# TRATAMIENTO ESTADISTICO DE UNA SERIE DE OBSERVACIONES

Supongamos que realizamos *n* observaciones de una misma magnitud, en forma independiente y bajo las mismas condiciones.
Los estimadores usados son los ya vistos:

| PARAMETRO | ESTIMADOR USADO |
|---|---|
| $\mu$= valor medio | $\Rightarrow \bar{X} = \frac{x_1+x_2+.....+x_n}{n}$ |
| $\sigma$ = desviación de cada obs. | $\Rightarrow s = \sqrt{\frac{1}{n-1}\sum(\bar{X} - X_i)^2}$ |
| $\sigma_{\bar{X}}$= desviación del promedio | $\Rightarrow \xi = \frac{s}{\sqrt{n}}$ |
| $\sigma_{XY}$= covarianza | $\Rightarrow s = \frac{1}{n-1}\sum_{i=1}^{i=n}(X_i - \bar{X})\,(Y_i - \bar{Y})$ |

## LIMITE DE CONFIDENCIA O TOLERANCIA – DETECCION OUTLIERS

Un valor atípico o OUTLIER es una observación que parece desviarse notablemente de otras observaciones en la muestra.

Un valor atípico puede indicar datos incorrectos. Por ejemplo, es posible que los datos se hayan codificado incorrectamente o que una observación no se anote correctamente. Si se puede determinar que un punto atípico es de hecho erróneo, entonces el valor atípico debe eliminarse del análisis (o corregirse si es posible).

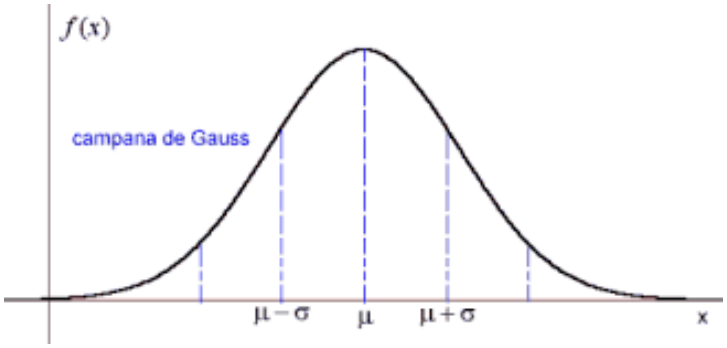
En algunos casos, puede que no sea posible determinar si un punto atípico son datos incorrectos. Los valores atípicos pueden deberse a una variación aleatoria o pueden indicar algo científicamente interesante.

Es importante tener un criterio para definir cuando estamos ante una muestra con un número de observaciones importantes o de lo contrario cuando las mismas son pocas.
No existe una definición teórica o cuya deducción sea matemática de este número, pero la experiencia y la diversa bibliografía al respecto nos lleva a convenir en considerar como **30 el número de observaciones** a partir del cual consideramos una muestra de gran tamaño o una pequeña.

## MUESTRAS GRANDES (n ≥ 30)

## CRITERIOS BASADO EN LA Distribución Normal

Recordemos la Distribución Normal: $X \approx N(\mu, \sigma)$




Ya vimos que si analizamos una medición de la magnitud X, ésta se comportará siguiendo este modelo.

Si calculamos la probabilidad de que la medición X caiga en el intervalo: $[\mu - \sigma, \mu + \sigma]$

Tenemos que:

$$P[\mu - \sigma \leq x \leq \mu + \sigma] = P\left[\frac{((\mu - \sigma) - \mu)}{\sigma} < z < \frac{((\mu + \sigma) - \mu)}{\sigma}\right] = P[-1 < z < 1]$$

$$P[\mu - \sigma \leq x \leq \mu + \sigma] = \Phi(1) - \Phi(-1) = 0.8413 - 0.1587 = 0.6826$$

