Algoritmo de Grubbs

Método de Grubbs para la Detección de Outliers:

El método de Grubbs es una técnica estadística utilizada para detectar valores atípicos (outliers) en un conjunto de datos. Este método se basa en la idea de que un outlier es una observación que es significativamente diferente del resto del conjunto de datos. El método de Grubbs calcula un estadístico llamado "G" para cada valor en el conjunto de datos y compara este valor con un valor crítico basado en el nivel de confianza y el tamaño de muestra.

Pasos para Aplicar el Método de Grubbs:

1. **Calcular el promedio y la desviación estándar del conjunto de datos.**
2. **Calcular el estadístico "G" para cada valor en el conjunto de datos:**

Para una observación "x" en el conjunto de datos:

$$Gi = \frac{x - promedio}{desviacion\ estandar}$$

3. **Comparar el valor calculado de "G" con el valor crítico de Grubbs:**

El valor crítico se obtiene de una tabla de valores críticos de Grubbs, que depende del nivel de confianza y el tamaño de muestra. Si el valor calculado de "Gi" es mayor que el valor crítico correspondiente, la observación "x" podría ser considerada un outlier.

El método de Grubbs puede utilizarse para detectar más de un outlier en un conjunto de datos. Sin embargo, debes tener en cuenta que el método de Grubbs se aplica de manera iterativa. Esto significa que, una vez que se detecta un outlier, se elimina del conjunto de datos y luego se recalcula el estadístico de Grubbs con el nuevo conjunto de datos reducido. El proceso se repite hasta que ya no se encuentren más outliers o hasta que no se cumplan los criterios de detección.

El método de Grubbs puede utilizarse para detectar más de un outlier en un conjunto de datos. Sin embargo, debes tener en cuenta que el método de Grubbs se aplica de manera iterativa. Esto significa que, una vez que se detecta un outlier, se elimina del conjunto de datos y luego se recalcula el estadístico de Grubbs con el nuevo conjunto de datos reducido. El proceso se repite hasta que ya no se encuentren más outliers o hasta que no se cumplan los criterios de detección.

Pasos para Detectar Múltiples Outliers con el Método de Grubbs:

1. Calcula el promedio y la desviación estándar del conjunto de datos original.
2. Calcula el estadístico de Grubbs (G) para cada valor en el conjunto de datos original.
3. Identifica el valor con el mayor valor de G. Si este valor es mayor que el valor crítico de Grubbs correspondiente (según el nivel de confianza y el tamaño de muestra), se considera como un outlier.

4. Elimina el valor identificado como outlier del conjunto de datos y recalcula el promedio y la desviación estándar con el conjunto de datos reducido.
5. Calcula nuevamente los valores de G para el nuevo conjunto de datos.
6. Repite los pasos 3 a 5 hasta que ya no se encuentren más outliers o hasta que no se cumplan los criterios de detección.

Es importante destacar que este proceso debe realizarse de manera cuidadosa y documentada, ya que la eliminación de observaciones puede afectar la validez de tus análisis y conclusiones. Además, este método asume que los datos siguen una distribución normal, por lo que si tus datos no siguen esta distribución, podría ser necesario considerar métodos alternativos.

En resumen, el método de Grubbs puede ser utilizado para detectar más de un outlier, pero el proceso debe repetirse iterativamente para cada outlier detectado, y se deben tomar precauciones para evitar la manipulación inadecuada de los datos.

Ejercicio: Detección de Outliers en Mediciones de Distancias

Supongamos que un grupo de estudiantes de agrimensura ha realizado mediciones de distancias entre dos puntos utilizando diferentes instrumentos de medición. Los datos obtenidos son los siguientes:

35.2,35.1,35.0,35.5,35.3,35.7,35.4,34.9,35.6,55.035.2,35.1,35.0,35.5,35.3,35.7,35.4,34.9,35.6,55.0

Instrucciones:

1. **Calcula el promedio y la desviación estándar de las mediciones.**
2. **Usa el método de Grubbs para identificar posibles outliers:**

a. Calcula el estadístico "G" para cada valor en el conjunto de datos.

b. Consulta una tabla de valores críticos de Grubbs para un nivel de confianza del 95% y un tamaño de muestra de 10 para obtener el valor crítico correspondiente.

c. Compara los valores calculados de "G" con el valor crítico para cada observación.

d. Determina si alguna observación podría considerarse un outlier según el método de Grubbs.

