تحديد القيم الشاذة باستخدام الطرق الاستكشافية ومقارنتها مع الطرق المعلمية

أ. م. د. محمود مهدي حسن البياتي — الباحث
كلية الادارة والاقتصاد/جامعة بغداد/ قسم الاحصاء — دلير صليو دغا

الخلاصة

تلعب البيانات الاحصائية دوراً مهماً في عملية التخطيط المركزي والدراسات العلمية ، وتأتي أهمية هذه الدراسة لتعاملها حول كيفية حفظ البيانات الاحصائية من الاخطاء المتوقعة والقيم الشاذة. الهدف من البحث هو تحديد القيم الشاذة بأستخدم الطرق الاستكشافية الحديثة (Exploratory Data Analysis) ومقارنتها مع الطرق المعلمية بعد تحديد القيم الشاذة بأستخدام الاسلوبين. وجد هناك أختلاف بين تحديد القيم الساذة بأستخدام الطرق الآستكشافية وطرق المعلمية وكذلك أختلاف بين الطرق المعلمية نفسها.تبين من خلال الدراسة أنه أفضل طريقة لتحديد القيم الشاذة هي طريقة الرسم الصندوقي (Box plot) .

Abstract

The availability of statistical data plays an important role in planning process. The importance of this research which deals with safety of statistical data from errors and outliers values. The Objective of this study is to determine the outlier values in statistical data by using modern exploratory data methods and comparing them with parametric methods. The research has been divided into four chapters ,the main important conclusions reached are:1-The exploratory methods and the parametric methods showed variation between them in determining the outlier values in the data.

2-The study showed that the box plot method was the best method used in determining outlier values in data.

---

* بحث مستل من أطروحة الماجستير/ قسم الاحصاء, 1996 الموسومة "تحديد القيم الشاذة بأستخدام الطرق الاستكشافية ومقارنتها مع الطرق المعلمية ".

1-1 المقدمـــة

التخطيط الشامل هو المفتاح المركزي لعملية التنمية ، وتلعب مسألة توفر البيانات الاحصائية دوراً مهماً في عملية التخطيط اذ ان عملية التخطيط المستقبلية تحتاج الى تحليل للواقع الحالي الذي يرتكز على دقة البيانات الاحصائية للمتغيرات المراد دراستها وما لها من اثر واضح على الخطة المرسومة من هنا تبرز اهمية هذا البحث الذي يتناول التأكيد من سلامة البيانات الاحصائية المستخدمة في المتغيرات قيد الدراسة ومحاولة تصفية هذه البيانات من الاخطاء والقيم الشاذة (Outlier values) } القيم الشاذة هي القيم الغير متجانسة مع بقية البيانات، ومنفصلة عن جسم البيانات{، والتي قد تؤدي في حالة تجاهلها الى تغيير نتائج التحليل الاحصائي الى نتائج غير دقيقة او انحراف النتائج عن واقعها الذي يجب ان تكون عليه فيما لو اجري التحليل الصحيح على البيانات الصحيحة او توجيهها الى اتجاه اخر بعيد عن الواقع، ومن ثم اعتماد اساليب وهمية لمعالجة متغيرات الظاهرة المدروسة.

ان الغرض من دراسة القيم الشاذة في البيانات هو محاولة تنقية البيانات من الشوائب والقيم الغريبة بأساليب وطرق علمية، لأن وجود القيم الشاذة سوف يؤثر بشكل سلبي على نتائج التحليل الاحصائي مما يقود الى عدم الاستفادة من نتائج التحليل كما هو مخطط له. أن دراسة القيم الشاذة (Outlier Values) ربما لا تقتصر على اكتشاف القيم الشاذة، وانما تتعدى ذلك لتشمل معالجة القيم الشاذة في البيانات وقد يعزى شذوذ البيانات كما اوردتها (Barnett 1978)[5] الى الاسباب التالية:

1- أخطاء الحساب (Calculation Error)
2- أخطاء القراءة (Reading Errors)3- أخطاء التسجيل (Recording Errors).

واذا تم تحديد السبب او الاسباب في شذوذ القيم المذكورة اعلاه او غيرها سهلت عملية المعالجة.

وقد يتم تصحيح القيم الشاذة او يتم حذف البيانات المسببة للاخطاء اما في حالة الشك بوجود قيم معينة شاذة دون دليل واضح عن سبب حدوثها فلابد من اعتماد الاساليب الاحصائية المناسبة لتحديد تباعدها (أبتعاد القيم عن جسم البيانات)، وأن عملية حذفها في هذه الحالة سيؤدي الى نقص في المعلومات المتوفرة للدراسة المعينة، لأنها قد تمثل حالة خاصة او ظرفاً طارئاً وأنها قيم صحيحة وأن هناك عدة احتمالات لمعالجة تلك القيم الشاذة كما جاء في دراسة المختار، سليمان[4] عام 1980 وهي:

1- رفض القيمة (القيم) الشاذة ثم اجراء التحليل الاحصائي على بقية البيانات او اللجوء الى تقدير تلك القيمة (القيم) ومن ثم اجرء التحليل الاحصائي المناسب على المجموعة الجديدة من البيانات.
2- ان وجود القيم الشاذة لايعني في جميع الاحوال بأنه حاله سلبية فقد تعطي القيم الشاذة اهتماماً خاصاً لأنها يمكن ان تمثل حالة خاصة او ربما يكون هذا الوجود مؤشراً جديداً لعوامل جديدة او دليلاً على تفسير البيانات وفق سياقات معينة قد تكون ذات اهمية بالغة وقد تسهم اسهاماً بالغاً في تفسيير الظاهرة المدروسة.
3- معالجة البيانات كما هي دون استثناء أي من قيمها بواسطة الاساليب الاحصائية المناسبة كأن يوضع لها نموذجاً معدلاً يأخذ بنظر الاعتبار حالة القيم الشاذة ومدى تأثيرها على طبيعة مثل هذا النموذج.وتظهر القيم الشاذة بشكل نتيجة لبعض العوامل، بعضها يتم السيطرة عليه من قبل الباحث والبعض الاخر لا يمكن السيطرة عليه وقد صنفتها Barnett 1978 كما يلي:

1- أخطاء المعاينـــة (Sampling Errors) وهي الاخطاء الناتجة عن اختيار العينة.
2- الاختلاف الاصلي (Inherent Variability) لايمكن السيطرة على هذا النوع من الاختلاف لكونه من الخصائص الطبيعية الملازمة للمجتمع الاصلي.3 -الخطأ الناتـج من سبـب طبيعـي تكون القيم الشاذة في هذه الحالة نتيجة طبيعية للاختلافات الموجودة في قيم المتغيرات الاصلية.

وهناك عدة تعاريف للقيم الشاذة حيث عرفها (Bross,1961)[9] على انها المشاهدات التي تظهر منحرفة بشكل كبير عن جميع قيم العينة. كذلك عرفها (Kishpaugh,1972)[18] بأنها المشاهدة التي يكون ابتعادها عن باقي قيم العينة يستحق الاهتمام، وبذلك يعني ان لأي شخصي مهتم في تحديد القيمة الشاذة. كما عرفها (AL-Jobouri,1976)[16] بأنها تلك القيمة التي تكون غير متجانسة او منسجمة مع بقية بيانات المجموعة لمتغير من المتغيرات لظاهرة معينة او مجموعة من الظواهر، او ان القيم الشاذة هي القيم التي تأتي من مجتمع يختلف عن مجتمع العينة قيد الدراسة. اما (Barnett,1978)[5] فقد عرفت القيمة الشاذة في العينة، بأنها تلك القيمة التي تبدو غير منطقية اذا ما قورنت بباقي القيم.

أن معظم الدراسات التي تناولت القم الشاذة كانت تتعامل مع القيم الشاذة كحالة (Parametric Case) وتحت افتراض ان البيانات تتبع التوزيع الطبيعي (Normal Distribution) والمشكلة تظهر عندما لا تتوزع البيانات توزيعاً طبيعياً، ولأجل الابتعاد عن مشكلة تحديد التوزيع الاحتمالي الملائم للبيانات وللتطور الذي حصل بأستخدام الحسابات والذي ادى الى دفع الاحصاء مئات السنين الى الامام، واكتشاف طرق حديثة لتحليل البيانات الاحصائية فأن الكثير من الدراسات في الفترة الاخيرة بدأت تأخذ منحنى اخر وهو التعامل مع الحالة الامعلمية (Non-Parametric Case) والطرق الاستكشافية الحديثة (Exploratory Data Analysis).

1.2 ـ الهدف من البحـــث

قبل البدء بعملية أي تحليل للبيانات الاحصائية لابد من التأكد من صحة وسلامة البيانات وخلوها من الاخطاء والقيم الشاذة، حيث ان نتائج البحث العلمي المبنية على احصاءات غير صحيحة لايمكن الاعتماد عليها بأي شكل من الاشكال لأنها سوف تكون بعيدة عن الواقع. أن هدف هذا البحث هو أستخدام طرق أستكشافية جديدة(Exploratory Data Analysis) لتحديد القيم الشاذة في البيانات الاحصائية بأستخدام الرسم وهي (Box plot, Stem and Leaf, and Rangefinder Box plot) وأجراء مقارنة بين الطرق الاستكشافية وبين الطرق المعلمية لتحديد القيم الشاذة وايهما افضل بالتطبيق العملي.

1.3 ـ الدراسات السابقـــة عن القيـــم الشاذة

بدائة فكرة وجود القيم الشاذة منذ منتصف القرن الثامن عشرعندما حاول (Boscovich,1755)[5] تحديد أهليجية الارض معتمداً على معدل القياسات التي سيحصل عليها وقد استطاع الحصول على عشرة قياسات، أستبعد اثنين منها لتطرفهما الشديد ثم وجد المعدل للثمانية الاخرى.

يعتبر (Peirce,1852)[4] اول من اعتمد اسلوب الاختبار للقيم الشاذة فقد نص اختباره على رفض (K) من المشاهدات الشاذة في عينة حجمها (n) اذا كان احتمال منظومة الاخطاء (System of Error) الناتج عن الاحتفاظ بـ (K) من المشاهدات الشاذة اقل من احتمال رفضها مضروباً بأحتمال الحصول على هذا العدد وليس اكثر من المشهدات وقد حدد (Peirce) الأحتمال الاخير بالصيغة التالية $\theta^k(1-\theta)^{n-k}$ اذا عرف (θ)بأنه احتمال وجود مثل هذه المشاهدات الشاذة المرفوضة بالنظر لكبر قيمتها ومن ثم حدد له القيمة k/n.

وأقترح (Wright,1913)[3] رفض أي مشاهدة شاذة تنحرف عن المتوسط بأكثر من ثلاثة أضعاف الانحراف المعياري.

بعد ذلك اقترح (Goodwin,1913)[16] رفض المشاهدات في عينة حجمها (n) اذا زاد انحرافها عن متوسط باقي المشاهدات (n-1) مشاهدة بأربعة أضعاف معدل انحراف (n-1) مشاهدة. ولم تكن جميع الاحصاءات المقترحة قبل عام 1925 تميز بين تباين المجتمع وتباين العينة.

وكان (lrwin,1925)[15] اول من استطاع ان يفرق بين هذين النوعين من التباينات، اذا شدد على ضرورة أستخدام الانحرف المعياري للعينة (s) كتقدير الى الانحراف المعياري للمجتمع عندما يكون الانحراف المعياري للمجتمع (σ)مجهولاً... وقد أقترح أستخدام مؤشرات احصائية (عندما تكون σمعلومة). وأقترح جداول بالقيم الحرجة لهذه المؤشرات الاحصائية لتحديد تباعد القيم الشاذة وللتفصيل(أرجع دلير 1996). واقترح Thompson[16] في عام 1935 مؤشراً احصائياً يعتمد على اختبار T.

ثم تطورت دراسة القيم الشاذة لتشمل بالاضافة الى دراستها كقيم شاذة ضمن العينة المفردة دراستها في تصميم التجارب والانحدار والبيانات متعددة المتغيرات.....الخ.

وفي عام 1950 اقترح Grubbs[13] بعض المعايير لاختبار معنوية الشذوذية للمشاهدة الكبرى في عينة بحجم n مسحوبة من مجتمع طبيعي (Normal Population)، كذلك اقترح معايير اخرى لاختبار معنوية الشذوذية لاختبار ما اذا كانت المشاهدات الكبرى الاولى والثانية في العينة متطرفتين بالكبر الى حد بعيد، او ان المشاهدتين الصغرى الاولى والثانية في العينة متطرفتين في الصغر الى حد بعيد...اما التوزيعات الاحصائية التي اقترحها Grubbs[13] لاحصاءاته فقد اعتمد اشتقاقها الى حالة المجتمع الطبيعي.