***→ La probabilidad de que la observación X se encuentre en ese intervalo es del 68.26%.***

Si multiplicamos a σ por un número K, aumentamos el tamaño del intervalo, por lo que aumentamos la probabilidad de que la observación caiga en el intervalo:

$$[\mu - k\sigma, \mu + k\sigma]$$

$$P[\mu - k\sigma \leq x \leq \mu + k\sigma] = P[-k < z < k]\Phi(k) - \Phi(-k)$$

Como $\Phi(-k) = 1 - \Phi(k)$

$$P[\mu - k\sigma \leq x \leq \mu + k\sigma] = \Phi(k) + \Phi(k) - 1 = 2\Phi(k) - 1$$

| Niveles de Confianza | | |
|---|---|---|
| k | ϕ (k) | Prob. |
| 0,674 | 0,7498 | ***0,500*** |
| 1 | 0,8413 | ***0,683*** |
| 2 | 0,9772 | ***0,954*** |
| 3 | 0,9989 | ***0,997*** |

La cantidad ± 0.674 σ es conocida como error probable

Al realizar una serie de mediciones de una magnitud dada, es posible que en algunos casos aislados se cometa un error no casual (o sea, no gaussiano), originado por un factor extraño (error de cálculo, mal funcionamiento del aparato de medición, equivocación personal, etc.). La distribución normal nos permite utilizar un criterio físico para rechazar un dato sospechoso.

**La probabilidad de que un dato se encuentre fuera del intervalo [$\overline{X}$ ± 3σ ] ≈ 0.3%.**
**O sea que recién en 1000 datos podría esperarse que hayan 3 datos fuera del intervalo.**

Supongamos haber hecho 100 mediciones de una magnitud, con un valor medio $\overline{X}$ y una dispersión estándar σ de cada dato, y que entre los 100 datos haya 3 que difieren de $\overline{X}$ en más de 3σ por ej.
El hecho de que los 3 datos aparezcan entre un número 10 veces menor de datos, es un indicio de que esas 3 mediciones padecen de un defecto "extra gaussiano" y deben rechazarse. De esta forma podemos fijar para cada serie de mediciones un ***"Límite de Confidencia o Tolerancia".***

Cuando son pocas las observaciones (n≤30) no es posible eliminar una observación afectada de un error grosero, por mas evidente que sea, aplicando el intervalo de tolerancia visto 3σ.
La razón es que la tolerancia de Gauss está basada en un número muy grande de observaciones, en cuyo caso si es posible su eliminación.

En caso de tener pocas observaciones al calcular: $s = \sqrt{\frac{1}{n-1}\sum(\bar{X} - X_i)^2}$ la observación afectada de error no aleatorio produce un valor exagerado de *s* y por ende al hacer *3s* la observación va a caer dentro del intervalo de tolerancia.

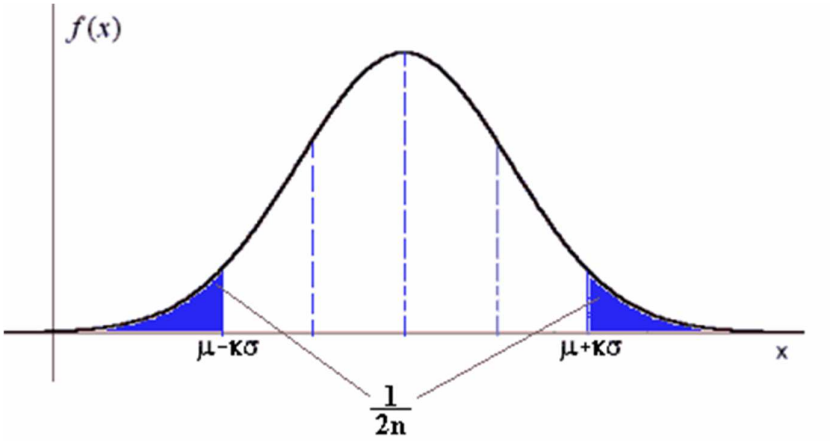
En estos casos se podría pensar que cuando es muy evidente un error grosero, debe excluirse directamente. Pero el problema consiste en "cuando es muy evidente" y en la dosis de subjetividad que en tal apreciación aporta el responsable de la medición.

