوال هو القيمة الأكثر شيوعًا للمتغير . وهو مفيد فقط للمتغيرات الاسمية أو الترتيبية ، لأن العديد من متغيرات النسبة إما لا تحتوي على منوال ، أو تحتوي على عدة منوال . إذا رُتبت قيم المتغير حسب الحجم (بما في ذلك الملاحظات التي لها قيمة متغير صفر) ، فإن الوسيط هو قيمة متغير الملاحظة الوسطى . وبالتالي ، يكون الوسيط أيضًا النسبة المئوية الخمسين . لا يمكن حساب الوسيط للبيانات الاسمية ، ولكنه مناسب للمتغيرات الترتيبية .

يبلغ وسيط .49.86 VTO من بين المقاطعات ، قيمة VTO في 50% منها أقل من أو تساوي 49.86، و50% منها أعلى من تلك القيمة . لا تقيس النسب المئوية عمومًا النزعة المركزية ، لكنها تُوفر مقياسًا للموقع في توزيع تكراري . تُقاس درجات اختبار SAT بطريقة لا يُمكن معها عد المتغير مقياسًا فاصلًا / نسبة . إذًا ، تُوضح نسبتك المئوية المعنى الحقيقي لدرجة 1100 من 1600. إذا كانت النسبة المئوية الثمانون هي 1100، فإن 80% من الذين خضعوا للاختبار حصلوا على درجات أقل من أو تساوي 1100، بينما حصل 20% منهم على درجات أعلى . تُسمى النسب المئوية أيضًا : العُشر الأول (النسبة المئوية العاشرة) ، والربيع الأول (النسبة المئوية الخامسة والعشرون ، أو أدنى 25%)، والربيع الثالث (النسبة المئوية الخامسة والسبعون). الربيع الأول لـ VTO هو 46.63 والربيع الثالث هو 56.07. الربع الثالث أبعد عن الوسيط من الربع الأول ، مما يدل على انحراف المتغير لليمين.

يُطلق على المتوسط الحسابي، كما هو شائع ، اسم المتوسط الحسابي، أو المتوسط الحسابي فقط. رمز متوسط المتغيرX ، المحسوب من بيانات العينة ، هو $\bar{X}$. $\bar{X}$ رمز المتوسط الحسابي الذي جُمعت منه بيانات المتغير X لكامل المجتمع (ويُسمى مُعامل المجتمع) هو μ_X يُنطق"mu" . يُحسب المتوسط الحسابي بجمع جميع قيم المتغير وقسمتها على عدد المشاهدات . ينطبق المتوسط الحسابي على بيانات الفترات / النسب، على الرغم من أن حساب رتبة المتوسط (متغير ترتيبي) شائع ، على سبيل المثال، المعدل التراكمي . غالبًا ما يكون الوسيط والمتوسط الحسابي مختلفين تمامًا في القيمة . إذا كان المتغير ملتويًا لليمين ، فسيكون المتوسط الحسابي أكبر من الوسيط . وسيكون المتوسط الحسابي أصغر من الوسيط الحسابي إذا كانت البيانات ملتوية لليسار .

يرجع ذلك إلى أن المتوسط الحسابي يتأثر بشكل كبير بالقيم الكبيرة أو الصغيرة جدًا ، بينما لا يتأثر الوسيط . إذا كان التوزيع متماثلًا ، فسيكونان متساويين . في البيانات الملتوية ، قد يكون الوسيط مقياسًا مفضلًا إذا كان القصد من "المتوسط الحسابي" وصف قيمة "نموذجية" ، ولكن بخلاف ذلك ، يكون المتوسط الحسابي هو المقياس المفضل لـ "المتوسط الحسابي" لأنه يحتوي على معلومات أكثر (كل قيمة للمتغير تساهم في قيمة المتوسط الحسابي) ، ولأنه مُعرّف رياضيًا ، فإن تباينه بين العينات المأخوذة من مجتمع واحد معروف . يبلغ متوسط 51.67 VTO، وهو أكبر من الوسيط ، مما يدل على الانحراف الأيمن في المتغير. إذا سُحبت عدة عينات مختلفة بنفس الحجم من مجتمع واحد ، فستختلف قيم متوسطات العينات المختلفة ، وكذلك متوسطات العينات المختلفة . إذا كان توزيع متغير المجتمع متماثلًا إلى حد ما ، فستختلف متوسطات العينات بدرجة أقل.

حول متوسطها السكاني الحقيقي ، ستكون متوسطات العينة حول متوسطها السكاني الحقيقي أكثر موثوقية من متوسطات العينة حول متوسطها السكاني الحقيقي . علاوة على ذلك ، فإن درجة تباين هذه المتوسطات معروفة . وبالتالي ، يُعد متوسط العينة مُقدِّرًا أكثر موثوقية لمتوسط المجتمع من متوسط العينة لمتوسط المجتمع.