كما اقترح جداول بالنسب المئوية (Percentage Points) التقديرية.

واقترح Mosteller[21] عام 1948 اختبار انزلاق (Silppage) K من العينات ثم حدد النسب المؤية لحالات معينة وعندما تكون حجوم العينات متساوية الا ان العينات قد لاتكون متساوية الحجوم ...لقد ادرك Mosteller & Tukey[22] في عام 1950 هذه النقطة واستطاعا تحديد مستويات معنوية دقيقة لعينات مختلفة الحجوم بأقتراحهما صيغة تأخذ بنظر الاعتبار تلك الاختلافات في الحجوم.

درس (Zinger 1961)[30] ظاهرة الشذوذية في عدة مجتمعات طبيعية (لها تباينات معلومة)... ثم اقترح مؤشراً احصائياً لاختبار شذوذية التباعد لمجتمعات يصل عددها الى سبع مجتمعات ...فقد اعتمد هذا المؤشر على الفرق المعياري بين اكبر متوسطي عينتين، وقد اعطى جداول خاصة بالقيم الحرجة.

وقدم Mcmillan[20] عام 1971 اسلوباً جديداً في اختباره لشذوذية مشاهدة واحدة او مشاهدتين في عينات طبيعية عندما يكون تباين المجتمع مجهولاً ويتضمن اسلوب Mcmillan ثلاثة طرق...

الطريقة الاولى وتتضمن التنفيذ المتسلسل لاحصاءة الباقي الاعظم (Maximum Residual). اما الطريقة الثانية فهي تعتمد على اختبار ما اذا كانت المشاهدتان الكبرى الاولى والثانية في العينة شاذتين.

اما الطريقة الثالثة والاخيرة فقد اعتبرت المشاهدتين الكبرى الاولى والثانية شاذتين ولمزيد من التفاصيل ارجع (دلير 1996). وكذلك اقترح احصاءة لاختبار التباعد لمشاهدة واحدة او مشاهدتين في عينات طبيعية على أن يكون تباين المجتمع معلوماً.

وقد لاحظ كل من Gnanadeskan and Kettenring[12] في عام 1972 ان ألاعتبارات المأخوذة للقيم الشاذة في عينة ذات المتغيرات المتعددة تكون اكثر تعقيداً منها في حالة المتغير الواحد.

ان اهداف Gnanadeskan and Kettenring الاولية هي تكوين تقديرات قوية لمواقع المتغيرات المتعددة (Multivariate Locations) وتحديد الاستجابات المتعددة للقيم الشاذة. ولقد تم اقتراح تصميم لرسم نقاط البيانات في الواجهات الذاتية (Eigen Vectors) لمتوسطات البيانات المعروضة لاكتشاف المشاهدات الشاذة المحتمل وجودها.

وقد عدل Kishpaugh[18] في عام 1972 على هذا الاقتراح وسميت بتصاميم المركبات الاساسية.

واقترح كل من Tietjen, Moore and Beckman[27] في عام 1973 اسلوباً لتحويل الاخطاء الى أخطاء معيارية (Standard Residual) لأختبار التباعد لنموذج الانحدار الخطي البسيط وقد حددوا القيم الحرجة بالاعتماد على دراسة المحاكاة، وقد وجدوا ان هذه القيم مقارنة للقيم التي حصل عليها Grubbs.

أقترح Rohlf[24] في عام 1975 صيغة لاكتشاف القيم الشاذة المتعددة (Multivariate Outliers) وسميت باختبار الفجوة العام (Generalized Gap Test)، حيث ذكر ان المشاهدات في حالة متعدد المتغيرات تتخذ شكل شجرة لها فروع متجمعة تقريباً وهناك مجموعة تكون خارجة عنها.

وقد اقترح كل من John and Prescott[24] في عام 1975 عدداً من الاحصاءات لأختبار التباعد في تصاميم التجارب ثم حدد القيم الحرجة لتلك الاحصاءات باستخدام اسلوب المحاكاة.

ناقش AL-Jobouri[16] في عام 1976 الطرق المعلمية لتحديد القيم الشاذة في حالة متغير ومتغيرين وعدة متغيرات في تصاميم التجارب وقد اعتمد Barnett and Lewis[5] في عام 1978 دراسة القيم الشاذة بفرض ان البيانات مأخوذة من توزيع طبيعي متعدد.

استعرض المختار، سليمان في عام 1980 مختلف الطرق التي عالجت القيم الشاذة، في التجارب المصممة ونماذج الانحدار وكذلك متعدد المتغيرات.

أقترح (الجبوري،1988)[1] طريقة لأكتشاف المشاهدات الشاذة في حالة متعدد المتغيرات بأعتماد طريقة الرسم الصندوقي Box plot.

وقد اقترحت (الجبوري، منى 1990)[2] طريقة للآكتشاف الجزئي للمشاهدات الشاذة وطرق التقدير في حالة متعدد المتغيرات وبأعتماد طريقة الرسم الصندوقي Box plot .

واخيراً اقترحت (المشنو،1993)[3] طريقة لأكتشاف المشاهدات الشاذة في تحليل تصاميم التجارب غير المتزنة وبأعتماد طريقة الرسم الصندوقي Box plot.

نجد مما تقدم هناك العديد من الدراسات تناولت القيم الشاذة وسوف نتطرق في البند اللاحق بأختصارالى مفهوم الطرق الاستكشافية .

1.4 الطرق اللاستكشافية الحديثة لتحليـــل البيانات الاحصائيـــة

Exploratory data analysis statistical data analysis

تعتبر هذا الطرق والاساليب التي اغلبها حديثة وتستخدم لعرض وتحليل البيانات. العالم **Tukey** كان له

الدور الكبير باستكشاف وتطوير هذه الاساليب واعطى الفكرة الاولية عليها عام **1970** وهذه، الاساليب يمكن ان تعطينا فكرة واضحة عن توزيع واتجاه البيانات، فهي تفصل وتشخص عناصرمكونات البيانات الاحصائية المهمة الى المحلل. ان المحللين الاحصائيين الجيدين في بادىء الامر يعرضون البيانات بشكل تفصيلي قبل البدء بتحليلها واختيارها للاطلاع على مكوناتها وعلى اتجاه توزيعها، وهذه الطرق تتعامل مع البيانات بمرونة عالية في التحليل ويمكن ان تخدم هذه الطرق في المرحلة الاولى من التحليل لتحديد المقياس المناسب والكفوء لوصف البيانات، ويمكن ان تستخدم هذه الطرق لتحوير البيانات لاستخدام المقياس المناسب بعد ان تقترب اوتتوزع توزيع طبيعي. وقد صممت هذه الطرق لاجراء مقارنات بسيطة او بشكل تفصيلي بين توزيعات البيانات او اجزائها الرئيسية المهمة.

اما الحالات التي لا تحتاج في التحليل للبيانات عرض كل الاجزاء، يمكن ان تأخذ هذه الطرق ملخصآ لعرض توزيع البيانات وهناك قسم من هذه الاساليب يمكن ان تنجز هذه المهمة بسهولة او بشكل جيد جداً مثل القيم الحرفية **(latter Values)** ممكن ان تعرض لنا خمس قيم من البيانات وهو الوسيط **(Median)** والربيع الاول **(First Quartile)** والربيع الثالث **(Third Quartile)** واكبر واصغر قيمة من البيانات وهذا الاسلوب بسيط جداً بحيث يمكن عمله باليد بسهولة. وتوجد طريقة اخرى مهمة جداً لتلخيص البيانات تعرف بالرسم الصندوقي **(Box plot)** تعطي هذه الطريقة انطباعاً بشكل سريع لاجزاء محدودة ومهمة من التوزيع ولشكل انتشار البيانات وعلى شكل رسم بياني. وهذا الاسلوب يعتمد على خمس قيم من البيانات ويمكن استخدامالرسم الصندوقي لاجراء مقارنات متعددة لعدة مجاميع من البيانات.

1.4.1 الطرق الحسابية والطرق الاستكشافية لتحليل البيانات [14]

ان الطرق الاستكشافية هي اكثر فائدة من الطرق الحسابية لتحليل البيانات ولتحديد اجزاء محددة من البيانات، اما الطرق الحسابية فهي اكثر فائدة من الطرق الاستكشافية في حالة صنع قرارات عامة تعتمد على نتائج البيانات، على سبيل المثال بناء نموذج من العينة واستخدامه بالمجتمع.

والطرق الاستكشافية تعتمد نتائجها على القرارات الشخصية للأشخاص وتفسيراتهم وهكذا يمكن ان يكون هناك فروقات بسيطة بين تفسير النتائج من شخص الى اخر بالاعتماد على هذه الطرق، ولكن الطرق الحسابية تعتمد على ملخص التحليل للبيانات الاحصائية المحددة على سبيل المثال الوسط الحسابي والانحراف المعياري لايتأثران بتصورات الاشخاص او تفسيراتهم.

1.4.2 **طــرق الرســـم** ((14)) (Graphical Methods)

كانت الحاجة لترتيب البيانات على شكل جداول او على شكل رسوم بيانية مع بداية الثورة الصناعية في اوربا لكثرة وجود البيانات وكان اكتشاف هالي Halliy طريقة الرسم لتحليل ضغط البارامیتر وهذا الاكتشاف قاد الى استخدام طريقة الرسم لعرض وتحليل البيانات الاحصائية، بعد ذلك تم اكتشاف طرق الرسم واحدة بعد الاخرى ولحد الان وبعد عام 1820 اصبح استخدام طرق الرسم معروفاً وشائعاً بأغلب المقالات والمجلات وكان لها وجود مثلاً في المؤتمر العلمي للاحصاء الذي انعقد في فينا 1852 (International statistical Congress) وفي عام 1872 خصص الكونغرس الامريكي مبلغ من المال لاول مرة لتطوير طريقة الرسم في تحليل البيانات. العقدين الاخيرين شهد اهتمام واسع جداً بطرق الرسم لتحليل البيانات الاحصائية وكان Tukey الاول في هذا المضمار حيث اكتشف عدة طرق لتحليل وتفسير البيانات الاحصائية، وتوجد الان حاجة ماسة لتحليل البيانات الاحصائية بالرسم لأن الرسم يساعد على تحليل البيانات وعرض العلاقات النظرية لها. وان الرسم هو الصورة للبيانات التي تعطي فكرة دقيقة عن المعلومات لهذه المجموعة من البيانات ولعبت الحاسبات دوراً كبيراً في تطوير واكتشاف طرق الرسم، ولهذا تعتبر طرق الرسم الان مهمة جداً في تحليليل البيانات الاحصائية وجزءاً اساسياً يربط بين العلم والتكنلوجيا وسوف يتم التطرق في الفصل الثاني الى اهم طرق الرسم.

2ـ الجانب النظـــري

2.1 **المقدمـــة** :

تم التطرق في الفقرة (1) الى بعض المفاهيم والدراسات السابقة عن القيم الشاذة ومفهوم الطرق الاستكشافية، وسوف يتم التطرق في هذه الفقرة الى الجانب النظري للبحث ويكون على مبحثين، المبحث الاول يتناول عرض بعض الطرق المعلمية (Parametric) المستخدمة لتحديد القيم الشاذة في البيانات، اما المبحث الثاني فيتناول عرض الطرق الاستكشافية الحديثة (Exploratory Data Analysis) التي تستخدم لتحديد القيم الشاذة في البيانات.

2.2 **الطــرق المعلميـة**

تم التطرق في هذا المبحث الى الطرق المعلمية التي تستخدم لتحديد القيم الشاذة في البيانات الاحصائية في حالة المتغير الواحد (Univariate case) والمتغيرين (Two Dimensions) ونماذج الانحدار (Regression models) التي تتطلب معرفة التوزيع الاحتمالي للبيانات الاحصائية.

2.2.1 **الطرق المعلمية المستخدمة لتحديد القيم الشاذة في حالة المتغير الواحد** (Vnivariate).

اولاً : طريقـــة (Grubbs) (13).