## CRITERIO DE CHAUVENET

Otro criterio para realizar una depuración de observaciones basado en la distribución normal es el ***criterio de Chauvenet***, este consiste en eliminar directamente las observaciones que caigan fuera del intervalo determinado de la siguiente manera: $[\mu - k\sigma, \mu + k\sigma]$ siendo:

P [μ - kσ < x < μ + kσ] = 1 - 1/(2n)

$$P[\mu - k\sigma \leq x \leq \mu + k\sigma] = 1 - \frac{1}{2n}$$




| Cuadro de Valores de Chauvenet | | | | | |
|---|---|---|---|---|---|
| n | k | n | k | n | k |
| 4 | 1,54 | 12 | 2,03 | 26 | 2,35 |
| 5 | 1,68 | 14 | 2,10 | 30 | 2,39 |
| 6 | 1,73 | 16 | 2,16 | 40 | 2,50 |
| 7 | 1,79 | 18 | 2,20 | 50 | 2,58 |
| 8 | 1,86 | 20 | 2,24 | 100 | 2,80 |
| 9 | 1,92 | 22 | 2,28 | 200 | 3,02 |
| 10 | 1,96 | 24 | 2,31 | | |

Cuando son relativamente pocas las observaciones (n $\leq$ 30) no es posible detectar una observación afectada de un error grosero, por mas evidente que sea, aplicando el intervalo de tolerancia visto 3σ.
La razón es que la tolerancia de Gauss está basada en un número muy grande de observaciones, en cuyo caso si es posible su detección.

En caso de tener pocas observaciones al calcular: $s = \sqrt{\frac{1}{n-1}\sum(\bar{X} - X_i)^2}$ la observación afectada de error no aleatorio produce un valor exagerado de s y por ende al hacer 3s la observación va a caer dentro del intervalo de tolerancia.

En el caso de Chauvenet, a pesar de definir el intervalo a partir del nro. de observaciones, se basa en la distribución normal, por lo cual su uso para muestras de pocas observaciones no es recomendable.

## MUESTRAS PEQUEÑAS (n ≤ 30) - Distribucion t-Student.

Cuando realizamos relevamientos, generalmente tomamos muy pocas observaciones generando muestras muy pequeñas. Las medias muestrales y las desviaciones estándar tienen niveles de incertidumbre más altos que los derivados de la distribución normal, que proviene de toda una población de observaciones. A menos que observemos más de 30 muestras de un solo valor, realmente deberíamos usar la distribución *t- Student* y no la distribución normal.

### TEST DE GRUBBS (Normas ISO)

El Test de Grubbs detecta un valor atípico o **outlier** a la vez. Este outlier se elimina del conjunto de datos y la prueba se repite hasta que no se detectan valores atípicos. Sin embargo, las iteraciones múltiples cambian las probabilidades de detección, y la prueba no debe usarse iterativamente para tamaños de muestra de seis o menos. En estos casos el criterio que seguiremos es el de repetir la serie en caso de detectarse un valor atípico.

El Test de Grubbs se define para las siguientes hipótesis:

$H_0$ : No hay valores atípicos en el conjunto de datos
H: Hay exactamente un valor atípico en el conjunto de datos

La estadística de prueba de Grubbs se define como: $G = \frac{max|Y_i - \bar{Y}|}{s}$

donde $\bar{Y}$ es la media y s la desviación estándar muestrales.

