الفصل الثامن

معالجة البيانات غير الطبيعية

Handling Non-Normal Data

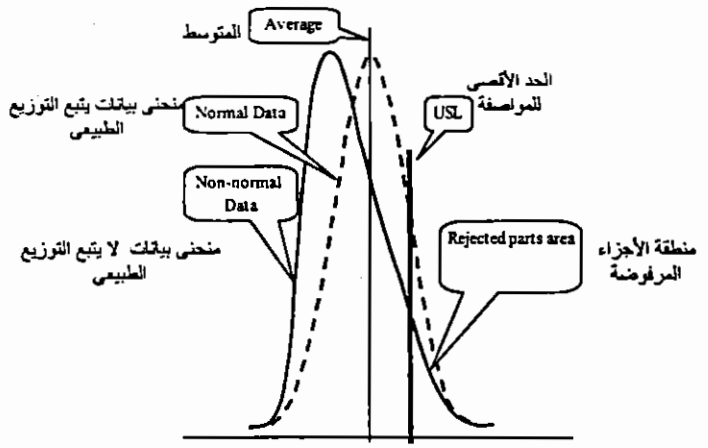
8.1. مقدمة Introduction

في الأجزاء السابقة عندما كنا نتعامل مع البيانات المتاحة بهدف دراسة مقدرة العملية Process Capability مثلا، كان من أهم الافتراضات الأولية هو كون هذه البيانات تتبع منحنى التوزيع الطبيعي، وهو افتراض كما نعلم مثالي، إذ أنه في الحقيقة قلما وجدنا مثل هذا النوع من البيانات، وإنما توجد البيانات بوفرة على غير هذه الصورة.

فماذا سنفعل ؟ هل سنتعامل مع هذه البيانات بنفس الطريقة وبنفس الأدوات التي كنا نتعامل بها مع البيانات التي تتبع التوزيع الطبيعي ؟ أم أن ذلك سلوك محفوف بالخطورة؟ خطورة الخطأ في التحليل، ومن ثم خطورة الخطأ في الاستنتاج وما يرتبط به من الخطأ في اتخاذ القرارات، وما يليه من نتائج غير مرغوب بها.

8.2. محاذير معالجة البيانات غير الطبيعية Precautions for Non normal Data Treatment

في كل ما ذكرناه فى الفصل السابق عن جودة العملية وقدرتها وأدائها ، كنا نفترض أن البيانات المتاحة هي بيانات ذات توزيع طبيعى Normal Data، ولكن ماذا سيكون الوضع لو كانت البيانات المتاحة هي البيانات غير طبيعية Non-Normal Data ؟



شكل رقم 8-1 اختلاف مناطق الرفض باختلاف المنحنى

ولبيان مدى خطورة التعامل مع البيانات غير الطبيعية Non-Normal Data على أنها بيانات ذات توزيع طبيعى Normal Data دعنا ننظر إلى الشكل 8-1، وفيه نرى بوضوح اختلاف

المساحة الممثلة للأجزاء المعيبة في حالة المنحنى الممثل لبيانات تتبع التوزيع الطبيعى ، عن المساحة الممثلة للأجزاء المعيبة في حالة المنحنى الممثل للبيانات غير الطبيعية،وبالتالي فان قيمة مستوى الجودة سيختلف باختلاف الفرض المتعلق بنوعية البيانات .

وللحيلولة دون الوقوع هذا الخطأ ، فسوف نقوم بمعالجة هذه النوعية من البيانات بخمس طرق غير تقليدية، (غير تلك المتبعة عند التعامل مع البيانات التى تتبع التوزيع الطبيعى والتى تنبنى عليها كافة النظريات الإحصائية)، تمكننا من التعامل معها بالرغم من كونها لا تتبع التوزيع الطبيعى Non-Normal Data .

8.3. العلاقة بين طرق المعالجة الإحصائية وطبيعة البيانات Relationship Between Statistical Treatment & Data Type

إذا كانت عملية معالجة البيانات غير الطبيعية عملية محفوفة بالمخاطر، وإذا كان الموضوع بهذه الأهمية، فمن المنطقى أن نتعرف على الطرق التي تتأثر بنوع البيانات المستخدمة، وتعطى تحليلات خاطئة إذا استخدمت مع البيانات غير الطبيعية Non-Normal Data، وهنا يجب ان نشير الى أن هذه الطرق هى التى تتعامل مع البيانات الفردية Individual Data و يمكن تلخيصها في الجدول 8-1:

جدول رقم 8-1 مخاطر معالجة البيانات التى لا تتبع التوزيع الطبيعى

تأثير استخدام البيانات غير الطبيعية Non-Normal Data	الطريقة	
تعطى قيما خاطئة لقيمة مستوى الجودة Sigma	Process Sigma Calculation	حساب مستوى السيجما
حيث سيتأثر كل من LCL & UCL حسابيا بنوع البيانات، وعليه فتحديد الأسباب الخاصة Special Causes سيكون خطأ ، وبالتالي فالعبث و التدخل الغير مطلوب في العملية وهو ما نطلق عليه Tempering أي محاولة إصلاح العملية في حين أنها سليمة ولا تحتاج للتدخل	Individual Control Charts	خرائط المدى المتحرك
تصور خاطئ للفروق بين Groups	Hypothesis testing	اختبار الفرضيات
تحديد غير سليم للعوامل الهامة واستنتاج غير سليم لتأثير هذه العوامل	Regression	الانحدار
تحديد غير سليم للعوامل الهامة واستنتاج غير سليم لتأثير هذه العوامل	Design of Experiment	تصميم التجارب