نشر Grubbs عام 1950 بحثاً حول مقياس العينة لاختبار المشاهدات الشاذة حيث اقترح فيه بعض المقاييس لاختبار دلالة المشاهدة الشاذة في عينة بحجم (n) تتوزع توزيعاً طبيعياً وقد استخدم Grubbs عدة مقاييس لاختبار القيم الشاذة بعد ترتيب المشاهدات في العينة تصاعدياً حيث تم استخدام المقياس في الصيغة رقم (2-1) لاختبار دلالة اكبر مشاهدة في العينة بحجم (n) فيما اذا كانت شاذة ام لا. اما بالنسبة للمشاهدة الصغيرة فقد استخدم الاحصاءة او الاختبار في الصيغة رقم (2-2)، ولاختبار فيما اذا كانت اكبر مشاهدتين شاذتين ام لا فقد استخدم الاحصاءة او الاختبار في الصيغة رقم (2-3)، وبالنسبة الى اصغر مشاهدتين فقد استخدم الاحصاءة او الاختبار في الصيغة رقم (2-4) وهذه الصيغ مدرجة أدناه. ونقارن هذه الاحصاءات او قيم الاختبارات بما يقابلها من القيم الحرجة في جداول خاصة، ولمزيد من التفاصيل ارجع الى (ديلر 1996).

$$\frac{S_n^2}{S^2} = \frac{\sum_{i=1}^{n-1}(X_i - X_n)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \qquad .....(2-1)$$

عندما

$$\overline{X}_n = \frac{1}{n-1}\sum_{i=1}^{n-1} X_i$$

$$\frac{S_1^2}{S_2^2} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}_1\right)^2}{\sum\left(X_i - \overline{X}\right)^2} \qquad .....(2-2)$$

عندما

$$\overline{X}_1 = \frac{1}{n-1}\sum_{i=2}^{n} X_i$$

$$\frac{S_{n,n-1}^2}{S^2} = \frac{\sum_{i=1}^{n-2}\left(X_i - X_{n,n-1}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \qquad .....(2-3)$$

عندما

$$\overline{X}_{n,n-1} = \frac{1}{n-2}\sum_{i=1}^{n-2} X_i$$

$$\frac{S_{1,2}^2}{S^2} = \frac{\sum_{i=3}^{n}\left(X_i - \overline{X}_{1,2}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \qquad .....(2-4)$$

عندما

$$\overline{X}_{1,2} = \frac{1}{n-2}\sum_{i=3}^{n} X_i$$

ثانياً : طريقـــة (Quesenberry and David)[23]

انصب عمل كل من (Quesenberry and David) في عام 1961 على تحديد القيم الشاذة لاكثر من مجموعة من المشاهدات (k من المجتمعات) لها توزيع طبيعي شريطة ان يكون التباين σ^2 للمجتمعات غير معلوم وقد استخدم المقاييس التالية في الاختبارات والتي ارقام صيغها كما يلي.

(2-5)، (2-6)، (2-7) حيث ان S^{*2} تساوي مجموع مربعات الانحرافات لـ K من المجتمعات.

استخدم (Quesenberry and David) الاحصائتين (2-6),(2-7) اثناء معالجتها لمسألة الانزلاق (التقويت) للمجتمعات الطبيعية. فبعد ان يتم اختيار احدى القيم (المشاهدات المشكوك فيها شاذة) لأي مجتمع كان واجراء الاختبار عليها لتحديدها شاذة ام لا، فأن الاجراء الاتي سيكون باستبعاد المشاهدة الشاذة من أي مجموعة كانت ومن ثم الاستمرار بعملية الاختبار للقيم الاخرى ، حسب الصيغ التالية ارقامها كما يلي: (2-8),(2-9),(2-10) عندما S_1^* تساوي مجموع مربعات الانحرافات لـ K من المجتمعات محسوب بدون المشاهدات الشاذة، وهكذا يستمر الاختبار وتقارن هذه الاحصاءات بما يقابلها من القيم الحرجة في جداول خاصة [16] وهذه الصيغ مدرجة أدناه. وللاطلاع على كل التفاصيل ارجع الى (ديلر 1996).

$$b_i = \left(X_i - \overline{X}\right)/S \qquad .....(2-5)$$

$$b = Max\left(b_i\right) = \frac{X_{\max} - \overline{X}}{S^*} \qquad .....(2-6)$$

$$b^* = Max\left|b_i\right| \qquad .....(2-7)$$

$$b_i = \left(\sqrt{n_i}\,\overline{X}_i - X_W\right)\Big/ S_1^* \qquad .....(2-8)$$

$$b = Max\left(b_i\right) = \frac{Max\sqrt{n_i}\,\overline{X}_i - \overline{X}_W}{S_1^*} \qquad .....(2-9)$$

$$b^* = Max\left|b_i\right| = Max\frac{\sqrt{n_i}\,\overline{X}_i - \overline{X}_W}{S_1^*} \qquad .....(2-10)$$

عندما

$$\overline{X}_W = \frac{1}{k}\sum_{i=1}^{k}\sqrt{n_i}\,\overline{X}_i$$

و

$$N = \sum_{i=1}^{k} n_i$$

ثالثاً: طريقـــة (Mcmillan) [20].

قدم (Mcmillan1970) طريقة لاختبار واحدة او اكثر من المشاهدات الشاذة لمجتمع يتوزع توزيعاً طبيعاً بمتوسط u وتباين σ^2 فقد أعتمد Mcmillan ثلاثة أجراءات لمعالجة البيانات المشكوك فيها والتي تحتوي على مشاهدات شاذة متعددة، الاجراء الاول عبارة عن تطبيق متسلسل لاختبار البواقي الاعظم (Maximum Residual Test) فاذا كانت القيمة المحسوبة في الصيغة (2.11) اكبر من الجدولة كما موضح في الصيغة فأن X_n يمكن اعتبارها مشاهدة شاذة ثم يتكرر الاختبار على بقية المشاهدات عندما $V_\alpha^{(n,v)}$ عبارة عن قيم حرجة يمكن الحصول عليها من جداول خاصة و S تمثل الانحراف المعياري للعينة. فاذا كانت ايضاً القيمة كما في الصيغة (2-12) اكبر من القيمة الجدولية فأن X_{n-1} يمكن اعتبارها مشاهدة شاذة ايضاً.

الاجراء الثاني لـ Mcmillan يتمثل في اعتبار القيمتين X_{n-1}, X_n شاذتين اذا كانت القيمة المحسوبة في الصيغة (2-13) اكبر من القيمة المحسوبة $\left(C_\alpha^n.S\right)$ عندما $C_\alpha^{(n)}$ تمثل قيمة حرجة معطاة في جداول خاصة.

اما الاجراء الثالث مشابه لما اقترحه Grubbs وهو كما في الصيغة (2-14) وتقارن هذه الاحصاءة مع القيم الحرجة في جداول خاصة وهذه الصيغ مدرجة أدناه.

$$X_n - \overline{X} > V_\alpha^{(n,V)}.S \qquad .....(2-11)$$

$$X_{n-1} - \overline{X} > V_\alpha^{(n-1,V)}.S_n \qquad .....(2-12)$$

عندما

$$S_n^2 = \sum_{i=1}^{n-1}\left(X_i - \overline{X}_n\right)\Big/ n-2$$

و

$$\overline{X}_n = \sum_{i=1}^{n-1} X_i \Big/ n-1$$

$$X_n + X_{n-1} - 2\overline{X} > C_\alpha^{(n)}.S \qquad .....(2-13)$$

$$\frac{S_{n,n-1}^2}{S^2} = \frac{\sum_{i=1}^{n-2}\left(X_i - \overline{X}_{n,n-1}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \qquad .....(2-14)$$

رابعـاً : طريقــة (Tietjen and Moore) [27]

اقترح كل من (Tietjen and Moore 1973) أحصاءتين L_k التي تعتمد على k من القيم (المشاهدات) الكبيرة المشكوك فيها في العينة (n) ، E_k التي تعتمد على البواقي العظمى (Largest Residuals) بقيمة مطلقة، فبعد ترتيب المشاهدات للعينة الطبيعية تصاعدياً وأختيار k من المشاهدات الكبيرة المشكوك فيها نستخدم الاحصاءة الاتيـة في الصيغة (2-15): واستخدام الاحصاءة L_k^* ايضاً في اختبار k من المشاهدات الصغيرة المشكوك فيها وبالصيغة المرقمة (2-16).

اما الاحصاءة E_k الصيغة (2-17) فقد اقترحت من قبل (Tietjen, Moore) لدعم وتأكيد على صحة ما تفرضه الاحصائتان L_k^*, L_k. والاحصائتان L_k^*, L_k تستخدمان على النحو الاتي اذا قررنا في عينة من

حجم n ان نختبر فيما اذا كانت k من القيم الكبيرة او الصغيرة المشكوك فيها شاذة ام لا فأننا نحسب L_k^*, L_k فاذا كانت قيمة هاتين الاحصاءتين اصغر من القيمة الحرجة المرغوبة فأننا نستنتج ان k من هذه القيم الكبيرة او الصغيرة هي بالفعل شاذة، واذا حسبنا E_k ثم قارناها بالقيمة الحرجة، فأذا كانت اصغر من القيم الحرجة المختارة فأن k من المشاهدات المشكوك فيها شاذة، وتقارن الاحصاءات E_k, L_k^*, L_k بما يقابلها من القيم الحرجة في جداول خاصة وهذه الصيغ مدرجة أدناه.

$$L_k = \frac{\sum_{i=1}^{n-k}\left(X_i - \overline{X}_k\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \qquad .....(2-15)$$

عندما

$$\overline{X}_k = \sum_{i=1}^{n-k} X_i/(n-k)$$

$$L^*_k = \frac{\sum_{i=k+1}^{n}\left(X_i - \overline{X}_k\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \qquad .....(2-16)$$

عندما

$$\overline{X}^*_k = \sum_{i=k+1}^{n} X_i/(n-k)$$

$$E_k = \frac{\sum_{i=1}^{n-k}\left(Z_i - \overline{Z}_k\right)^2}{\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2} \qquad .....(2-17)$$

عندما

$$\overline{Z}_k = \sum_{i=1}^{n-k} Z_i/(n-k)$$

و

$$Z_i = |di| = \left|X_i - \overline{X}\right|$$

خامساً : طريقــة (Rosner) [25].

اقترح (Rosner,1975) لتحديد واختبار اكثر من مشاهدة شاذة المشكوك فيها الصغيرة او الكبيرة او الاثنان معاً كعينة من التوزيع الطبيعي بحجم (n) والاختبار كما في الصيغة (18-2). ولاختبار معنوية الاحصاءات المحسوبة R_K -- الى R_1, R_2 تقارن مع القيم الحرجة في جداول خاصة [16].

$$R-statistic = \left|Max(X_i) - a\right| / b^2 \qquad .....(2-18)$$

عندما

$$a = \sum_{i=k+1}^{n-k} X_i \Big/ (n-2k)$$

و

$$b^2 = \sum_{i=k+1}^{n-k} (X_i - a)^2 \Big/ (n-2k-1)$$

$$I_o = (X_1, ...X_n)$$

$$R_1 = \left|X^{(o)} - a\right| \Big/ b, X^o \in I \ , \qquad .....(2-19)$$

$$I_1 = I_0 - X^{(o)}$$

$$R_k = \left|X^{(1)} - a\right| \Big/ b \qquad .....(2-20)$$

$$R_i = \left|X^{(k-1)} - a\right| \Big/ b \quad X_{(k-1)} \in I_{k-1} \qquad .....(2-21)$$

2.2.2 الطرق المعلميــة المستخدمة لتحديد القيــم الشاذة في حالة المتغيرين (Two Dimensions) ونماذج الانحدار.

اولاً : اختبار T^2 Hotelling .