التباين، الانتشار، أو التشتت

يُقدم المتوسط معلومات حول الاتجاه المركزي للتوزيع ، ولكنه لا يُخبرنا شيئًا عن مقدار التباين في متغير ما . ولذلك ، نحتاج إلى إحصاءات تصف التباين . تعتمد قيمة معظم مقاييس التباين على حجم أعداد المتغير، أي على وحدة قياس المتغير . وبالتالي ، فإن مقارنة تباين متغيرين بوحدات قياس مختلفة تمامًا تكون غير مُجدية ما لم تُوحَّد مقاييس التباين هذه بطريقة ما بحيث يتم التعبير عن التباينات بالنسبة إلى متوسط حجم الأعداد. أحد الاستخدامات المهمة لمقاييس التباين هو الاستدلال الإحصائ ي. إذا سُحبت عدة عينات بحجم nمن مجتمع واحد ، فإن متوسطات تلك العينات ستختلف بشكل أكبر مع زيادة التباين في المتغير . وبالتالي ، تؤثر درجة التباين في قيم المتغير على مدى موثوقية إحصائية عينة واحدة مثل متوسط العينة، $\bar{X}$ كمقدّر لمعامل المجتمع.(μ_X)

يُعد المدى أبسط مقياس للتباين . ويُحسب بأخذ الفرق بين أصغر وأكبر قيمتين للمتغير . أصغر قيمة لـ VTO هي 23.92، وأكبرها هي 91.14. الفرق البالغ 67.22 هو المدى . يعتمد المدى كليًا على القيمتين المتطرفتين للمتغير . لذا ، على الرغم من بساطته ، إلا أنه غير مفيد نسبيًا . المدى بين الربعين هو المدى بين النسبتين المئويتين 25 و75 56.07 – 46.63 = 9.44 لـVTO . الانحراف العشري هو المدى بين النسب المئوية العاشرة والتسعين . تُزيل هذه المقاييس تأثير القيم المتطرفة ، ولكن هناك مقاييس أكثر فائدة للتباين في متغير ما، والتي تستخدم المعلومات الموجودة في جميع قيمه .

الانحراف هو المسافة بين قيمة المتغير والإحصاء ، وعادةً ما يكون المتوسط . يُكتب جبريًا على النحو التالي$x = X - \bar{X}$: للمتغير.$\bar{X}$ القيم التي تقل عن المتوسط ستكون لها قيم سالبة . يُعد متوسط المسافة بين كل قيمة ومتوسطها على جانبي المتوسط مقياسًا بديهيًا للتباين . **ولكن مجموع الانحرافات حول المتوسط يساوي صفرًا** . ويمكن جعل الانحرافات حول المتوسط موجبة بتربيعها . يُسمى متوسط (أو متوسط) الانحراف التربيعي حول المتوسط بالتباين . يُسمى مجموع الانحرافات التربيعية بالتباين الكلي للمتغير أو مجموع المربعات . الصيغة الأولى أدناه مخصصة لتباين العينة (s^2)، والصيغة الثانية مخصصة لتباين المجتمع (σ^2). يُفضل القسمة على n – ... تباين VTO هو 81.79. مقياسه هو النسبة المئوية التربيعية .

لتحويل التباين إلى مقياس مماثل للمتغير الأصلي ، نستخدم الانحراف المعياري ، وهو الجذر التربيعي للتباين . رموزه هي s لإحصاء العينة ، و σ(سيجما) لتباين المجتمع . وهو عادةً أصغر بكثير من التباين ، إلا إذا كان التباين أقل من واحد . الانحراف المعياري لـ VTO هو 9.04 (نسبة مئوية) . ينطبق الانحراف المعياري فقط على بيانات الفترات/النسب ، وهو المقياس الأكثر شيوعًا للتباين لهذه البيانات ، على الرغم من صعوبة فهم معناه البديهي . قيمته حساسة للقيم المتطرفة في البيانات بسبب تربيع الانحرافات . ومع ذلك ، لدينا فكرة جيدة عن عدد قيم المتغير التي يجب أن تكون ضمن عدد معين من الانحرافات المعيارية عن المتوسط (انظر الملحق 17.1) .

معامل التباين

غالبًا ما يكون من المستحسن مقارنة التباين في متغيرين لهما نفس وحدات القياس ولكن بمتوسطات مختلفة تمامًا . على سبيل المثال ، قد نرغب في معرفة ما إذا كانت تفاوتات الدخل ، كما تُقاس بالانحراف المعياري لدخل الفرد في مقاطعة ما ، تتناقص أو تتزايد بمرور الوقت . ومع ذلك ، على مدى فترة زمنية تبلغ 30 عامًا ، زاد دخل الفرد بمقدار 3 أضعاف في المتوسط ، وبالتالي يجب أن يكون الانحراف المعياري.