La hipótesis de ausencia de valores atípicos se rechaza en el nivel de significancia α si:

$G \geq \frac{n-1}{\sqrt{n}}\sqrt{\frac{t^2}{n-2+t^2}}$ valor crítico

donde t es el valor de la distribución t con n − 2 grados de libertad y un nivel de significancia de α/(2n), es decir T Student de 2 colas.
De esta manera si:

G < G crítico entonces se mantiene el valor en el conjunto de datos. No es un **caso atípico.**
G > G crítico entonces se **rechaza el valor por ser atípico o OUTLIER.**

**Aplicación libre:** https://www.graphpad.com/quickcalcs/Grubbs1.cfm

## UTILIZACION DE LA t- Student DIRECTAMENTE

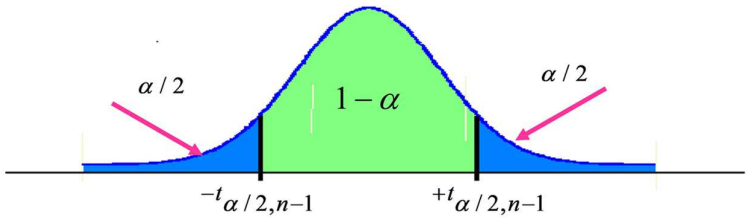
La observación de muestras pequeñas es muy común donde los ángulos a menudo se observan solo dos o cuatro veces, y las observaciones de distancia repetidas simplemente significan presionar el botón de medición mas de una vez. Las distribuciones de muestreo se basan en la media y la varianza de una muestra, así como en el numero de observaciones redundantes conocidas como sus ***grados de libertad***.
Los grados de libertad son todas las observaciones realizadas mas allá de lo necesario para determinar un valor. Entonces, por ejemplo, si hicimos cuatro observaciones de un angulo, cualquiera de estos valores podría usarse para indicar el tamaño del angulo. Las tres observaciones adicionales restantes se consideran observaciones ***redundantes*** y representan los grados de libertad de la observación.

A medida que aumenta el numero de observaciones redundantes, los valores críticos (valores que definen el intervalo de confianza) de la distribución de muestreo se aproximan a los valores para el mismo nivel de confianza de la distribución normal.

Por ejemplo, a 30 grados de libertad, el valor critico de la distribución *t- Student* es 2,04, que no esta lejos de su valor de distribución normal de 1,96. Tener en cuenta que la ***distribución t-Student*** se convierte en la **distribución normal** con un numero infinito de observaciones.

Fijando un nivel de confianza, considerando los grados de libertad (n-1) y las observaciones realizadas podemos construir un intervalo de confianza para una sola observación que nos ayudará a determinar si en alguna de las observaciones ocurrió un error no aleatorio.

$$\bar{X} - s.t_{\alpha/2}(n-1), \bar{X} + s.t_{\alpha/2}(n-1)$$




Observación: en caso de querer construir un intervalo de confianza para la media, debemos multiplicar el valor de t por $\frac{s_n}{\sqrt{n}}$ (desviación estándar de la media).

## RESUMEN EJECUTIVO DEL TRATAMIENTO DE UNA SERIE DE OBSERVACIONES

**1) Medir *n*** veces la magnitud, expresando los valores con sus cifras significativas (determinadas por el instrumento de medición) y según apuntes.

**2) Tratar y Depurar** la serie:

a) **Si n ≥ 30: usar la Distribución Normal (Criterio 3S)**

a.1) Calcular el promedio $\overline{X}$ y la desviación S de cada observación.

a.2) Fijar un intervalo de tolerancia para un nivel de confianza determinado: $\overline{X}$ ± 3S
Rechazar todos los datos que estuvieren fuera del intervalo y que de acuerdo al modelo gaussiano no deberían estarlo.

a.3) Repetir la prueba (a.1) y (a.2), hasta que ningún valor quede fuera del intervalo siempre y cuando hasta n=30.