اقترح AL-Bayati[6] في عام 1970 هذا الاختبار لمعرفة فيما اذا كان الزوج (X_o, Y_o) المشكوك فيه من المشاهدات شاذاً ام لا على النحو الاتي:-

لتكن n عينة من ازواج المشاهدات (X_i, Y_i) حيث ان $(i = 1,...n)$، مأخوذة من مجتمع طبيعي ثنائي (Bivariate Normal Distribution) له معالم محددة. اقترح AL-Bayati أختبار الفرضية الاتية

$$H_0 : P = P'$$

$$H_1 : P \neq P'$$

لاختبار الزوج المشاهد (X_o, Y_o) المشكوك فيه شاذة ام لا حيث أن P' يمثل معامل الارتباط محسوب بدون القيمة المشكوك فيها (X_o, Y_o)، وباستخدام احصاءة القطع الناقص (E) كما في الصيغة (22-2).وتقديرها للعينة $(\hat{E})$كما في الصيغة (23-2). والاحصاءة $\hat{E}$ لها توزيع T^2 Hotelling المستخدم عادة لآختبار متجه من المتوسطات ل P من المتغيرات المرتبطة هنا في هذه الحالة P=2 ، وباستخدام العلاقة بين ال T^2 Hotelling وتوزيع F عند مستوى α، والتي هي كما في الصيغة (24-2) وأن قيمة F_α يمكن الحصول عليها من جداول F بمستوى معنوية α ودرجة حرية (n-2) وعند اختبار ازواج القيم مثلاً (X_o, Y_o) فأن الاحصاءة (23-2) تصبح كما في

الصيغة (2-25)، ونقارن على النحو التالي، اذا كانت $T_\alpha^2 > T^2$ نرفض (X_o, Y_o) على انها مشاهدة شاذة واذا كانت $T_\alpha^2 \le T^2$ تقبل (X_o, Y_o) على انها واحدة من المشاهدات الشاذة وهذه الصيغ مدرجة أدناه.

$$E = \frac{1}{1-p^2}\left[\frac{(X-u_x)^2}{\sigma_x^2} + \frac{(Y-u_y)^2}{\sigma_y^2} - \frac{2p(X-u_x)(Y-u_y)}{\sigma_x \sigma_y}\right] \quad \ldots\ldots\ldots(2-22)$$

وتقديرها للعينة

$$\hat{E} = \frac{1}{1-r^2}\left[\frac{(X-\bar{X})^2}{S_x^2} + \frac{(Y-\bar{Y})^2}{S_y^2} - \frac{2r(X-\bar{X})(Y-\bar{Y})}{S_x S_y}\right] \quad \ldots\ldots\ldots\ldots(2-23)$$

$$T^2 = \frac{P(n-p)}{n-(p+1)} F_\alpha \quad \ldots\ldots\ldots\ldots(2-24)$$

وبوضع $p = 2$ تصبح

$$T^2 = \frac{2(n-2)}{n-3} F_\alpha$$

$$T_\alpha^2 = \frac{1}{1-r^2}\left[\frac{(X_0-\bar{X})^2}{S_x^2} + \frac{(Y_0-\bar{Y})^2}{S_y^2} - \frac{2r(X_o-\bar{X})(Y_o-\bar{Y})}{S_x S_y}\right] \quad \ldots\ldots\ldots\ldots(2-25)$$

ثانياً : القيــم الشاذة في نماذج الانحدار (Outliers in a Regression Madels)

a- القيم الشاذة في النموذج الخطي البسيط[4] Outliers in a Regression Models تصاغ معادلة النموذج الخطي البسيط كما يلي

$$Y_i = B_o + B_1 X_i + U_i \quad \ldots\ldots(2-26)$$

حيث Y_i يمثل المتغير المعتمد (Dependent Variable) في النموذج و χ_i يمثل المتغير المستقل (Independent Variable) في النموذج، B_1, B_o تمثل معالم (Parameters) النموذج و U_i يمثل الخطأ العشوائي (Random Error) وان

$$U_i \sim N(O, \sigma^2), \forall i = 1,\ldots n \quad \text{cov}(\mu_i, \mu_j) = o \qquad \forall i \ne j$$

فاذا قدرنا القيم B_1, B_o باحدى طرق التقدير ولتكن طريقة المربعات الصغرى (Least Square Estimations) ستكون صيغة المعادلة (2-26) على النحو الاتي:

$$e_i = y_i - b_o - b_1 X_i \quad \ldots\ldots\ldots\ldots(2-27)$$

للبحث عن القيم الشاذة في الانحدار الخطي البسيط فأنه من المناسب دراسة واختبار حجوم e_i وخواصهما والصيغة (2-28) لحساب التباين $V(e_i)$ الى الخطأ. ونستنتج ان القيم الشاذة لها تأثير كبير على الانحراف المعياري للاخطاء ولهذا فأن اختبار تباعد القيم يعتمد على الاخطاء وكما في الصيغة (2-29) وهذه الصيغ مدرجة أدناه .

$$V(e_i)=\sigma^2\left[\frac{n-1}{n}+\frac{(X_i-\overline{X})^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right] \qquad .....(2-28)$$

$$\frac{e_i}{s_i}=e_i/s\sqrt{\frac{n-1}{n}+(X_i-\overline{X})^2\Big/\sum_{i=1}^{n}(X_i-\overline{X})^2} \qquad .....(2-29)$$

وقد اقترح Daniel[11] في عام 1960 استخدم الصيغة |Max ei| / s (Maximum Absolute studentized Residual) اساساً لتحديد واختبار تباعد قيمة شاذة واحدة ... في حين اتخذ كل من Behnken and Draper[8] في عام 1972 الاحصاءة في المعادلة المرقمة (2-30) لاختبار تباعد قيمة شاذة واحدة في النموذج البسيط، فاذا كانت قيمة الاحصاءة في المعادلة (t) كبيرة بشكل كاف، عند ذلك سنحكم على المشاهدة انها قيمة شاذة متباعدة (Discordant Outlier) في حين ان الاستخدام العلمي لهذه الاحصاءة يقضي بمعرفة توزيعها وهذا امر يصعب تحقيقه، لذلك اعتمد كل من Moore and Beckman اسلوب المحاكاة حيث قاما بعمل الالاف من التجارب لعينات يصل حجمها n =100 وقد استخدما مستويات المعنوية $(\alpha = 0.1, 0.05 , 0.01)$ وفي النهاية استطاعا تحديد القيم الحرجة للاحصاءة t واستنتجا ان القيم التي سنحصل عليها عند افتراض تساوي تباينات البواقي سوف لا تختلف كثيراً عن القيم التي ستحصل عليها بعد ان تؤخذ حالة عدم تساوي تباينات الاخطاء بنظر الاعتبار كما في الصيغة (30- 2) .

$$t=Max|e_i/S_i \qquad .....(2-30)$$

ثم اقتراح Prescott[17] في عام 1975 أهمال التباينات المختلفة للاخطاء المقدرة e_i واحلال $\overline{S}$ محل S_i في الاحصاءة للمعادلة (2-30) حيث $\overline{S}$ تمثل معدل التباين.

وبين كل من Behken and Draper[8] في عام 1972 ان معدل التباين للاخطاء غير المتحيز هو $(n-q)\sigma^2/n$ حيث ان (q) تمثل عدد المعالم (Parameters) في النموذج (وهنا $q=2$) وعليه فأن تقدير معدل التباين سيكون كما في المعادلة (2-31) وبناءاً على ذلك فأن (t^*) بعد الاشتقاق سوف تكون كما في الصيغة المرقمة (2-32). اما بالنسبة للقيم الحرجة للاحصاءة تبين t^*,t فيمكن الحصول عليها من جداول خاصة والصيغ كما مدرجة أدناه (وللمزيد من التفاصيل انظر ديلر 1996).

$$\overline{S} = (n-q)S^2/n \qquad .....(2-31)$$

$$t^* = Max|e_i| \Big/ \sqrt{(n-q)S^2/n}$$

$$= Max|e_i| \Big/ \sqrt{\sum_{i=1}^{n} e_i^{\,2}/n}$$

$$= \sqrt{n}\ \ Max|e_i| \Big/ \sqrt{\sum_{i=1}^{n} e_i}$$

$$t^* = \sqrt{n}\ \ Max|z_i| \qquad .....(2-32)$$

$$Z_i = e_i \Big/ \sqrt{\sum_{i=1}^{n} e_i^{\,2}}$$

b- القيم الشاذة في النموذج الخطي العام [4]

Outliers in a General Linear Models.

تعرف معادلة النموذج الخطي العام بصفة المصفوفات كالاتي:-

$$\underline{Y} = \underline{X}\,\underline{B} + \underline{U} \qquad .....(2-33)$$

حيث ان

$\underline{Y}$ **: يمثل متجه المتغير المعتمد**

(Vector of Dependent Random variables) بدرجة $n \times 1$

$\underline{X}$ **: تمثل مصفوفة المتغيرات المستقلة**

(Matrix of Independent Random variables) بدرجة $[n \times (k+1)]$

$\underline{B}$ **: يمثل متجه معالم النموذج (Vector of Parameters) بدرجة** $[(k+1) \times 1]$

$\underline{U}$ **: يمثل متجه الاخطاء العشوائية (Vector of Random Errors) بدرجة** $(n \times 1)$**.**

أن تقدير معالم النموذج B للمعادلة (2-33) بطريقة المربعات الصغرى هو كما في المعادلة (2-34). وتقدير تباينه كما في صيغة المعادلة (2-35). اما تقدير الخطأ (e_i) **فهو كما في المعادلة (2-36) وعليه فأن تقدير تباين الخطأ يكون كما في صيغة المعادلة (2-37) ومن المعادلة (2-38) نلاحظ ان الاخطاء المقدرة** (e_i) **ستمتلك تباينات مختلفة مع وجود ارتباطات فيما بينها ويمكن توضيحها كما في الصيغة (2-39) وبناء على ذلك يمكن تقدير مصفوفة التباين والتباين المشترك للخطأ V-COV(e) كما في المعادلة (2-40) وسيكون التباين المقدر الى الخطأ لـ** (e_i) **كما في المعادلة (2-41) والمعادلات مدرجة كما في الصيغ أدناه. (وللتفاصيل ارجع الى ديلر 1996).**

$$\hat{B} = (X'X)^{-1}X'Y \qquad .....(2-34)$$

$$V = Cor(\underline{\hat{B}}) = (X'X)^{-1}\sigma^2 \qquad .....(2-35)$$

$$\underline{e} = [I-R]\underline{U} \qquad .....(2-36)$$

وعليه فأن تقدير تباين الخطأ يكون

$$V-Cor(\underline{e})=(I-R)\sigma^2 \qquad .....(2-37)$$

من المعادلة (2-37) نلاحظ ان الاخطاء المقدرة e_i ستمتلك تباينات مختلفة مع وجود أرتباطات فيما بينها ويمكن توضيحها كالاتي

$$Var(e_i)=\left[1-X_i'(X'X)^{-1}X_i\right]\sigma^2$$
$$=(1-r_{ii})\sigma^2 \qquad .....(2-38)$$

حيث يمثل X_i' الصف i من المصفوفة X وان

$$r_{ii}=X_i'(X'X)^{-1}X_i$$

و σ^2 مجهولة القيمة بصورة عامة وان التقدير غير المتحيز لها هو

$$S_e^2=\underline{e}'\,\underline{e}/(n-q)$$
$$=\underline{U}'[\underline{I}-\underline{R}]\underline{U}/(n-q) \qquad .....(2-39)$$

حيث ان : الكمية $[I-R]$ تمثل مصفوفة صماء [Idempotency Matrix] وبناء على ذلك يمكن تقدير ($\underline{e}$) V-Cov كالاتي:

$$S_e^2=(I_n-r)\hat{\sigma}^2 \qquad .....(2-40)$$
$$S_i^2=(I-r_{ii})\hat{\sigma}^2$$
$$=(I-r_{ii})\underline{e}'\,\underline{e}/(n-q) \qquad .....(2-41)$$

وبما أنه الاخطاء المقدرة (e_i) تمتلك تباينات مختلفة فيجب استخدام صيغة اخرى تأخذ بنظر الاعتبار عدم مساواة التباين، والصيغة المناسبة هي $M_i=e_i/s_i$ وتعتمد اغلب الاحصاءات المقترحة على (Maximum studentized Residual) حيث تعتبر المشاهدة التي تعطينا اكبر قيمة من قيم (m_i) مشاهدة شاذة متباعدة.

أحصاءة Strikantan [26] عام 1961 احدى الاحصاءات المقترحة لهذا الغرض ووصيغتها كما في المعادلة (2-42). وتقارن الاحصاءة T لهذه المعادلة مع القيم الحرجة في جداول خاصة اعدها Lund[19] عام 1975 لهذا الغرض.
وسيكون التباين المقدر لـ e_i مساوي الى

$$T=Max|e_i|S_i|>n_\alpha \qquad .....(2-42)$$

2.3 **الطرق الاستكشافيـــة**. Exploratory Data Analysis

يتم التطرق في هذا المبحث الى ثلاث طرق من الطرق الاستكشافية الحديثة (Exploratory Data Analysis)التي تستخدم لتحديد القيم الشاذة في البيانات وهي طريقة الغصن والورقة (steam and leaf)، الرسم الصندوقي (Box plot) ، وطريقة الرسم الصندوقي المزدوج للمتغيرين (χ,y) (Rangefinder Box plot).

1.3.2 الغصـــن والورقـــة (Steam and Leaf)

من الطرق الحديثة لعرض البيانات الاحصائية هي طريقة الغصن والورقة، وهذا الاسلوب من العرض اسهل عند تكوينه من جداول التوزيع التكراري وكذلك من المدرج التكراري وبصورة عامة فهو يعرض معلومات اكثر.

فهو يعرض نفس معلومات المدرج التكراري (histogram) وكذلك يعرض معلومات جداول التوزيع التكراري بالاضافة الى ذلك فهو يعرض الارقام بشكلها الاعتيادي عند ربط كل قيمة بين الغصن والورقة الخاص بها لهذا يسمى بالشكل الهجين (hybrid) لأن معلوماته تمثل الرسم وكذلك الارقام في الجدول في أن واحد كما هو موضح لاحقآ. لتوضيح الغصن والورقة لاحظ المثال التالي:-

مـــثال1 : اوجد الغصن والورقة للمشاهدات التالية

((100 98 70 39 78 69 32 44 90 89 60 200 40 50 10 6))

الحـــل : بناء او تكوين الغصن والورقة والذي هو في نفس الوقت يوضح البيانات على شكل مجاميع تشبه جدول التوزيع التكراري وكذلك يعرضها على شكل رسم يشبه المدرج التكراري تقوم بالخطوات التالية:-

1. نختار ارقام من البيانات على انها الارقام التي تكون امام البيانات او تقود البيانات (Leading digits) والتي تشمل الجزء الاول مـــن الارقام والتي تمثـل العشرات وهذا ينتج الارقـــام التاليـــة (10 .....6 5 4 3 1 0)

نضعهم على شكل عمود كما في الشكل اللاحق (الغصن steam).