نقاطZ

تعبر الدرجات المعيارية أو درجات Z عن قيمة متغير ما بعدد وحدات الانحراف المعياري التي تكون القيمة فيها من المتوسط . ينص هذا على أن القيمة تساوي المتوسط الحسابي زائد z انحرافات معيارية . إذا كانت القيمة أقل من المتوسط الحسابي ، تكون z سالبة ، وتساوي القيمة المتوسط الحسابي ناقص z انحرافات معيارية . على سبيل المثال ، تعني قيمة – 1.5 لـ z أن قيمة X التي حُسبت لها أقل بمقدار 1.5 انحراف معياري من متوسطها . يسمح تحويل قيم متغير إلى درجات z بمقارنة المتغيرات التي لا تختلف في متوسطها الحسابي فحسب ، بل تختلف أيضًا في وحدات قياسها . يُشار أحيانًا إلى درجات z بالدرجات المعيارية ، ولكن هذا خطأ في التسمي ة، لأنه إذا لم يكن للمتغير توزيع طبيعي ، فإن درجات z لا تتبع التوزيع الطبيعي أيضًا

أصغر قيمة لـ VTO هي 23.92. (23.92 – 51.67)/9.044 يساوي – 3.068، وهو الدرجة المعيارية للقيمة 23.92. وكما ذُكر سابقًا ، فإن ثلاثة انحرافات معيارية تحت المتوسط هي 51.67 – 3(9.044) ، أي ما يعادل 24.54. 23.92 هي القيمة الوحيدة التي تقل عن ثلاثة انحرافات معيارية تحت المتوسط . تُحسب ثلاثة انحرافات معيارية فوق المتوسط على أنها 51.67 + 3(9.044)، أي ما يعادل 78.80. أكبر قيمة لـ 91.14 هي (91.14 – 51.67)/9.044، أي ما يعادل الدرجة المعيارية 4.364. إذن ، فهي 4.364 انحراف معياري فوق المتوسط ، وهي واحدة من أربع قيم تزيد عن ثلاثة انحرافات معيارية فوق المتوسط .

يوضح النقاش أعلاه أنه يمكن استخدام درجات Z لتحديد ما إذا كان توزيع التكرار لمتغير ما يقارب توزيع التوزيع الطبيعي . درجاتZ ، المحسوبة لمتغير له توزيع طبيعي ، لها متوسط يساوي 0 وانحراف معياري يساوي 1 ، ومضلع التكرار لمنحنيات z هذه له مساحة إجمالية تساوي 1. أشار نقاش مضلعات التكرار إلى أن المساحة بين قيمتين لـX ، معبرًا عنها كنسبة إجمالية من المساحة تحت مضلع التكرار، تساوي نسبة جميع قيم X الواقعة بين هاتين القيمتين . المساحة بين أي درجتين Z في توزيع طبيعي بمتوسط يساوي صفر وانحراف معياري يساوي واحدًا ، تساوي نسبة جميع درجات Z الواقعة بين هاتين القيمتين . تُقدَّم المساحات، الإحصاء الوصفي وبالتالي النسب، بين أي قيمة z ومتوسط القيمة 0 في جميع كتب الإحصاء التمهيدية . كما تُجدول هذه النسب (المساحات تحت المنحنيات) لتوزيعات تكرارية مهمة أخرى ، بما في ذلك توزيع t ، وتوزيع F، وتوزيع ثنائي الحد ، وتوزيع مربع كاي .(χ2)

وصف العلاقات بين المتغيرات

يمكن تحليل العلاقات بين المتغيرات باستخدام الجداول والرسوم البيانية والإحصاءات . طُوِّرت إجراءات مختلفة لأنواع مختلفة من البيانات . غالبًا ما نستخدم جداول الاحتمالات للبيانات الاسمية والترتيبية ، ومخططات التشتت ومعاملات الارتباط لبيانات النسبة والفاصل .