3. **Presenta tus resultados y discute si hay outliers en las mediciones de distancias. Explica cómo llegaste a tu conclusión utilizando el método de Grubbs.**
4. **Pregunta de discusión:** ¿Qué factores podrían haber contribuido a la aparición de cualquier outlier identificado? ¿Cómo podrían los agrimensores minimizar o evitar la presencia de outliers en sus mediciones?

RESOLUCION Ejercicio: Detección de Outliers en Mediciones de Distancias

Datos: 35.2, 35.1, 35.0, 35.5, 35.3, 35.7, 35.4, 34.9, 35.6, 55.0

1. **Calcula el promedio y la desviación estándar:**

Promedio: $\frac{35.2+35.1+35.0+35.5+35.3+35.7+35.4+34.9+35.6+55.0}{10}$=37.28

Desviación Estándar: ≈6.76

Aplicar el método de Grubbs:

a. Calculamos el estadístico "G" para cada valor en el conjunto de datos:

$$G = \frac{x - promedio}{desviacion\ estandar}$$

Para cada valor en el conjunto de datos:

$$G = \frac{x - 37.28}{6.76}$$

Calculamos "G" para cada valor:

G(35.2)≈0.103, G(35.1)≈0.204, G(35.0)≈0.306, G(35.5)≈0.102, G(35.3)≈0.005, G(35.7)≈0.349, G(35.4)≈0.163, G(34.9)≈0.543, G(35.6)≈0.056, G(55.0)≈3.182

b. Consultamos una tabla de valores críticos de Grubbs para un nivel de confianza del 95% y un tamaño de muestra de 10. El valor crítico es aproximadamente 2.821.

c. Comparamos los valores calculados de "G" con el valor crítico:

Ninguna observación tiene un valor calculado de "G" mayor que el valor crítico, excepto la última observación (55.0).

2. **Presenta tus resultados y discute:**

Según el método de Grubbs, la observación 55.0 podría ser considerada un outlier, ya que su valor calculado de "G" (3.182) es mayor que el valor crítico (2.821).

Discusión adicional:

Se podría argumentar que la observación 55.0 es significativamente diferente de las otras mediciones de distancias. Podría haber varios factores que contribuyan a este outlier, como errores en la medición, calibración incorrecta del equipo o incluso factores externos que afectaron la medición.

EXTREME STUDENTIZED DEVIATE TEST

Name:

EXTREME STUDENTIZED DEVIATE TEST

Type:

Analysis Command

Purpose:

Perform a generalized extreme studentized deviate (ESD) test for outliers.

Description:

The generalized extreme Studentized deviate (ESD) test is used to detect one or more outliers in a univariate data set that follows an approximately normal distribution.

The primary limitation of the Grubbs test and the Tietjen-Moore test is that the suspected number of outliers, *k*, must be specified exactly. If *k* is not specified correctly, this can distort the conclusions of these tests. On the other hand, the generalized ESD test only requires that an upper bound for the suspected number of outliers be specified.

Given the upper bound, *r*, the generalized ESD test essentially performs *r* separate tests: a test for one outlier, a test for two outliers, and so on up to *r* outliers.

The generalized ESD test is defined for the hypothesis:

H_0:	There are no outliers in the data set
H_a:	There are up to *r* outliers in the data set
Test Statistic:	Compute $R1=\max_i \lvert x_i-\bar{x}\rvert/s$ with $\bar{x}$ and *s* denoting the sample mean and sample standard deviation, respectively. Remove the observation that maximizes $\lvert x_i-\bar{x}\rvert$ and then recompute the above statistic with *n* - 1 observations. Repeat this process until *r* observations have been removed. This results in the *r* test statistics $R_1, R_2, \ldots, R_r$.
Significance Level:	α
Critical Region:	Corresponding to the *r* test statistics, compute the following *r* critical values $\lambda_i=t_{n-i-1,p}(n-i)(n-i-1+t^2_{n-i-1,p})(n-i+1)\nu$ where *i* = 1, 2, ..., *r*, $t_{\nu,p}$ is the 100*p* percentage point from the t distribution with ν degrees of freedom and $p=1-\alpha 2(n-i+1)$.