2. بعدها نبدأ بالمرور على جميع الارقام لكتابة الجزء الثاني من كل رقم مقابل الجزء الاول منه (Final digit) ونضعه الى اليمين من الجزء الاول وهو الاحاد.

الرقم الاول في المثال هو (6)، ولهذا نحتاج ان نضع الرقم (6) على يمين الرقم (0) ، وبعدها نقرأ الارقام في المثال ولهذا الرقم الثاني هو (10) ولهذا نحتاج لوضع (0) الى يمين الرقم (1) ونستمر على هذه الطريقة نضع (0) الى يمين الرقم (5) وهكذا نستمر.

ولهذا سوف يكون شكل الغصن والورقة لبيانات المثال(1) كما موجود في الشكل التالي. وكما موضح في الشكل الارقام التي تتبعها الارقام الاخرى على اليمين تدعى الارقام البداية الغصن (stem) والارقام النهائية او المكملة تسمى الاوراق (Leaves) والشكل هو (stem and Leaf) او الغصن والورقة.

الشكـل رقـم (1)

```
Stem-and-leaf of C1          N   = 16
          Leaf Unit = 1.0
     1     0 6
     2     1 0
                        2     2
     4     3 29
     6     4 04
     7     5 0
   (2)     6 09
     7     7 08
     5     8 9
     4     9 08
     2    10 0
     HI   200;
```

ويمكن ملاحظة اكبر رقم في الشكل رقم(1) الرقم (200) منفصل هذه القيمة يحددها الغصن والورقة على انها قيمة شاذة وبمزيد من التفاصيل أرجع ديلر .1996

2.3.2 الرســـم الصندوقـــي (Box plod)

الرسم الصندوقي box plot يعتبر من افضل الرسوم الاحصائية لعرض البيانات الاحصائية ولاجراء المقارنات بين عدة مجاميع من البيانات والذي اكتشف من قبل العالم (Tukey 1977).

ان الفكرة الاساسية للرسم الصندوقي بسيطة وهي عرض بالرسم لخمس قيم تؤخذ من البيانات تسمى الملخصات الخمسة (5-Number Summeries) انظر الشكل رقم (6-2) وهي كما يلي (Lower Extreme) وهي اصغر قيمة ترتبط بل الصندوق غير شاذة، قيمة الربيع الاول او الربيع الادنى (Q_1 = Lower Quartile)، قيمة الوسيط (Q_2 = Median)، قيمة الربيع الثالث او الربيع الاعلى (Q_3 = Upper Quartile)، وكذلك اكبر قيمة غير شاذة مرتبطة مع الصندوق (Upper Extreme)، القيم التي تكون اكبر من اكبر قيمة تعتبر قيم شاذة وكذلك القيم التي هي اصغر من اصغر قيمة تعتبر قيم شاذة ولهذا كل القيم الشاذة تؤشر بشكل منفصل اذا كانت كبيرة او صغيرة.

اما كيفية بناء الرسم الصندوقي او ابعاد الصندوق فهي كما يلي، بعد ترتيب البيانات تصاعدياً.

1. طول الصندوق او المستطيل فهو الفرق بين الربيع الثالث Q_3 والربيع الاول Q_1 أي $(Q_3 - Q_1 =$ IQR) ويمثل المدى الربيعي (IQR= Interquartile Range) والذي يمثل 50% من البيانات.
2. الخط داخل المستطيل او الصنوق يمثل الوسيط (Median) الى البيانات.
3. الخط الذي يوصل الصندوق با (Upper Extreme , Lower Extreme) يسمى (Whisker Lengths) وطوله بالنسبة الى الادنى يمثل (Q_1 - 1.5*IQR) وطوله بالنسبة الى الاعلى فهو (Q_3 + 1.5*IQR) القيم التي تقع خارج هذين الامتدادين من الادنى والاعلى تؤشر بشكل منفصل وتسمى قيم شاذة (Outlier values).
4. عرض الصندوق(box width) فقد استخدمTukey قاعدة تتناسب مع الجذرالتربيعي لحجم العينة $\left(\sqrt{n}\right)$.

5. حدود الشقة حول الوسيط (notch) وبمستوى 5% تحسب وفقاً للقانون التالي.

$$M_e \pm \frac{1.58(IQR)}{\sqrt{n}}$$

حيث ان:

n **: يمثل حجم العينة.**

M_e **: الوسيط للبيانات.**

ويستخدم الرسم الصندوقي لاجراء المقارنات بين عدة مجاميع في ان واحد وكذلك لتحديد القيم الشاذة في البيانات فكل قيم تقع خارج (Upper extreme and lower extreme) تعتبر قيم شاذة،وكما هو واضح من الرسم الصندوقي (Box plot ,Example no.2) في نهاية البحث شكل رقم (1AA) وللاطلاع على التفاصيل (ارجع الى ديلر 1996).

4-3-2 **الرسم الصندوقي المزدوج** (Range Finder Box plot)

يعتبر الرسم الصندوقي المزدوج من احدث الطرق الاحصائية لتحليل العلاقة بين توزيعي المتغيرين لشكل الانتشار(Scatter plot) وتحديد الاجزاء المهمة لهما وكذلك القيم الشاذة لكل منهما في نفس الوقت. وهو مفيد جداً حيث يربط بين شكل الانتشار الاعتيادي والرسم الصندوقي الاعتيادي ويجمع المعلومات الخاصة بالشكلين، والرسم الصندوقي المزدوج (Rangefinder Box plot) يمثل رسمين صندوقين متقاطعين في نقطة الوسيط ولكل رسم صندوقي هناك ثلاث خطوط تمثل الرسم الصندوقي ثلاثة للمتغير المعتمد (Y) وثلاثة للمتغير المستقل (X).

ويمكن توضيح فكرة الرسم هو هناك خطين وسطيين (عمودي للكل متغير) طول كلا منهما يمثل طول الصندوق في الرسم الصندوقي الاعتيادي ويتقاطعان في موقع الوسيط وهناك فراغ يمثل طول الامتداد الى اصغر قيمة واكبر قيمة (whisker lengths) وهناك خطان افقيان لكل صندوق موقعهما يمثل اكبر قيمة واصغر قيمة (Extreme values) لكل صندوق وطولهما يمثل عرض الصندوق لكل متغير وهذان الرسمان الصندوقيان العمودي للمتغير (y) والافقي للمتغير (x). أي قيمة تقع خارج (Extreme values) لكل متغير او رسم صندوقي تؤشر على انها قيمة شاذة وتكون بشكل منفرد، وسوف يكون هناك رسم صندوقي مزدوج في الجانب العملي في الفصل التالي وللتفاصيل رجع الى (ديلر 1996).

3ـ الجانـــب التطبيقـــي

3ـ1 المقدمـــة:ـ

سوف نتطرق في هذا الجانب الى استخدام الطرق السابقة لتحديد القيم الشاذة (Outlier values) في البيانات واجراء المقارنة بين هذه الطرق جميعها طرق الرسم والطرق المعلمية (parametric methods) المستخدمة لتحديد القيم الشاذة في البيانات،الجانب العملي يكون باستخدام الامثلة المستخدمة في الدراسات السابقة، ثم تطبيق طرق الرسم الحديثة الاستكشافية (Exploratory Data Analysis) على تلك الامثلة لتحديد القيم الشاذة فيها وتحديد الفرق بينهما ومقارنة النتائج.

3ـ2 القيم الشاذة في حالة المتغير الواحـــد

هناك ست امثلة في الاطروحة تمثل البيانات في حالة المتغير الواحد سوف نتناول ثلاثة منهم فقط لتحديد المجال الى البحث عند النشر، ومن ثم تحديد القيم الشاذة لكل مثال اولاً باستخدام الطرق المعلمية، وثانياً باستخدام الطرق الاستكشافية.

مثال رقم (2) 0.39 -0.05 -0.13 -0.24 - 0.44 0.18 -0.22 0.63 0.48 - 0.3
1.01 0.06 -1.4 0.20 0.10

a- تحديد القيــم الشاذة باستخــدام الطرق المعلمية.

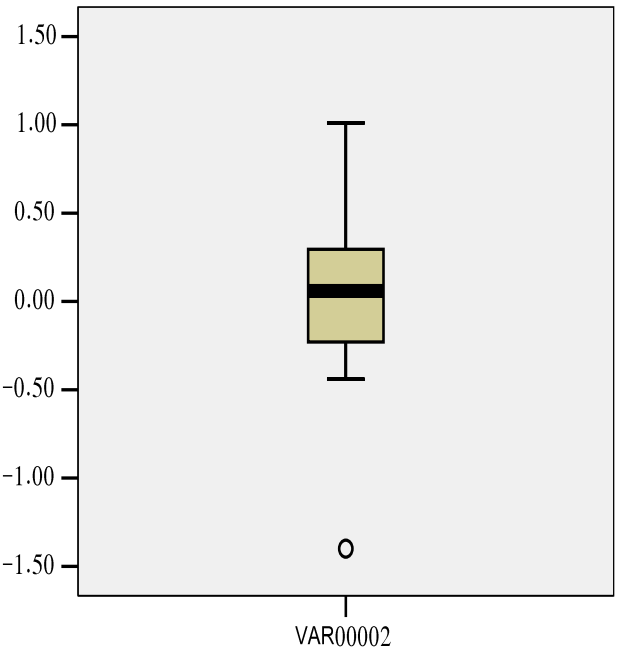
استخدام AL-Jobouri [16] 1976 طريقتين لاختبار وتحديد القيم الشاذة في المثال رقم (2) الطريقة الاولى طريقة Grubbs وقد وجد من نتيجة الاختيار بأن القيمة (-1.40) شاذة، اما الطريقة الثانية فهي طريقة Tietjen and moore وباستخدام الاحصاءة (E_k) وقد وجد بأن القيم (-1.40) و (1.01) هي قيم شاذة. كما استخدم الباحث طريقة Rosner وهي طريقة معلمية أخرى لتحديد القيم الشاذة لنفس المثال رقم (2) وقد طبق على القيمتين (-1.40) ,(1.01)وقد وجد بأن (R_1=2.573) وان (R_2=2.218) والقيم الحرجة للاحصاءتين R_2,R_1 على التوالي هي 2.72*,2.68* وبالمقارنة مع القيم المحسوبة لاتعتبر اياً من القيمتين معنوية وهذا يعني لا توجد قيم شاذة في هذا المثال حسب اختبار Rosner.

b- تحديد القيم الشاذة باستخدام الطرق الاستكشافية.

اولاً : بطريقـة الرسـم الصندوقـي (Box plot)

من خلال النظر الى الشكل رقم (2) التالي وهو الرسم الصندوقي لهذا المثال نجد ان هناك قيمة شاذة واحدة وهي (1.40-) . شكل رقم (2) حيث تم رسمها بشكل منفصل خارج الرسم الصندوقي.

شكل رقم(2)



ثانياً: باستخدام طريقة الغصن والورقة وهي موضحة في الشكل رقم (3)

شكل رقم(3)

```
Stem-and-leaf of C1          N
= 15

                    Leaf Unit = 0.10
                    LO  -14;
                     2    -0 4
                     5    -0 322
                     7    -0 10
                    (3)    0 011
                     5     0 23
                     3     0 4
                     2     0 6
                     1     0
                     1     1 0
```

ايضاً باستخدام هذه الطريقة نلاحظ ان هناك قيمة شاذة واحدة فقط وهي (1.40-) حيث تم فصلها عن باقي البيانات وقد اشرت تحت حقل (LO–1.40) وهذا يعني وجود تطابق بين الطرق الاستكشافية في تحديد القيم الشاذة وتختلف عن الطرق المعلمية كما هناك اختلاف ايضاً بين الطرق المعلمية في تحديد القيم الشاذة فحسب طريقة Grubbs هناك قيمة شاذة واحدة وهي (1.40-) وقيمتان شاذة وهما (1.40-) و(1.01) حسب طريقة Tietjen and Moore ولا توجد قيم شاذة حسب طريقة Rosner وهذا يعني ان الاختلافات واضحة بين الطرق المعلمية بعكس الطرق الاستكشافية بالرسم فهي واضحة جداً ومتطابقة في هذا المثال.