جداول الاحتمالات

يُظهر جدول الاحتمالات ثنائي المتغير كيفية توزيع قيم متغير بناءً على قيم متغير آخر . عدد الملاحظات في فئة متغير واحد "مشروط" بالفئة التي تنتمي إليها كل ملاحظة من المتغير الآخر . يمكن إظهار أكثر من متغيرين ، ولكن الحد الأقصى عمومًا هو ثلاثة ، نظرًا للحاجة إلى عدد كبير من الملاحظات لإظهار توزيع متغير يعتمد على قيم متغيرين آخرين . في الجدول 17.3، أُعيد تجميع الأنواع الوظيفية للجدول 17.1 في ثلاثة أنواع من المدن . كما تم تجميع معدلات النمو السكاني للمناطق الإدارية الخاصة من عام 1950 إلى عام 1980. ترددات الخلايا (33، 11، إلخ) هي عدد الملاحظات . وبالتالي ، تم تصنيف 33 مدينة من أصل 52 مدينة ذات معدلات نمو أقل من 40% على أنها صناعية أو متنوعة / صناعة .

Table 17.3 US urban growth rates, 1950–80, by urban functional type

Population growth	*Functional type, 1950*			*Total*
	Manufacturing	*Service*	*Government/Prof*	
<40%	33	11	8	52
	(58.93)	(36.67)	(13.56)	(35.86)
40.01%–75.0%	20	11	21	52
	(35.71)	(36.67)	(35.59)	(35.86)
>75.0%	3	8	30	41
	(5.36)	(26.67)	(50.85)	(28.28)
Total	56	30	59	145

Notes: Proportions (%) within categories are shown in parentheses under numeric counts.
Source: Data from Subramaniam-Bryson (1991)

النوع الوظيفي هو متغير اسمي . وقد أدى تجميع معدلات النمو، وهو متغير نسبة ، إلى تحويله إلى متغير ترتيبي . تُستخدم الجداول مثل هذه لإظهار العلاقة بين متغيرين حيث يقع عدد كبير من الملاحظات في كل فئة من فئات البيانات ، المتغيرات . لذلك ، تُستخدم جداول الطوارئ عمومًا لإظهار العلاقات بين المتغيرات الاسمية والترتيبية . معدل النمو هو المتغير التابع في الجدول 17.3، والنوع الوظيفي هو المتغير المستقل (ينظر الفصل 4). وذلك لأن النوع الوظيفي في عام 1950 لا يمكن أن يكون "مُسببًا" بمعدلات نمو لاحقة . يجب أن تكون فئات المتغير التابع دائمًا في أسفل الجدول مع فئات المتغير المستقل في الأعلى (تمامًا مثل الرسوم البيانية) . وهذا يجعل العلاقة الموضحة في الجدول أكثر وضوحًا .

تُحسب النسب المئوية في الجدول 17.3 لمعرفة ما إذا كان هناك فرق بين الأنواع الوظيفية من حيث سرعة نمو المدن . وبما أن المتغير التابع في أسفل الجدول ، تُحسب ترددات الخلايا كنسبة مئوية من إجمالي الأعمدة . لذا ، فإن 33 كنسبة مئوية من 56 تساوي 58.93. تُظهر النسب المئوية أن المدن الصناعية شهدت أقل معدل نمو، كما يتضح من نسبة 59% في أدنى فئة نمو و5% في أعلى فئة نمو. مالت المدن الحكومية والمهنية إلى النمو الأسرع ، بينما اتسمت مدن الخدمات بنمط أكثر توازناً . يبدو أن النوع الوظيفي في عام 1950 كان له تأثير قوي على معدلات النمو اللاحقة (ينظر الملحق 17.2).

من المحتمل أن تكون العلاقة زائفة . يمكن أن يؤثر الموقع الإقليمي في الولايات المتحدة على كل من النوع الوظيفي ومعدلات النمو الحضري . لتقييم التأثير النسبي للموقع الإقليمي ، يمكن تقسيم البلاد إلى ثلاث مناطق (مثل الشمال والجنوب والغرب) ، ويمكن بعد ذلك تضمين المنطقة في جدول ثلاثي الاتجاهات لمعرفة ما إذا كان النوع الوظيفي يؤثر على معدلات النمو عند التحكم في المنطقة . كمتغير مستقل ، ستكون الفئات

الثلاث للمنطقة أيضًا في أعلى الجدول . تُقسّم كل فئة من فئات المناطق إلى ثلاثة أنواع وظيفية ، مما ينتج عنه تسع فئات في أعلى الجدول ، ومع ثلاث فئات نمو ، سيكون هناك 27 خلية إجمالية . عند هذا المستوى من التفصيل، قد لا تُسجّل بعض الخلايا أي ملاحظات . مع ذلك ، يُمكن حساب التكرارات النسبية كنسبة مئوية من إجمالي الأعمدة . إذا كانت نسبة عالية من المدن الصناعية ذات معدل نمو منخفض في كل منطقة ، فإن النوع الوظيفي يُمارس تأثيرًا مستقلًا على معدلات النمو. إذا كانت المدن الصناعية ذات معدل النمو المنخفض تميل إلى التواجد في منطقة واحدة فقط ، فإن المنطقة هي التي تُؤثر على معدلات النمو ، في الواقع ، كان لكلا المتغيرين تأثيرات مستقلة قوية على معدلات النمو الحضري.