b) **Si n ≤ 30, usar el Test de Grubbs** (basado en la Distribución t-Student) de la siguiente manera:

b.1) **Si n > 6**: Para un nivel de significancia del 95% (α = 5%), calcular el promedio $\overline{X}$, la desviación S de cada observación y el valor de **G crítico**. Rechazar todos los datos cuyo valor de $\boldsymbol{G_i}$ superara el crítico.

b.1.1) Repetir la prueba b.1 hasta que ningún valor supere el de G crítico.

b.2) **Si n ≤ 6**: Para un nivel de significancia del 95% (α = 5%), calcular el promedio $\overline{X}$, la desviación S de cada observación y el valor de G crítico. *Si por lo menos se detecta 1 valor de G que supera el crítico, entonces la serie debe observarse nuevamente.*

**3) Recalcular** $\overline{X}$ y S corregidos.

**4) Calcular** el error estándar del promedio: $\xi = \frac{s}{\sqrt{n}}$

**5) Expresar el resultado**: $\overline{X}$ ± ξ interpretándolo correctamente:

a) ***El valor mas probable de X es $\overline{X}$.***

b) La probabilidad de que el ***verdadero valor esté en el intervalo $\overline{X}$ ± ξ es del 68%*** (la probabilidad de que al analizar una nueva serie de medidas de X, su promedio $\overline{X}$ caiga en $\overline{X}$ ± ξ es del 68%).

*Obs.*: Si expresamos $\overline{X}$ ± S decimos:

- ***el valor mas probable de X es $\overline{X}$.***
- ***la probabilidad de que una nueva observación caiga en el intervalo $\overline{X}$ ± S es del 68%.***

## EJEMPLOS

Para determinar la longitud entre dos puntos A y B de una base de replanteo para un trabajo de precisión, se ha realizado una serie de observaciones obteniendo los siguientes resultados:

### Muestra 1.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 127,834 | 127,839 | 127,832 | 127,831 | 127,830 | 127,852 | 127,801 | 127,832 | 127,833 | 127,835 |

Como n = 10, aplicamos el criterio de la t-Student directamente considerando un nivel de confianza del 95%.

$\overline{X}$ = 127.8319 $\quad s = \sqrt{\frac{1}{n-1}\sum(\bar{X} - X_i)^2}$ = 0.0126

n = 10; α = 0,05 → G critico = 2,28995

Calculando $G = \frac{|X_i - \bar{X}|}{s}$ para cada valor observado, se considera outlier aquel que supere al crítico.

→ descartamos 127.801

### Muestra 2.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 127,852 | 127,839 | 127,835 | 127,834 | 127,833 | 127,832 | 127,832 | 127,831 | 127,830 |

$\overline{X}$ = 127.8353 S = 0.00678

n = 9; α = 0,05 → G critico = 2,2150

Calculando $G = \frac{|X_i - \bar{X}|}{s}$ para cada valor observado, se considera outlier aquel que supere al crítico.

→ descartamos 127.852

### *Muestra 3.*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 127,839 | 127,835 | 127,834 | 127,833 | 127,832 | 127,832 | 127,831 | 127,830 |

$\overline{X}$ = 127.8333 S = 0.00282

n = 8; α = 0,05 → G critico = 2,12664

Calculando $G = \frac{|X_i - \bar{X}|}{s}$ para cada valor observado, se considera outlier aquel que supere al crítico.

**→ No descartamos ninguno**

$\xi = \frac{s}{\sqrt{n}}$= 0.00282 / √ 8 = 0.000997

Resultado: 127.833 ± 0.001 m ($\overline{X}$ ± ξ )

127.833 ± 0.003 m ($\overline{X}$ ± S )