The number of outliers is determined by finding the largest *i* such that $R_i > \lambda i$.

Simulation studies by Rosner indicate that this critical value approximation is very accurate for $n \geq 25$ and reasonably accurate for $n \geq 15$.

Note that although the generalized ESD is essentially Grubbs test applied sequentially, there are a few important distinctions:

- The generalized ESD test makes approriate adjustments for the critical values based on the number of outliers being tested for that the sequential application of Grubbs test does not.
- If there is significant masking, applying Grubbs test sequentially may stop too soon. The example below identifies 3 outliers at the 5% level when using the generalized ESD test. However, trying to use Grubbs test sequentially would stop at the first iteration and declare no outliers.
- Grubbs test allows one-sided tests (i.e., you can specify a minimum test or the maximum test) in addition to two-sided tests (both the minimum and the maximum value are tested). The generalized ESD test is restricted to two-sided tests.

Generalized ESD Test for Outliers

Purpose: Detection of Outliers

The generalized (extreme Studentized deviate) ESD test (Rosner 1983) is used to detect one or more outliers in a univariate data set that follows an approximately normal distribution.

The primary limitation of the Grubbs test and the Tietjen-Moore test is that the suspected number of outliers, *k*, must be specified exactly. If *k* is not specified correctly, this can distort the conclusions of these tests. On the other hand, the generalized ESD test (Rosner 1983) only requires that an upper bound for the suspected number of outliers be specified.

Definition

Given the upper bound, *r*, the generalized ESD test essentially performs *r* separate tests: a test for one outlier, a test for two outliers, and so on up to *r* outliers.

The generalized ESD test is defined for the hypothesis:

H_0: There are no outliers in the data set

H_a: There are up to *r* outliers in the data set

Test Statistic:	Compute $$Ri=maxi\|xi-x^{-}\|s$$ with x^{-} and ***s*** denoting the sample mean and sample standard deviation, respectively. Remove the observation that maximizes $\|xi-x^{-}\|$ and then recompute the above statistic with *n* - 1 observations. Repeat this process until *r* observations have been removed. This results in the *r* test statistics R_1, R_2, ..., R_r.
Significance Level:	α
Critical Region:	Corresponding to the *r* test statistics, compute the following *r* critical values $$\lambda i=(n-i)tp,n-i-1(n-i-1+t2p,n-i-1)(n-i+1)\forall i=1,2,\ldots,r$$ where $t_{p,v}$ is the 100*p* percentage point from the *t* distribution with *v* degrees of freedom and $$p=1-\alpha 2(n-i+1)$$ The number of outliers is determined by finding the largest *i* such that $R_i > \lambda_i$. Simulation studies by Rosner indicate that this critical value approximation is very accurate for $n \geq 25$ and reasonably accurate for $n \geq 15$.