مخططات التشتت ومعاملات الارتباط

تُعدّ الجداول قيّمة عند دراسة العلاقات بين المتغيرات الترتيبية والاسمية . ولكن عندما تكون لدينا بيانات نسبية أو فاصلة ، فإن مخططات التشتت (وتسمى أيضًا مخططات التشتت) ومعاملات الارتباط البسيطة تكون أكثر ملاءمة . مخططات التشتت هي رسوم بيانية تستفيد من حقيقة أن متغيرات الفاصل / النسبة تحتوي على معلومات حول حجم الاختلافات بين جميع أزواج الملاحظات . ترسم هذه الرسوم البيانية قيم المتغيرات لكل ملاحظة في بُعدين . وكما هو معتاد ، عند رسم هذه المتغيرات بيانيًا ، نضع المتغير التابع ، Y، على المحور الرأسي ، والمتغير المستقل ، X، على المحور الأفقي . **يُعد تصور هذه الرسوم البيانية طريقة مباشرة لتقييم وجود العلاقات من عدمه** . كما يمكن أن يخبرنا أيضًا بشيء عن شكل العلاقة (موجبة ، سالبة ، خطية ، غير خطية) . إذا بدت النقاط في مخطط التشتت وكأنها تتبع خطًا مستقيمًا ، فيمكن عندئذٍ رسم خط يصف العلاقة المتوسطة لهذه النقاط . سنتناول هذه التقنية في الفصل 18؛ لأغراض التوضيح ، ومع ذلك ، تُظهر مخططات التشتت في هذا الفصل خط "أفضل ملاءمة" عبر نمط النقاط.

هناك عدد من إحصاءات الارتباط المختلفة لمتغيرات ذات مستويات قياس مختلفة . وأكثرها شيوعًا لبيانات الفترات والنسب هو معامل ارتباط بيرسون لحاصل الضرب . وهو مقياس بسيط وحيد يُخبرنا بدرجة "التباين" أو التحرك حول متوسطي متغيرين ، بطرق متشابهة أو مختلفة (أو لا يتغيران على الإطلاق) . يتراوح المقياس، المعروف باسم r، من – 1.0 إلى 0، وصولًا إلى +1.0. تُشير نقاط النهاية السالبة والموجبة إلى وجود ارتباط مثالي بين المقياسين ؛ وهذا عندما يكون المتغيران متماثلين بالفعل ، وإذا واجهت مثل هذه النتيجة ، فستحتاج إلى إعادة التحقق من بياناتك . بين هذين الحدين ، توجد درجات ارتباط بين المتغيرات : على سبيل المثال ، بين تساقط الثلوج من نوفمبر إلى فبراير وارتفاعات الأنهار في مارس وأبريل (الفرضية: إيجابية) ؛ بين معدلات الفقر ومستويات التعليم (الفرضية: سلبية) ؛ بين الارتفاع وكثافة الأشجار (الفرضية: سلبية)؛ بين سعر القدم المربع من مساحات البيع بالتجزئة والمسافة إلى مركز المدينة (الفرضية: سلبية) ؛ بين ساعات دراسة الجغرافيا ومعدل التراكمي الجغرافي (الفرضية: إيجابية).

صيغة ارتباط بيرسون اللحظي هي : هذا أحد أهم المقاييس في العلوم، ولهذا السبب، من المفيد تخصيص بضع فقرات لمعرفة كيفية عمله . أولًا، تذكر أن X و Y هما المتغيران المستقل والتابع على التوالي ، وأن s_x و s_y هما الانحرافان المعياريان لهما . تذكر، ثانيًا، أن المقياس يتراوح بين – 1.0 و+1.0 ، حيث تشير القيم الموجبة لـ r إلى ارتباط موجب ، بينما تشير القيم السالبة لـ r إلى علاقة عكسية.

كيف تعمل الصيغة لإنشاء هذه النتيجة ؟ لكل قيمة ملاحظة فردية على X وY، نأخذ انحرافها عن متوسط X وY، ونحصل على حاصل ضرب تلك القيمة . سيكون حاصل الضرب صفرًا إذا وقعت أي من القيمتين على متوسطهما الخاص ؛ إذا كانت كلتا القيمتين أكبر من متوسطهما ، فإن القيمة المعادة تكون موجبة ؛ وينطبق

الشيء نفسه إذا كانت كلتا القيمتين سالبتين (لأن حاصل ضرب سالبين ينتج قيمة موجبة)؛ وإذا كان أحدهما أعلى والآخر أقل من متوسطيهما، فإن القيمة المرتجعة ستكون سالبة.