***Teoria del Caso 1***

**Caso Grupo 3, distancia R2-R4**

| R2 - R4 |
|---|
| 30,036 |
| 30,031 |
| 30,032 |
| 30,028 |
| 30,016 |
| 30,300 |

$\overline{X}$ = 30,074 $s = \sqrt{\frac{1}{n-1}\sum(\bar{X} - X_i)^2}$= 0.111

| R2 - R4 | G |
|---|---|
| 30,036 | 0,3408 |
| 30,031 | 0,3859 |
| 30,032 | 0,3769 |
| 30,028 | 0,4129 |
| 30,016 | 0,5210 |
| 30,300 | 2,0374 |

n = 6; α = 0,05 → G critico = 1,8871 → descartamos 30,300

**La serie debe observarse nuevamente**

**NOTAS:**

1) Ya vimos que σ es la desviación de una observación. La cantidad $\eta = \frac{\sigma}{\overline{X}}$ se llama *Error o Desviación relativa (o unitaria) de cada medición*. Y 100.η se llama error porcentual.

   La desviación η no tiene dimensiones. Cuando decimos que un error es del 10% tenemos una información sobre la calidad de la medición.
   Si decimos que el error estándar de una medición de una longitud es de 10 cm, ello puede representar una medición excelente (si la medición es de centenares de metros) o una mala (si el objeto medido tiene 20cm).
   Cuando nos dicen: precisión de 1/1000 significa que en 1000 unidades el error tolerable es de 1 unidad.

2) Si deseamos conocer la exactitud de un resultado de una observación topográfica recordemos que:

   ε = τ – x (con τ = Verdadero valor).

   Cuando trabajamos con una serie de observaciones como en el ejemplo anterior:

   >> Exactitud = $\tau - \overline{X}$

   Dado que τ es desconocido, se conviene en adoptar el razonable criterio de considerar como valor exacto de la magnitud medida, al obtenido con orden de precisión muy superior al de aquel que pretendemos analizar.

**EJEMPLO:**

Medimos un ángulo usando:

Teodolito T1 >> α = 52º35'20" y $\sigma_{\alpha'}$ = 10"
Teodolito T2 >> α = 52º35'35" y $\sigma_{\alpha'}$ = 1"

Con T1:
$\eta_{T1}$ = 10" / 52º35'20" = 1 / 18932

Con T2:
$\eta_{T2}$ = 1" / 52º35'35" = 1 / 189335

= => Es probable que la precisión sea de 10 veces más
= => Tomamos como τ = 52º35'35"

Para evaluar la exactitud = ± 15", se utiliza un criterio de tomar como tolerancia $3\sigma_{\alpha'}$, es decir:
tolerancia = | τ – X' | = | $3\sigma_{\alpha'}$ |

En nuestro ejemplo:
| $3\sigma_{\alpha'}$ | = 30", o sea que la exactitud lograda con el T1 es satisfactoria.

# DETECCIÓN DE OUTLIERS

Resumiendo lo visto para el tratamiento de una serie de observaciones debemos dar los siguientes pasos:

**1)** Mídase *n* veces la magnitud, expresando los valores con sus cifras significativas (determinadas por el instrumento de medición).

**2)** Depuración de la serie:

a) **Si n ≥ 30: usar la Distribución Normal.**

a.1) Calcular el promedio $\overline{X}$ y la desviación S de cada observación.
a.2) Fijar un intervalo de tolerancia para un nivel de confianza determinado: $\overline{X}$ ± 3S
Rechazar todos los datos que estuvieren fuera del intervalo y que de acuerdo al modelo gaussiano no deberían estarlo.
a.3) Repetir la prueba (a.1) y a.2)), hasta que ningún valor quede fuera del intervalo.

b) **Si n ≤ 30**, usar el criterio de Grubbs (basado en la Distribución t-Student) de la siguiente manera:

b.1) **Si n > 6**: Para un nivel de significancia del 95% (α = 5%), calcular el promedio $\overline{X}$, la desviación S de cada observación y el valor de G crítico. Rechazar todos los datos cuyo valor de G supere el crítico.
b.1.1) Repetir la prueba b.1 hasta que ningún valor supere el de G crítico.

b.2) **Si n ≤ 6**: Para un nivel de significancia del 95% (α = 5%), calcular el promedio $\overline{X}$, la desviación S de cada observación y el valor de G crítico. Si por lo menos se detecta un valor de G que supera el crítico, entonces la serie debe observarse nuevamente.