بيانات المثـال رقـم (3)

-1.056	-1.008	-0.34	0.533	0.109	0.661	1.638	-0.413	-0.667	-0.57
1.207	-0.550	2.290	0.504	-2.215	2.139	-0.048	-0.909	0.967	-0.143

a- تحديد القيم الشاذة باستخدام الطرق المعلمية.

استخدام AL-Jobouri[16] عام 1976 طريقة Rosner لتحديد القيم الشاذة في المثال رقم 3 اعلاه، وقد وجد بأنه لاتوجد اية قيمة من القيم شاذة باستخدام هذه الطريقة.

في حين استخدام الباحث اضافة الى ذلك ثلاث طرق معلمية اخرى لتحديد القيم الشاذة في المثال رقم 3 وعلى النحو التالي:-

1. الطريقة الاولى:

باستخدام الاحصاءة التي افترضها Mcmillan لاختبار اعلى قيمتين في البيانات وهي:

$$X_n + X_{n-1} - 2\overline{X} > C_\alpha^n \, S$$

$$(2.29 + 2.139) - 2(0.1064) > (1.145)(0.637)^*$$

$$4.2101 > 0.729$$

وبذلك تعتبر القيمتان 2.139 , 2.2906 شاذتان حسب اختبار Mcmillan.

2-الطريقة الثانية :- باستخدام اختبار Grubbs
لاختبار اعلى قيمتين 2.139 , 2.290

نستخدم الاحصاءة

$$\frac{S^2_{n,n-1}}{S^2} = \frac{15.0129}{24.899} = 0.6029$$

ولاختبار ادنى قيمتين 1.056- , 2.215-

$$\frac{S^2_{1,2}}{S^2} = \frac{21.896}{24.896} = 0.879$$

وبالمقارنة مع القيمة الحرجة للاحصاءتين اعلاه عند مستوى $\alpha = 0.05$ والتي تساوي 0.5269 ، نجد بأنا القيم اعلاه ليست شاذة.وهذا يعني عدم وجود قيم شاذة في المثال رقم (3) باستخدام طريقة Grubbs المعلمية.

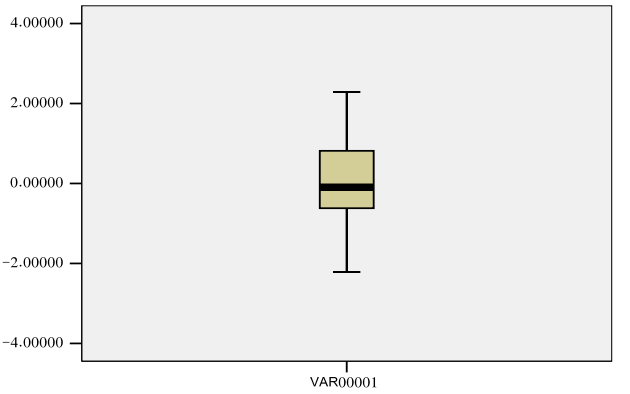
3- الطريقة الثالثة باستخدام اختبار Tietjen and Moore

وقد وجد بأن الاحصاءة $(E_2 = 0.537)$. وبالمقارنة مع القيمة الحرجة والتي تساوي $(E^*_2 = 0.416)$[16]، أي ان القيمة غيرمعنوية وهذا يعني لاتوجد قيم شاذة للمثال رقم(3) اعلاه باستخدام هذه الطريقة.

b- تحديد القيم الشاذة للمثال رقم (3) باستخدام الطرق الاستكشافية.

اولاً: بطريقة الرسم الصندوقي الشكل رقم (4) لا توجد قيم شاذة لهذه البيانات للمثال رقم (3).

شكل رقم(4)



ثانياً: بطريقة الغصن والورقة (Stem and Leaf)

وهي موضحة في الشكل رقم (5) للغصن والورقة ونلاحظ من خلال الشكل لاتوجد قيم شاذة أيضاً.

وهنا توجد فروقات بين الطرق المعلمية وطرق الرسم في تحديد القيم الشاذة، حيث نرى بوضوح عدم وجود قيم شاذة للمثال رقم (3) اعلاه باستخدام ثلاث طرق معلمية مختلفة في حين هناك قيمتان شاذتان باستخدام اختبار Mcmillan اما بالنسبة الى طرق الرسم فنلاحظ هناك تطابق بين طريقة (stem and leaf) و طريقة الرسم الصندوقي (Box plot) حيث لاتوجد قيم شاذة.

شكل رقم (5)

```
                                   Stem-and-leaf     of     C1
N  = 20
                                                 Leaf  Unit  =
0.10

                                                       1    -2
2
                                                       1   -1
                                                       3    -1
00
                                                       7    -0
9655
                                                      (4)   -0
4310
                                                       9     0
1
                                                       8     0
5569
                                                       4     1
2
                                                       3     1
6
                                                       2     2
12
```

بيانات المثـــال رقم (4)

4.57 5.62 4.12 5.29 4.64 4.31 4.30 4.39 4.45 5.67 4.39 4.52 4.26
4.26 4.40 5.78 4.73 4.56 5.08 4.41 4.12 5.51 4.82 4.63 4.29 4.60

a- تحديد القيم الشاذة باستخدام الطرق المعلمية.

تم استخدام ثلاث طرق معلمية لتحديد القيم الشاذة في المثال اعلاه وعلى النحو التي:-

اولأ: باستخدام طريقة (Grubbs)

لاختبار القيم العليا : تأخذ اعلى قيمة 5.78

وباستخدام الاحصاءة

$$\frac{S_n^2}{S^2} = \frac{4.91}{6.165} = 0.79$$

ولاختبار القيم الصغرى: لنأخذ اصغر قيمة 4.12 وحساب

$$\frac{S_1^2}{S^2} = \frac{5.837}{6.165} = 0.94$$

وبالمقارنة مع القيمة الحرجة 0.71 نرى بأن القيمتين 4.12 ,5.78 ليست شاذتين وهذا يعني عدم وجود قيم شاذة في مثال رقم 4 اعلاه باستخدام اختبار Grubbs.

ثانياً: **باستخدام اختبار**(Tietjen and Morre) :

وحساب

$$R_1 = |5.78 - 4.681| / 0.49$$

$$R_1 = 2.24$$

وبالمقارنة مع الحد الاعلى للاحصاءة وهي $R_1^* = 2.93$ **وهذا معناه عدم وجود قيم شاذة في هذا المثال حسب الاختبار (Rosner).**

b- تحديد القيم الشاذة باستخدام الطرق الاستكشافية

اولاً: باستخدام طريقة الرسم الصندوقي.

ابعاد الرسم الصندوقي هي:-

الوسيط (Median) – 4.54

الربيع الاول (Q_1) – 4.31

الربيع الثالث (Q_3)- 4.82

ادنى قيمة غير شاذة (L) – 4.12

اعلى قيمة غير شاذة (U)- 5.51

والشكل رقم (6) يمثل الرسم الصندوقي لبيانات المثال رقم (4).

شكل رقم (6)

وكما مبين من الشكل اعلاه وجود ثلاث قيم شاذة يمكن ملاحظتها بوضوح تم رسمها بشكل منفصل خارج الرسم الصندوقي وهي (5.62 ,5.67 , 5.78).

ثانياً: باستخدام طريقة الغصن والورقة وهي موضحة بالشكل رقم (7)

شكل رقم (7)

```
Stem-and-leaf of C1        N   =
26

Leaf Unit = 0.10
    2     4 11
    9     4 2223333
  (6)     4 444555
   11     4 6667
    7     4 8
    6     5 0
    5     5 2
    4     5 5

                HI      56;
56; 57;
```

ايضاً باستخدام هذه الطريقة هناك ثلاث قيم شاذة حيث تم فصلها عن باقي البيانات وقد حددة تحت حقل (HI)، وهي نفس القيم ظهرت باستخدام الرسم الصندوقي ، وهذا يعني هناك تطابق بين الطرق الاستكشافية لتحديد القيم الشاذة وهناك تطابق أيظآ بين الطرق المعلمية لتحديد القيم الشاذة حيث لم تؤشر اياً منها هناك قيم شاذة وهذا يوضح لنا هناك فروقات كبيرة بين الطرق المعلمية والطرق الاستكشافية لتحديد القيم الشاذة.

3.3 : القيم الشاذة في حالة المتغيرين ونماذج الانحدار

سنتناول في هذا المبحث مثالين من مجموع اربع امثلة تمثل البيانات في حالة المتغيرين، ثم تحديد القيم الشاذة لكل مثال باستخدام الطرق المعلمية اولاً والطرق الاستكشافية ثانياً.

بيانات المثال رقم (5)

X: 3.25 3.79 3.4 3.68 3.76 3.3 2.4 2.94 2.68 3.27 2.46 3.46 2.14 3.22 2.91 1.98

Y: 2.85 3.91 2.33 3.29 3.71 3.52 4.07 2.98 3.18 2.72 2.41 3.01 2.56 2.89 2.64 2.22

a- تحدد القيم الشاذة لبيانات المثال رقم(5) باستخدام الطرق المعلمية.

استخدم AL-Jobouri [16] عام 1976 اختبار البياتي (Hotelling T^2 test) لتحديد القيم الشاذة في بيانات المثال (رقم 5) فقد وجد من نتيجة الاختبار وجود زوج من قيم المشاهدات شاذة وهي (2.4,4.07) عن بقية قيم المشاهدات كما استخدم الباحث طريقة معلمية اخرى لاختبار وجود القيم الشاذة باستخدام الاحصاءة $t=Max\ |ei/s_i|$ وعلى النحو التالي:-

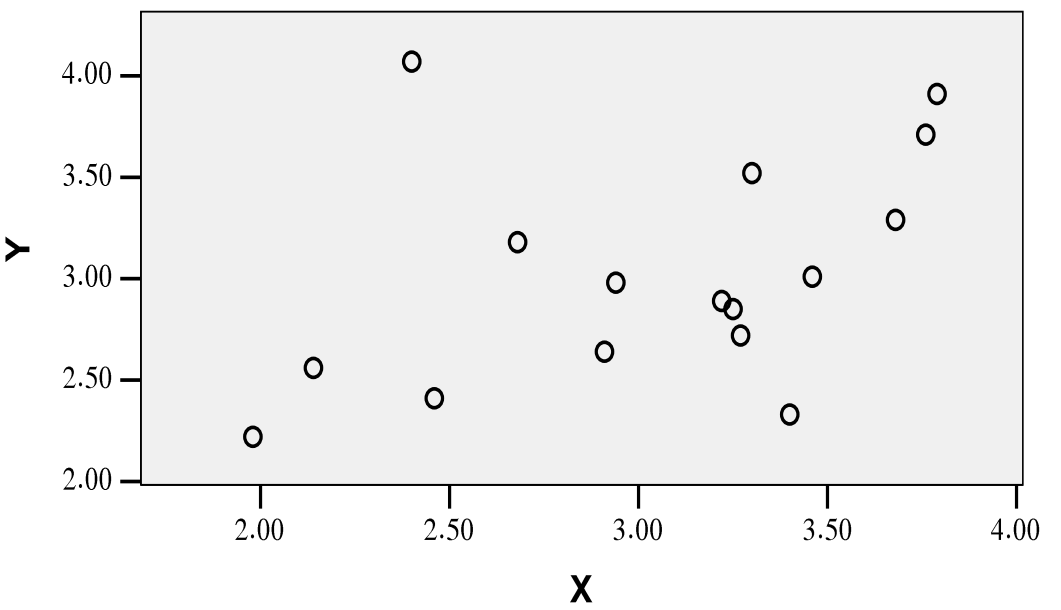
لحساب قيم e_i تحتاج الى تقدير معادلة خط الانحدار وتقديرها بطريقة المربعات الصغرى.

$$\hat{y}_i = 1.9 + 0.36 X_i$$

وبعد حساب قيـم $|e_i / s_i|$ وجد الباحث بأن أكبر قيمة هي 2.61 وهـي القيمـة التي تقابـل زوج المشاهدات (x = 2.4, y = 4.07) وبالمقارنـة مع القيمة الحرجة للاحصـاءة اعلاه والتـي تساوي $(t^* = 2.6)^{16}$ وبناءاً على ذلك يعتبر زوج القيم (2.4, 4.07) شاذاً عن باقي القيم حسب هذا الاختبار وهي نفس النتيجة التي حصل عليها الجبوري في اختباره.

b- تحديد القيم الشاذة باستخدام طريقة الرسم الصندوقي المزدوج للمتغيرين (x,y) (Rangefinder Box plot) لبيانات المثال رقم (5) اعلاه وهي موضحة بالشكل رقم (8) التالي.

شكل رقم (8)



ونلاحظ من الشكل اعلاه الرسم الصندوقي المزدوج للمتغيرين (x, y) عدم وجود اية قيمة شاذة للمتغيرين.