من حيث العلاقات ، عندما يتم إرجاع القيم الموجبة باستمرار، فهذا يعني أن مجموعة البيانات تتميز بملاحظات تُوجد فيها قيم عالية في أحد المتغيرات مع قيم عالية للمتغير الآخر، وأنه عندما تظهر قيم منخفضة ، تظهر أيضًا قيم منخفضة في المتغير الآخر. هذه حالة علاقة موجبة أو مباشرة . عندما تهيمن المنتجات السالبة ، ينعكس الوضع : توجد قيم منخفضة في أحد المتغيرات جنبًا إلى جنب مع قيم عالية للمتغير الآخر. لاحظ أن "موجب" و"سلبي" يشيران إلى اتجاه العلاقة ، وليس قوتها . الارتباط +0.6 أضعف من الارتباط -0.7.

نتبع هذا الحساب لكل ملاحظة في مجموعة البيانات ، مع الاحتفاظ بإجمالي جارٍ كما هو موضح بعلامة الجمع . يخبرنا المجموع الناتج باتجاه العلاقة قيد الدراسة :

(أ) إذا كانت موجبة، فإن معامل الارتباط يكون موجبًا؛

(ب) إذا كانت سالبة، فإن العلاقة تكون سالبة؛ و

(ج) إذا كانت القيمة صفرًا، أو منخفضة جدًا، فإن النتائج الإيجابية قد أُلغيت طوال الإجراء بالنتائج السلبية، ولا يوجد اتساق في مدى تباين المتغيرين فيما يتعلق بمتوسطاتهما الفردية.

تهدف العمليات المتبقية أساسًا إلى قياس r إلى عدد بلا أبعاد يمكن مقارنته عبر المتغيرات ، بل ومجموعات البيانات . أولًا، يُنشئ قسمة البسط على n متوسطًا لانحرافات حاصل الضرب لكل حالة على حدة . ومع ذلك ، فإن مقياس هذه القيمة يكون في X · Y ؛ لتحويله إلى r ، نقسم ببساطة على s x · s y ، الانحرافات المعيارية لـ X و Y الموجودة بالفعل في مقياس X · Y. وبالتالي، تقسم هذه العملية حاصل ضرب Xو Y على حاصل ضرب آخر له، مما يعطينا مقياسًا مستقلًا عن قيم X أو Y. وبعد إعادة قياسه لإزالة التباين في كلا المتغيرين، يتراوح r بالتالي من − 1.0 إلى +1.0.

وبشكل عام ، بالنسبة للتحليلات التي تكون فيها n > 50 ، فإن قيم r التي تزيد عن +0.25 أو تقل عن − 0.25 تميل إلى أن تكون "ذات دلالة إحصائية" - أي كبيرة بما يكفي لنفترض وجود علاقة بالفعل ، حتى وإن كانت مخفية بفعل عوامل أخرى تعمل على إحداث تباين في Y مشكلة المتغير المفقود . يوضح الشكل 17.3 أدناه العلاقة بين مخططات التشتت ومعاملات الارتباط . تم عرض أفضل خطوط التوافق (الموضحة في الفصل 18) للتيسير . وهي تلخص تشتت النقاط ، ولكن كما ترى ، من المفترض أن تكون مفيدة إذا كنت ترغب في التنبؤ بقيم المحور الرأسي (Y) بناءً على قيم المحور الأفقي(X) .

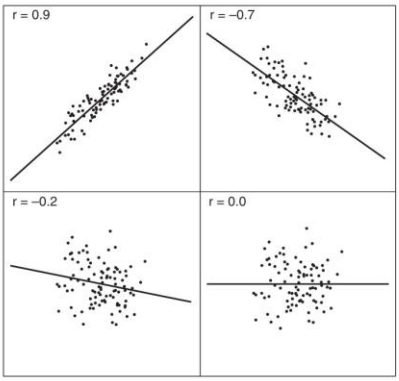

Figure 17.3 Scatter diagrams, correlation coefficients, and best fit lines

وأخيرًا، يجب أن تضع في الحسبان أنه في حين يركز الباحثون غالبًا على العلاقات الخطية ، وفي هذه الحالة تُوجَّه الملاحظات حول خط مستقيم يمر بها (كما في اللوحين العلويين في الشكل 17.3)، إلا أن العلاقات في الواقع يمكن أن تكون أكثر تعقيدًا بكثير. ولهذا السبب، تحتاج إلى تصور البيانات ، بحيث لا تعتمد على مقياس واحد مثل معامل الارتباط (كلارك 1951). يمكن أن تساعد مخططات التشتت في الكشف عن جميع أنواع العلاقات الأخرى . على سبيل المثال ، مع زيادة قيم المحور الرأسي (X)، قد تميل قيم المحور الرأسي (Y) إلى الانخفاض ، ثم تزداد مرة أخرى مع زيادة قيم المحور الأفقي (X) . وتُسمى هذه العلاقة التي تأخذ شكل حرف U بالعلاقة التربيعية.