**3)** Recalcular $\overline{X}$ y S corregidos.
**4)** Calcular el error estándar del promedio: $\xi = \frac{s}{\sqrt{n}}$

**5)** Escríbase el resultado: $\overline{X}$ ± ξ interpretándolo correctamente:
a) El valor mas probable de X es $\overline{X}$.
b) La probabilidad de que el verdadero valor esté en el intervalo $\overline{X}$ ± ξ es del 68% (la probabilidad de que al analizar una nueva serie de medidas de X, su promedio $\overline{X}$ caiga en $\overline{X}$ ± ξ es del 68%).

*Obs.*: Si escribimos $\overline{X}$ ± S decimos:

- el valor mas probable de X es $\overline{X}$**.**
- la probabilidad de que una nueva observación caiga en el intervalo $\overline{X}$ ± S es del 68%.

**EJEMPLOS**

**1)** Para determinar la longitud entre dos puntos A y B de una base de replanteo para un trabajo de precisión, se ha realizado una serie de observaciones obteniendo los siguientes resultados:

**Muestra 1.**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 127,834 | 127,839 | 127,832 | 127,831 | 127,830 | 127,852 | 127,801 | 127,832 | 127,833 | 127,835 |

Como n = 10, aplicamos el criterio de la t-Student directamente considerando un nivel de confianza del 95%.

$\overline{X}$ = 127.8319 $\quad s = \sqrt{\frac{1}{n-1}\sum(\bar{X} - X_i)^2}$ = 0.0126

n = 10; α = 0,05 → G critico = 2,28995

Calculando $G = \frac{|X_i - \bar{X}|}{s}$ para cada valor observado, se considera outlier aquel que supere al crítico.

→ descartamos 127.801

**Muestra 2.**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 127,852 | 127,839 | 127,835 | 127,834 | 127,833 | 127,832 | 127,832 | 127,831 | 127,830 |

$\overline{X}$ = 127.8353 S = 0.00678

n = 9; α = 0,05 → G critico = 2,2150

Calculando $G = \frac{|X_i - \bar{X}|}{s}$ para cada valor observado, se considera outlier aquel que supere al crítico.

→ descartamos 127.852

***Muestra 3.***

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 127,839 | 127,835 | 127,834 | 127,833 | 127,832 | 127,832 | 127,831 | 127,830 |

$\overline{X}$ = 127.8333 S = 0.00282

n = 8; α = 0,05 → G critico = 2,12664

Calculando $G = \frac{|X_i - \bar{X}|}{s}$ para cada valor observado, se considera outlier aquel que supere al crítico.

→ No descartamos ninguno

$\xi = \frac{s}{\sqrt{n}}$= 0.00282 / √ 8 = 0.000997

Resultado: 127.833 ± 0.001 m ($\overline{X}$ ± ξ )

127.833 ± 0.003 m ($\overline{X}$ ± S )

**2) Caso Grupo 3, distancia R2-R4**

| **R2 - R4** |
|---|
| 30,036 |
| 30,031 |
| 30,032 |
| 30,028 |
| 30,016 |
| 30,300 |

$\overline{X}$ = 30,074 $s = \sqrt{\frac{1}{n-1}\sum(\bar{X} - X_i)^2}$= 0.111

| **R2 - R4** | **G** |
|---|---|
| 30,036 | 0,3408 |
| 30,031 | 0,3859 |
| 30,032 | 0,3769 |
| 30,028 | 0,4129 |
| 30,016 | 0,5210 |
| 30,300 | 2,0374 |

n = 6; α = 0,05 → G critico = 1,8871 → descartamos 30,300

**La serie debe observarse nuevamente**