وهذه النتيجة تختلف عن نتيجة الطرق المعلمية حيث لاحظنا ان الطريقتين المعلمتين اعطت نفس النتائج وهي وجود قيمتين شاذة ، وهذا مؤثر قوي على انه هناك فروقات معنوية بين الطرق المعلمية والطرق الاستكشافية في تحديد القيم الشاذة.

بيانات المثـال رقـم (6)

X : 51.3 49.9 50 44.2 48.5 47.8 47.3 45.1 46.3 42.1 44.2 43.5 42.3 40.2 31.8 34.0

Y : 102.5 104.5 100.4 95.9 87.0 95 88.6 29.2 78.9 84.6 81.7 72.2 65.1 68.1 67.3 52.5

a- تحديد القيم الشاذة لبيانات مثال رقم (6) باستخدام الطرق المعلمية:-

استخدم الباحث طريقتين لاختبار وجود القيم الشاذة، في هذا المثال وهي كما يلي: الطريقة الاولى باستخدام اختبار البياتي وخطواته على النحو التالي:

اولاً: ايجاد قيمة $\hat{E}$

وعند اختبار القيمة $(X_0 = 31.8, Y_0 = 67.3)$

فأن $\hat{E} = 8.86$

وباستخدام العلاقة

$$T^2 = \frac{2(n-2)}{n-3} F_\alpha$$

$$T^2 = 8.05$$

أي ان $\hat{E} > T^2$

وهذا يعني ان زوج القيم اعلاه $(x_0 = 31.8, Y_0 = 67.3)$ **يعتبر شاذاً باستخدام طريقة البياتي.**

والطريقة الثانية باستخدام الاحصاءة $t = Max|e_i|s_i|$ **على النحو التالي:**

ان تقديرات معالم معادلة خط الانحدار بطريقة المربعات الصغرى هي كما يلي:

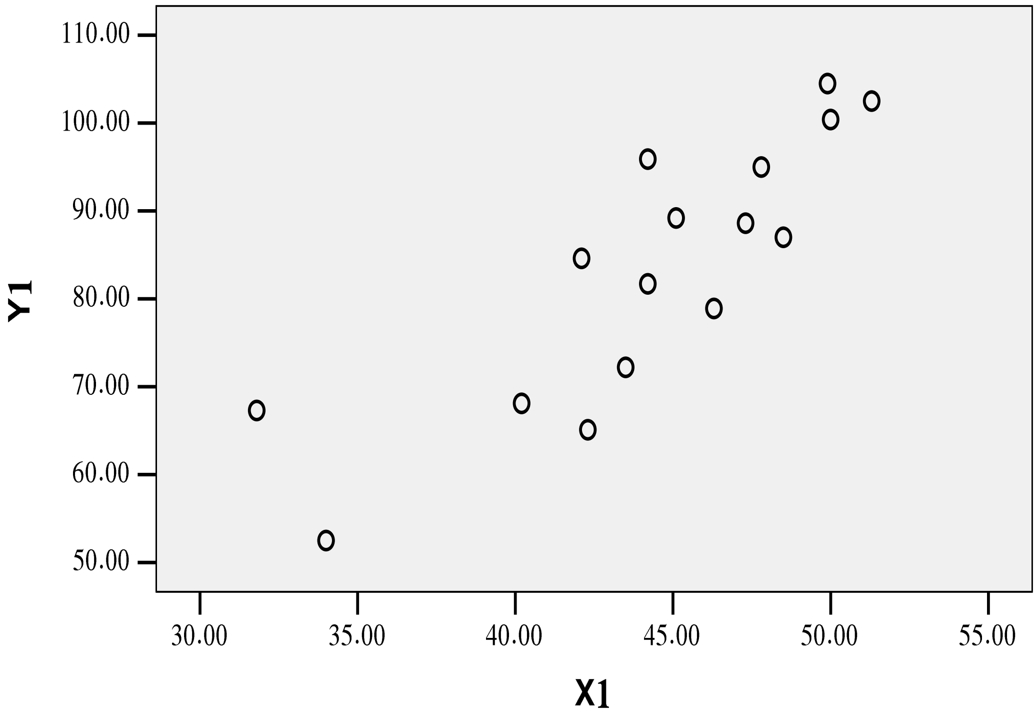
$$\hat{Y} = -21.8 + 2.4\ X_i$$

ولحساب المعلومات المطلوبة لحساب الاحصاءة t اعلاه.

وجد في جدول المعلومات في العمود $|e_i|s_i|$ **بأن اعلى قيمة لـ** e_i / s_i **هي (2.00)وهي القيمة التي تقابل كل من x =42.3 , y = 65.1 وبالمقارنة مع القيمة الحرجة** $(t^* = 2.9)^{16}$ **وبناءاً على ذلك لا توجد قيم شاذة في هذا المثال حسب هذا الاختبار ولمزيد المعلومات يمكن الرجوع الى المصدر (ديلر 1996).**

b- تحديد القيم الشاذة باستخدام طريقة الرسم الصندوقي المزدوج للمثال رقم (6) اعلاه وهي موضحة بالشكل رقم (9) التالي:

الشـــــــــــــــــــــــــــــــكل رقـــــــــــــــــــــــــــــــــم (9)



ونلاحظ من الشكل اعلاه بوضوح وجوه قيمة شاذة واحدة من ضمن القيم التي يمكن رؤيتها اذا رسمت بشكل منفصل خارج الرسم الصندوقي المزدوج وهي (x = 31.8 , y = 67.3) وهذه النتيجة مطابقة الى النتيجة التي تم التوصل اليها باستخدام احدى الطريقتين المعلمتين ومخالفة لنتيجة الطريقة المعلمية الاخرى، ومرة اخرى ظهرت فروقات في النتائج بين الطرق المعلمية والطرق الاستكشافية.

4. الاستنتاجات والتوصيات

4.1 – المقدمة:

سنتطرق في هذا المبحث اولاً الى اهم النتائج التي تم التوصل اليها في الجانب العملي:-

في المثال الاول والذي يمثل البيانات الخاصة بمتغير واحد كانت الطرق الاستكشافية متطابقة فيما بينها في تحديد القيم الشاذة حيث اظهرت كل من طريقتين الرسم الصندوقي والغصن والورقة بوجود قيمة شاذة واحدة، في حين ان الطرق المعلمية اختلفت فيما بينهما في تحديد القيم الشاذة، فحسب طريقة Grubbs هناك قيمة شاذة واحدة وقيمتان شاذتان حسب طريقة Tietjen and Moore وعدم وجود قيم شاذة حسب طريقة Rosner.

في المثال الثاني في الاطروحة والذي لم تتطرق اليه في الجانب العملي تطابقت الطرق الاستكشافية ايضاً في تحديد القيم الشاذة، حيث اظهرت كل من طريقتين الرسم الصندوقي والغصن والورقة عدم وجود قيم شاذة اما الطرق المعلمة فكانت الاختلافات بينها واضحة، وكانت النتائج وجود قيمتين شاذتين حسب طريقة Grubbs وعدم وجود اية قيمة شاذة باستخدام طريقة Tietjen and Moore، وهذا يعني ان الاختلاف كبير بين الطرق الاستكشافية والطرق المعلمية.

في المثال الثالث في الاطروحة الذي تم التطرق اليه في الجانب العملي مثال رقم 2، اختلفت الطرق الاستكشافية في تحديد القيم الشاذة حيث اظهرت طريقة الرسم الصندوقي بعدم وجود قيم شاذة، في حين عند استخدام طريقة الغصن والورقة كانت هناك ثلاث قيم شاذة، وكذلك الطرق المعلمية اختلفت ايضاً في تحديد القيم الشاذة وكانت نتيجة ثلاث طرق منها وهي طريقة Ronser وطريقة Grubbs وطريقة Tietjen and Moore متشابهة وهي عدم وجود قيم شاذة ووجود قيمتين شاذتين باستخدام طريقة McMillan. اما في المثال الرابع في الاطروحة والذي لم تتطرق اليه في الجانب العملي وموضح بشكل مفصل في المصدر (ديلر 1996) فقد استخدم طريقة معلمية واحدة لتحديد القيم الشاذة نظراً لطبيعة البيانات والتي كانت تتكون من ثلاث متغيرات ومجموعات وكانت النتائج وجود قيمة شاذة واحدة في كل متغير او مجموعة بينما في الطرق الاستكشافية كانت نتائجها مختلفة بالنسبة للمجموعة الاولى باستخدام الرسم الصندوقي لم تكن هناك قيمة شاذة ووجود قيمة شاذة واحدة باستخدام الغصن والورقة وللمجموعة الثانية وجود قيمة شاذة واحد باستخدام طريقة الغصن والورقة واخيراً للمجموعة الثالثة وجود قيمة شاذة واحدة باستخدام الرسم الصندوقي وقيمتين شاذتين باستخدام طريقة الغصن والورقة.

وهذا ايضاً مؤثر جديد لهذه الحالة من البيانات، حيث كانت هناك اختلافات كبيرة في النتائج باستخدام الطرق المعلمة والطرق الاستكشافية.

اما في المثال الخامس في الاطروحة والذي تم التطرق الى نتائجه في المثال رقم (3) في الجانب العملي في هذا المبحث فكانت نتائج الطرق الاستكشافية متشابهة نوعاً ما فيما بينها في تحديد القيم الشاذة حيث اظهرت ثلاث قيم شاذة باستخدام طريقة الرسم الصندوقي واربع قيم شاذة باستخدام طريقة الغصن والورقة، اما الطرق المعلمية فكانت نتائجها في هذا المثال متطابقة حيث لم تظهر اياً من الطرق الثلاث المستخدمة طريقة Rosner وطريقة Tietjen and Moore وطريقة Grubbs وجود أي قيمة شاذة.

في المثال السادس في الاطروحة والذي لم نتطرق اليه في هذا البحث في الجانب العملي والذي يمثل بيانات متغير واحد فقد كانت النتائج متشابهة نوعاً ما بين الطرق المعلمية والطرق الاستكشافية حيث اظهرت جميع الطرق نفس النتائج وهي عدم وجود قيمة شاذة فقط الاختلاف كان في طريقة الغصن والورقة حيث هناك قيمة شاذة واحدة فقط.

في المثال السابع في الاطروحة والذي يمثل الرابع في هذا المبحث في الجانب العملي والذي يمثل البيانات الخاصة بالمتغيرين (x ,y) ثم استخدام طريقتين معلمتين طريقة اختبار T^2 وطريقة الاحصاءة t لتحديد القيم الشاذة وكانت النتيجة متطابقة في الطريقتين وهي وجود زوج من

المشاهدات شاذاً اما باستخدام الرسم الصندوقي المزدوج للمتغيرين (x ,y) فلم يكن هناك أي زوج من القيم الشاذة.

اما المثال الثامن في الاطروحة والذي لم نتطرق اليه في البحث في الجانب العملي وهو ايضاً يمثل حالة العلاقة بين المتغيرين (x ,y) كانت النتيجة وجود زوج من قيم المشاهدات شاذاً باستخدام احد الطريقتين المعلميتن المستخدمتين وعدم وجود قيمة شاذة بأستخدام الطريقة الاخرى وعدم وجود أي زوج من قيم المشاهدات شاذاً باستخدام الرسم الصندوقي المزدوج للمتغيرين (x,y). في المثال التاسع في الاطروحة والذي لم نتطرق اليه في هذا البحث في الجانب العملي والذي يمثل حالة العلاقة بين المتغيرات (y , x_1 , x_2) كانت النتيجة متطابقة بين الطريقة المعلمية المستخدمة وطريقة الرسم الصندوقي المزدوج في تحديد القيم الشاذة وهي وجود قيمة شاذة بالنسبة لكل من المتغيرات (y , x_1 , x_2).

اما في المثال العاشر والذي تم التطرق اليه في هذا المبحث والذي يمثل العلاقة بين (x,y) فقد اختلفت الطريقتان المعلميتان المستخدمتان في تحديد القيم الشاذة فقد اظهرت قيمة من المشاهدات شاذة باستخدام احدى الطريقتين ولم يظهر أي من القيم شاذاً باستخدام الطريقة المعلمية الاخرى، اما باستخدام الرسم الصندوقي المزدوج فكانت هناك قيمة من قيم المشاهدات شاذة وهذا يعني وجود اختلافات بين الطرق المعلمية والاستكشافية وكذلك بين الطرق المعلمية نفسها.