يمكن أن تكون العلاقة أيضًا علاقة عكسية (U)، وهي علاقة تربيعية تميل فيها قيم (Y) إلى الزيادة أولاً ثم الانخفاض. كما أن مخططات التشتت مهمة. تُستخدم هذه الطريقة لتحديد المنحنيات على شكل حرف Jأو على شكل حرف J المقلوب، وللكشف عن المنحنى الشائع على شكل حرف S ، كما نجده في دراسات الانتشار، من بين دراسات أخرى.

الخلاصة

تدمج الحسابات الوصفية للبيانات الرقمية المصفوفات الجدولية والرسوم البيانية والمقاييس الإحصائية الفردية (المتوسطات ، الانحرافات المعيارية ، معاملات الارتباط) . تُعد معرفة بياناتك وصفيًا الخطوة الأولى والأهم في تحليل البيانات . فهي تُمكّنك من تحديد المشاكل في جمع البيانات (مثل القيم المتطرفة أو الشاذة) ؛ وتمنحك فهمًا دقيقًا للملاحظات ومتغيراتها (بشكل فردي وبالاشتراك مع متغيرات أخرى) ؛ وتساعدك في توجيهك في طرح أسئلة من مستوى أعلى حول السببية أو التفسير، بما في ذلك تلك التي تُدرك عالمًا متعدد المتغيرات إلى حد كبير - أو مُسببًا بشكل مُضاعف . ومع ذلك ، فإن الخطوة الأولى في تفسير هذا العالم المترابط بشكل معقد هي معرفة اتجاهات متغيرك ، وتقلباته ، وعلاقاته البسيطة مع المتغيرات الأخرى.

التمرين ١٧,١:

مخططات التشتت ومعاملات الارتباط البسيطة

البيانات أدناه خاصة بمقاطعات أريزونا الخمس عشرة (مكتب تعداد الولايات المتحدة لعام ٢٠٠٠). المتغير الأول هو نسبة إجمالي الأشخاص الذين يعيشون تحت خط الفقر . المتغير الثاني هو نسبة السكان الذين تزيد أعمارهم عن ٢٥ عامًا والحاصلين على شهادة أقل من الثانوية العامة . لنفترض في هذه المسألة أن الفقر تابع (Y) وأن التعليم مستقل .(X) نريد تصوّر العلاقة بين هذين المتغيرين وحساب معامل الارتباط . باستخدام ورقة رسم بياني ، ارسم قيم المقاطعات الخمس عشرة في فضاءي Y و .X هل يبدو أن هذه البيانات تقدم علاقة موجبة ، أم سالبة ، أم لا ؟ بناءً على مخططات التشتت في الشكل ١٧,٣، ما هو معامل الارتباط الذي تتوقعه ؟

الآن، احسب معامل الارتباط لهذه المجموعة من البيانات . يمكنك استخدام المساحة المتاحة في الجدول المُرفق ، أو نسخ الصفحة واستخدامها كورقة عمل . يُسمح باستخدام الآلة الحاسبة اليدوية ، ولكن الهدف من التمرين هو فهم العمليات الحسابية المُستخدمة . بعد حساب r، هل يبدو أن هناك علاقة قوية أم ضعيفة في البيانات ؟ لو كنتَ مسؤولًا في ولاية أريزونا ، فماذا ستستنتج بناءً على هذا التمرين القصير؟

الملحق 17.1

متباينة تشيبيشيف

تشير متباينة تشيبيشيف (المعروفة أيضًا باسم نظرية تشيبيشيف) إلى أنه ، بغض النظر عن شكل التوزيع التكراري ، فإن ما لا يقل عن 1 1 100 2 %k ( ) – من القيم تقع ضمن) k لأن (1 < k انحراف معياري على جانبي المتوسط . لذا ، يجب أن تقع 88.89% على الأقل من قيم المتغير ضمن ثلاثة انحرافات معيارية عن المتوسط . عادةً ما تقع ما يقرب من 100% من القيم ضمن ثلاثة انحرافات معيارية عن المتوسط . إذا كان للمتغير توزيع طبيعي ، فإن 68.26% من القيم تقع ضمن انحراف معياري واحد عن المتوسط ، و95.44% ضمن انحرافين معياريين عن المتوسط، و99.74% ضمن ثلاثة انحرافات معيارية .