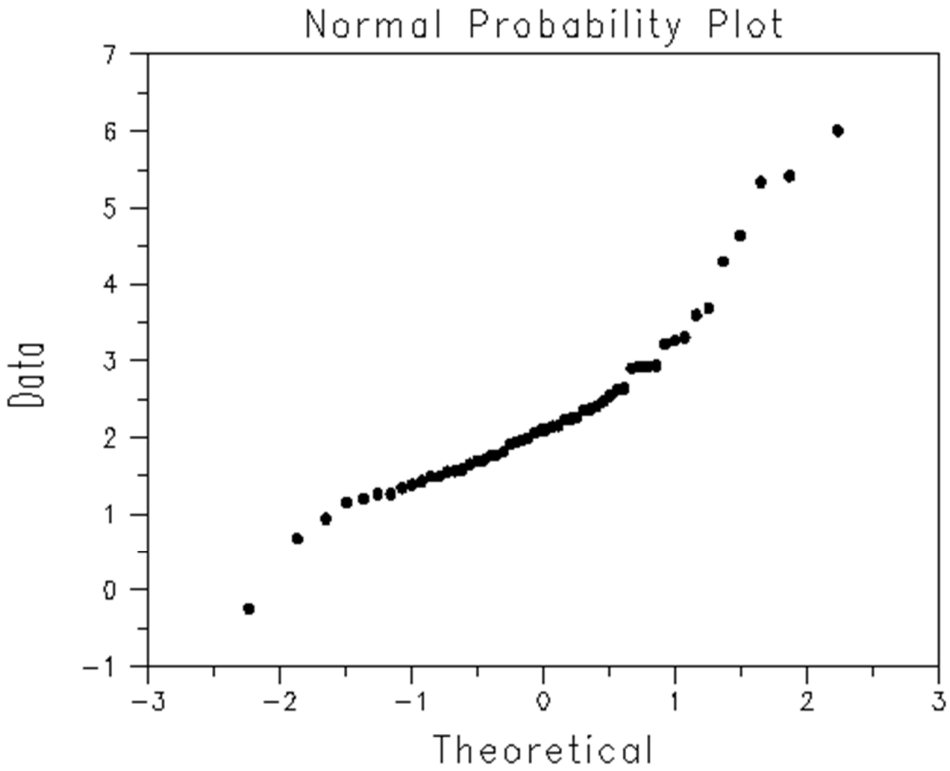
Note that although the generalized ESD is essentially Grubbs test applied sequentially, there are a few important distinctions:

- The generalized ESD test makes approriate adjustments for the critical values based on the number of outliers being tested for that the sequential application of Grubbs test does not.
- If there is significant masking, applying Grubbs test sequentially may stop too soon. The example below identifies three outliers at the 5 % level when using the generalized ESD test. However, trying to use Grubbs test sequentially would stop at the first iteration and declare no outliers.

Generalized ESD Test Example

The Rosner paper gives an example with the following data.

-0.25 0.68 0.94 1.15 1.20 1.26 1.26



1.34 1.38 1.43 1.49 1.49 1.55 1.56
1.58 1.65 1.69 1.70 1.76 1.77 1.81
1.91 1.94 1.96 1.99 2.06 2.09 2.10
2.14 2.15 2.23 2.24 2.26 2.35 2.37
2.40 2.47 2.54 2.62 2.64 2.90 2.92
2.92 2.93 3.21 3.26 3.30 3.59 3.68
4.30 4.64 5.34 5.42 6.01

As a first step, a normal probability plot was generated

This plot indicates that the normality assumption is questionable.

Following the Rosner paper, we test for up to 10 outliers:

H_0: there are no outliers in the data
H_a: there are up to 10 outliers in the data

Significance level: $\alpha = 0.05$
Critical region: Reject H_0 if R_i > critical value

Summary Table for Two-Tailed Test

Exact Number of Outliers, i	Test Statistic Value, R_i	Critical Value, λ_i 5 %
1	3.118	3.158
2	2.942	3.151
3	3.179	3.143 *
4	2.810	3.136
5	2.815	3.128
6	2.848	3.120
7	2.279	3.111
8	2.310	3.103
9	2.101	3.094
10	2.067	3.085

For the generalized ESD test above, there are essentially 10 separate tests being performed. For this example, the largest number of outliers for which the test statistic is greater than the critical value (at the 5 % level) is three. We therefore conclude that there are three outliers in this data set.

Questions

The generalized ESD test can be used to answer the following question:

1. How many outliers does the data set contain?

Importance

Many statistical techniques are sensitive to the presence of outliers. For example, simple calculations of the mean and standard deviation may be distorted by a single grossly inaccurate data point.

Checking for outliers should be a routine part of any data analysis. Potential outliers should be examined to see if they are possibly erroneous. If the data point is in error, it should be corrected if possible and deleted if it is not possible. If there is no reason to believe that the outlying point is in error, it should not be deleted without careful consideration. However, the use of more robust techniques may be warranted. Robust techniques will often downweight the effect of outlying points without deleting them.

Related Techniques

Several graphical techniques can, and should, be used to help detect outliers. A simple normal probability plot, run sequence plot, a box plot, or a histogram should show any obviously outlying points. In addition to showing potential outliers, several of these graphics also

help assess whether the data follow an approximately normal distribution.

- Run Sequence Plot
- Histogram
- Box Plot
- Normal Probability Plot
- Lag Plot