4.2 : الاستنتاجـــات

1. اظهرت الدراسة تفاوت الطرق المعلمية في تحديد القيم الشاذة في حالة المتغير الواحد وهي Grubbs وطريقة Tietjen وطريقة Rosner وطريقة McMillan فيما بينهما بشكل واضح جداً. حيث كانت طريقة Grubbs اكثر حساسية لتحديد القيم الشاذة من غيرها من الطرق المعلمية وبعكس طريقة Rosner التي كانت اقل الطرق المعلمية حساسية لتحديد القيم الشاذة.
2. بينت الدراسة وجود تطابق ملحوظ في النتائج بين الطريقتين المعلمتين المستخدمتين طريقة اختبار T2 وطريقة اختبار الاحصاءة t وبين طريقة الرسم الصندوقي المزدوج Rangefinder Box plot في تحديد القيم الشاذة في حالة المتغيرين في اغلب الامثلة المستخدمة في البحث.
3. اظهرت الدراسة اختلاف بين الرسم الصندوقي والغصن والورقة في تحديد القيم الشاذة وان الغصن والورقة اكثر حساسية من الرسم الصندوقي.
4. بينت الدراسة هناك اختلافات واضحة بين الطرق المعلمية والطرق الاستكشافية في تحديد القيم الشاذة.
5. اظهرت الدراسة وجود خلاف كبير في النتائج بين الطرق المعلمية وطريقة الغصن والورقة (stem and Leaf) في اغلب الامثلة المستخدمة في البحث.
6. وجد من خلال النتائج للدراسة هناك تشابه في النتائج بين الطرق المعلمية (Tietjen and Moore) وطريقة الرسم الصندوقي (Box plot).
7. الطرق الاستكشافية تحدد القيم الشاذة مباشرة بينما الطرق المعلمية تتطلب تحديد قيم مشكوك فيها شاذة كأجراء اولي، ومن ثم الاعتماد على الاختبارات الاحصائية الخاصة بتحديد فيما اذا كانت تلك القيم شاذة ام لا.
8. توصل الباحث الى ان الرسم الصندوقي (Box plot)هو أفضل الطرق الاستكشافية وان طريقة (Tietjen and Moore)أفضل الطرق المعلمية وهناك تشابه بين نتائج الطريقتين كما مبين في الفقرة 6.
9. توصل الباحث بأن طريقة الرسم الصندوقي (Box plot) هي أفضل طريقة لتحديد القيم الشاذة بالنسبة للطرق الاستكشافية والمعلمية.

10. اثبتت الدراسة كفاءة ونجاح طريقة الرسم الصندوقي المزدوج (Rangefinder box plot) في تحديد القيم الشاذة في حالة المتغيرين، حيث انها تقوم بتحديد القيم الشاذة في كل من المتغيرين (x, y) على انفراد والقيم الشاذة في المتغيرين (x, y) معاً وتجمع المعلومات الخاصة بالمتغيرين (x, y) في رسم يمثل رسمين صندوقيين معاً.

4.3 : التوصيـــات

على ضوء الاستنتاجات السابقة يوصي الباحث بما يلي:

1. ضرورة اجراء عرض للبيانات قبل القيام بأية عملية تحليل احصائي حيث ان مثل هذا الاجراء يؤدي الى الحصول على عوامل ومتغيرات جديدة ستسهم في تطوير الظاهرة تحت الدراسة، كذلك الحصول على مؤثر جديد لمدى دقة العمل، وكلا الامرين سيسهمان في دقة النتائج.
2. استخدام الطرق الاستكشافية Box plot, stem and leaf, and Rangefinder Box plot لتحديد القيم الشاذة في البيانات في حالة المتغير الواحد والمتغيرين وتفضيلهما على الطرق المعلمية الاخرى.
3. اذا كان لابد من استخدام الطرق المعلمية لتحديد القيم الشاذة يوصي الباحث باستخدام طريقة (Tetjen and Moore) في حالة المتغير الواحد وطريقة اختبار الاحصاءة $\left(t = Max|e_i|s_i|\right)$ في حالة المتغيرين (x, y) وتفضيلهما على الطرق المعلمية الاخرى.
4. اجراء دراسات على بيانات مختلفة ومتنوعة يستخدمها الباحث باستخدام الطرق المعلمية والاستكشافية لتحديد بالضبط أي من الطرق سوف يكون اكثر كفاءة.

المصـــــادر

المصادر العربيـــة

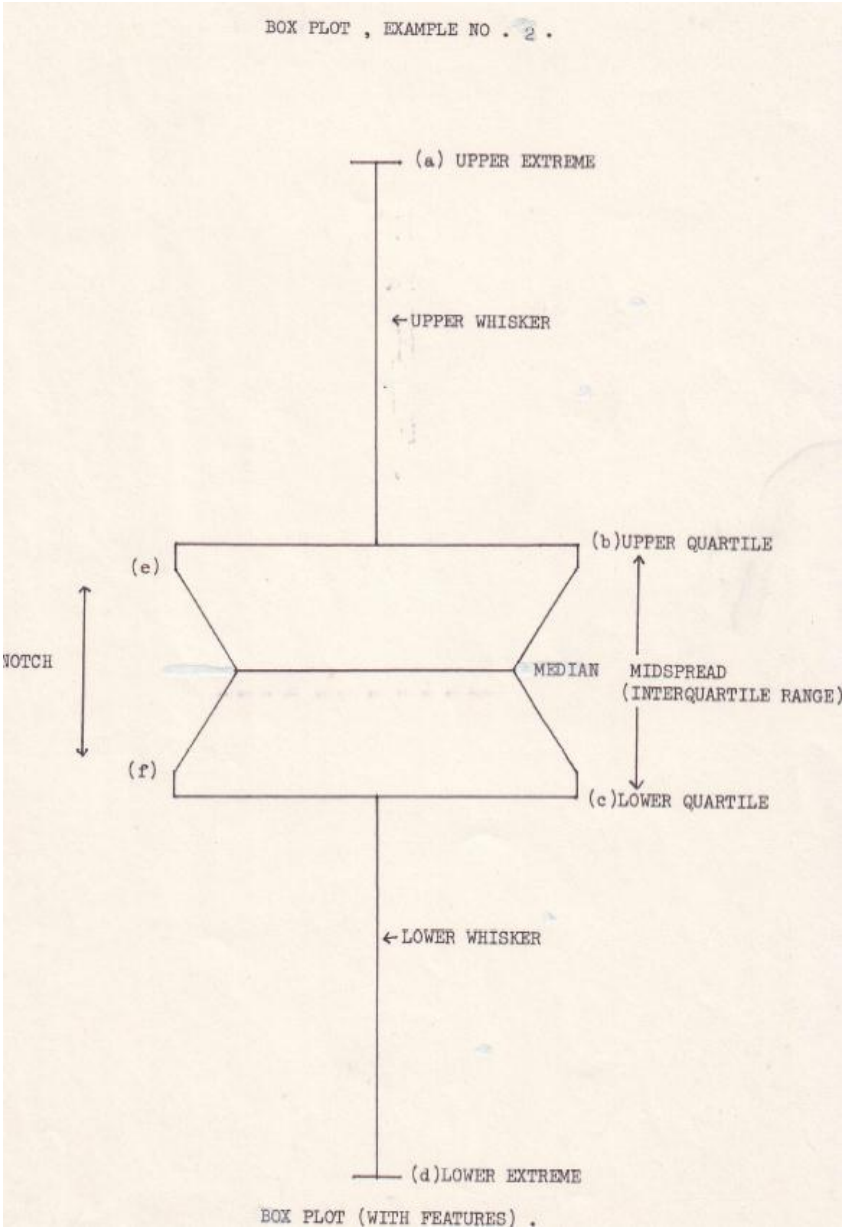
1. الجبوري، شلال حبيب (1988) "اسلوب جديد لاكتشاف وتقدير المشاهدات الشاذة في حالة متعدد المتغيرات"، بحث القي في المؤتمر الثاني للجمعية العراقية للعلوم.
2. الجبوري، منى حسين(1990) "الاكتشاف الجزئي للمشاهدات الشاذة وطرق التقدير في حالة متعدد المتغيرات"، رسالة ماجستير في الاحصاء، الجامعة المستنصرية.
3. المثنو، نغم مسلم(1993) " تقويم البيانات المفقودة والشاذة في تحليل تصميم التجارب غير المتزنة، رسالة ماجستير في الاحصاء، جامعة بغداد.
4. المختار، سليمان محمد امين (1980) " القيم الشاذة واثرها في تحليل البيانات الاحصائية، رسالة ماجستير في الاحصاء، جامعة بغداد.

المصادر الاجنبيـــة

5. **Barnett, V. & Lewis, T.(1978) "Outlier in Statistical Data", John Wiley and Sons, New York.**
6. **Al-Bayati, H.A.(1973)"Procedure for Detecting Observation in Samples two related Variables", The Proceeding of the Ninth Conference of Statistics and Computation to Multivariate Statistical Analysis, John Wiley and Sons,New York.**
7. **Becketti, S. & Gould, W.(1987) "Rangefinder Box plot Anote", The American Statistician, V.41,No.2,p.(149).**
8. **Behken, D. W. & Draper, N. R.(1972) "Residuals and their Variance Patterns", Technometrics, V.17,pp(127-128).**
9. **Bross, L. D. J.(1961) "Outliers in Patternend Expremints a Strategic reappraisal", Technometrics, V.3,pp(91-102).**
10. **Cleveland, W.S.(1985) "The Elements of Graphing Data", Monterey, CA; Wads Worth.**
11. **Daniel, C. (1960) "Location Outliers in Factorial Exeperiments", Technometrics, V.2,pp(140-150).**
12. **Gnanadesikan, R.& kettenring, J.R.(1972) "Robust Estimates, Residuals and Outlier Detection with Multiresponse Data", Biometrics,V.28,pp.(81-124).**
13. **Grubbs,F.E.(1950) "Sample Cirteria for Testing Outling Observation", Ann. Math. Stat., V.s1,pp(27-58).**
14. **Hussin, M. M. (1989) "Some studies of Graphical Methods in Statistical Data Analysis ; subjective Judgements in the Interpretation of Box plot", Unpublished Ph. D. Thesis, Keele University, Uk.**
15. **Irwin, J.O. (1925) "On a Criterion for the Rejection of Outlying Observation", Biometrika, V.17,pp.(238-250).**
16. **Al-Jobouri, S.(1976) "Test of Outliers", Unpublished M.Sc Thesis, University of Baghdad.**

17. **John & Prescott, P. (1975) "Critical Values of a Test to detect Outliers in Factorial Experiments", Appl. Sta., V.24,pp.(56-59).**
18. **Kishpaugh, J.R.L.(1972) "Experiments in Outliers Detection in Multivariate Data", M. Sc. Thesis, State University of New York, Stony Brook, NY.**
19. **Lund,R.E.(1975) "Table for an Approxiamte Test for Outliers in Linear Models", Technometrics, V.17.pp.(473-476).**
20. **McMillan, R.G. (1971) "Tests for one or Two Outliers in Normal Samples with unknown variance", Technometrics, V.13,No.1,pp.(87-100).**
21. **Mosteller, F.(1948) "AK-Sample Slippage Test for an Extreme Population", Ann. Math.Stat.,Vol.19,pp.(58-65).**
22. **----------& Tukey,J.W.(1950) "Significance Levels for a k-Sample Slippage Test", Ann.Math.Stat.,Vol.21,pp.(120-123).**
23. **Quesenberry, C. P. & David(1961) "Some Tests for Outliers", Biometrika,V.48,pp.(379-387).**
24. **Rohlf, F.J.(1975) "Generalization) of Gap Test for the Detection of Multivariate Outliers", Biometrics,V.31,pp.(93-101).**
25. **Rosner,B.(1975)"On the Detection of many Outliers",Technometrics,Vol.17,No.2,pp.(120-135).**
26. **Strikantan, K.S.(1961) "Testing for Single Outliers in a Regression Model", Sankhya,A.Vol.23,pp.(251-260).**
27. **Tietjen, G.L., Moore, R.H.& Beckman, R.J.(1973) "Testing for a Single Outlier in a Simple Linear Regression", Technometrics,Vol.15,pp.(717-721).**
28. **Tukey, J.W.(1977) "Exploratory Data Analysis", Addison-Wesley Publishing Company.**
29. **Velleman,P.F.& Hoaglin, D.C.(1981) "Applications, Basics and Computing Exploratory Data Analysis", Boston, M. Sc., Duxbury Press.**
30. **Zinger, A.(1961) "Detection of Best and Outline Normal Distributions with known Variances", Biometrika,vol.48,p.(457).**

شكل رقم(1AA)


BOX PLOT , EXAMPLE NO . 2 .
(a) UPPER EXTREME
←UPPER WHISKER
(b)UPPER QUARTILE
(e)
NOTCH
MEDIAN
MIDSPREAD
(INTERQUARTILE RANGE)
(f)
(c)LOWER QUARTILE
←LOWER WHISKER
(d)LOWER EXTREME
BOX PLOT (WITH FEATURES) .