لذا ، فإن القيم التي تزيد عن 3 انحرافات معيارية عن المتوسط غير شائعة جدًا ، وقد تشير إلى خطأ في البيانات . في بيانات VTO، تقع ثلاثة انحرافات معيارية أسفل وأعلى من المتوسط عند 24.54 و78.80، وينتج عن انحرافين معياريين عن المتوسط قيم 33.58 أو 69.75 . في البيانات الفعلية، تكون إحدى قيم VTO أقل من 24.54 وأربع قيم أكبر من 78.80. لذا، فإن 98.03% من القيم تقع ضمن ثلاثة انحرافات معيارية عن المتوسط . خمس قيم أصغر من انحرافين معياريين أسفل المتوسط ، وسبع قيم أكبر من انحرافين معياريين أعلى من المتوسط . لذا ، فإن 95.28% تقع ضمن انحرافين معياريين عن المتوسط . ولكن قيمة واحدة أكبر من أربعة انحرافات معيارية أعلى من المتوسط ، وبالتالي لا يمكن عد التوزيع طبيعيًا.

الملحق 17.2

معامل كرامرV

معامل كرامر V هو معامل ارتباط قائم على مربع كاي ، ويعتمد بدوره على الفرق بين ترددات الخلايا الفعلية والترددات المتوقعة إذا كان المتغيران مستقلين تمامًا عن بعضهما البعض . في الجدول 17.3، سجّلت 35.86% من جميع المدن معدلات نمو أقل من 40% . إذا لم يكن المتغيران مرتبطين ، فيجب أن يكون لدى 35.86% من كل نوع من أنواع المدن معدلات نمو أقل من 40% . من بين 56 مدينة في مجموعة التصنيع ، تبلغ نسبة 35.86% 20.08، وبالتالي يجب أن يكون لدى 20.08 من هذه المدن معدلات نمو أقل من 40.% . يجب أن يكون لدى مجموعة الخدمات 10.76 بمعدلات نمو أقل من 40%، ويكون التكرار المتوقع لمجموعة الحكومة / المهن هو 21.16.

صيغة مباشرة لحساب التكرارات المتوقعة هي: إجمالي الصفوف للخلية ، إجمالي الأعمدة للخلية n حيث n هو إجمالي عدد المشاهدات. للخلية الأولى، f e = (52)(56)/145 = 20.08

كما في السابق . في الواقع ، 33 مدينة من مدن مجموعة التصنيع لديها نمو منخفض مقارنةً بالقيمة المتوقعة البالغة 20.08، و8 مدن حكومية / مهنية فقط لديها معدلات نمو منخفضة ، حيث كان من المتوقع أن تبلغ 21.16 إذا كان المتغيران مستقلين عن بعضهما البعض . يكشف حساب جميع التكرارات المتوقعة أن المدن الحكومية / المهنية تميل إلى أن تكون في الفئة الثالثة (أعلى معدل نمو) للمتغير التابع ، وأن المدن الصناعية تميل إلى أن تكون متركزة في الفئة الأولى من معدل النمو. هذا يصف "طبيعة" العلاقة . مربع كاي (χ2) هو إحصائية تُشتق من الفرق بين الترددات المرصودة والترددات المتوقعة في جميع الخلايا.

حيث f o و f e هما الترددان المرصود والمتوقع في كل خلية ، ويكون المجموع على جميع الخلايا . **تنص المعادلة على أن الفرق بين الترددات المرصودة والمتوقعة يُربع ثم يُقسم على التردد المتوقع للخلية**.

ثم تُجمع هذه الكميات لجميع الخلايا . في الجدول 17.3، يكون مجموعها 37.572. إذن، قيمة V لكرامر هي مؤشر بين 0 و1 (من عدم وجود علاقة إلى قيمة واحدة مثالية) ، وتُحسب على النحو الاتي : حيث r هو عدد الصفوف ، وc هو عدد الأعمدة . يعني المقام الموجود أسفل الجذر أن عدد المشاهدات (n) مضروب في أصغر قيمة بين 1 – r و 1 – c. يمكن حساب قيمة V لكرامر للجداول غير المتماثلة ، أي حيث r لا يساوي c . لذا ، قيمة V لكرامر هي الجذر التربيعي لـ 37.572/((145)(2))، والذي يساوي 0.36. العلاقة ليست قوية تمامًا كما يوحي الجدول ، لأن التكرارات المتوقعة والتكرارات المرصودة لا تختلف اختلافًا ملحوظًا إلا في الخلايا الأربع في زوايا الجدول . وإلا ، فإن النوع الوظيفي ليس مؤشرًا جيدًا على فئة النمو التي تندرج تحتها الملاحظة.