وإذا كان الموضوع بهذه الأهمية أيضا، فما أنواع العمليات التي تعطى نتائجها بيانات من خصائصها أنها بيانات غير طبيعية Non-Normal Data ؟ والإجابة على هذا السؤال أيضا نجد أن العمليات التي تعطى نتائجها بيانات غير طبيعية كثيرة ومتعددة ومنها على سبيل المثال زمن انتظار العملاء Customer Waiting Time، وعدد المكالمات التليفونية فى الساعة Calls per Hour ،والدورات الزمنية Cycle Times، وزمن وصول الموظفين .

8.4. اختبار تبعية البيانات للتوزيع الطبيعى (الطبيعية) Tests for Normality

وإذا كان الموضوع بهذه الأهمية أيضا، فكيف يمكننا تحديد ما إذا كانت البيانات تتبع التوزيع الطبيعى Normal أم أنها لا تتبع التوزيع الطبيعى Non-Normal ؟

و يتم ذلك من خلال عدة وسائل كما يلي:

1. من خلال رسم الأشكال البيانية Graphs مثل الهستوجرام Histogram، ومخطط الصندوق مخطط الصندوق Box Plot، ومخطط الورقة والساق Stem and leaf ، واختبار الطبيعية..... Normality Test
2. من خلال المقاييس الرقمية Measure مثل القيمة بى P Value، ولكي تكون البيانات طبيعية Normal، فيجب أن تكون P>0.05.

ومن أجل نتائج موثوق فيها يجب ألا نكتفي بأي من الأشكال البيانية Graphs أوالمقاييس الرقمية Measures فقط، ولكن ينبغي اختبارهما معا، أي يفضل دائما إثبات أي فرضية بأكثر من طريقة، ولكي لا ننسى هذا المفهوم ينبغي أن نتذكر دائما مثال الصاروخ الذي يتبع الطائرة ويريد إصابتها، فلكي نتأكد من إصابة الطائرة نزود الصاروخ بالأشعة فوق الحمراء Infra Red، ولمزيد من التأكد ولزيادة احتمال الإصابة نزوده كذلك بالكاشف الليزرى Laser Detector، ولمزيد من التأكد ولزيادة احتمال الإصابة نزوده كذلك بالكاشف الحرارى Thermal Detector، فإذا أخطأت إحدى هذه الطرق، كان هناك احتمال للاصابه بطريقة أخرى، وهذا ما يجب أن يتم في دراستنا، فنحن نقوم بإطلاق فرضية ما، ثم نحاول إثبات هذه الفرضية، وكلما زادت الطرق التي نحاول بها ذلك كلما كان احتمال الدقة في الاستنتاج أكبر وكلما كان احتمال الخطأ أقل.

8.5. طرق معالجة البيانات غير الطبيعية Non-Normal Data Treatment Methods

وإذا كان الموضوع بهذه الأهمية أيضا، فهل هناك طرقا أخرى للتعامل مع مثل هذا النوع من البيانات غير الطبيعية Non-Normal Data؟

نعم هناك طرق أخرى للتعامل مع البيانات غير الطبيعية ويمكن إجمال هذه الطرق في الجدول 8-2:

جدول رقم 8- 2 الطرق المختلفة لمعالجة البيانات التي لا تتبع التوزيع الطبيعي

الطريقة	خصائص هذه الطريقة	
Subgroup averaging	وفيها يتم حساب متوسطات البيانات المتاحة، وطبقا لنظرية Central Limit theorem فإن البيانات الناتجة تتبع Normal data. ويتم تطبيق ذلك خصوصا في X-bar	
Segmenting or Stratification	في كثير من الحالات، فإن تقسيم البيانات إلى مجاميع وشرائح وفئات يؤدى إلى تحويلها إلى بيانات طبيعية .	
Transforming data	ويتم التحويل باستخدام إحدى الدوال واشهرهم Square root & Lognormal & inverse& Binomial& Poisson، ويتم تحديد نوع التوزيع بعد تحديد قيمة	
Using different distributions	يتم تحديد نوع التوزيع الذي تتبعه البيانات، وذلك برسم Probability للتوزيعات المختلفة، واختيار النوع الذي له أكبر P value وإجراء الحسابات بناءا عليه.	
Non parametric statistics	طرق حسابية لا تعتمد في حساباتها على فرضية أن تكون البيانات Normal Data، وفيها	
	Normal Data	**Non Normal Data**
	One Sample Z test	One Sample Sign
	One Sample t test	One Sample Wilcoxon
	Two Sample t test	Mann-Whitney
	One Way ANOVA	Kruskal-Wallis (Moods Medians)
	Randomized Blocks (Two Way ANOVA)	The Fridman test

وعلى ما سبق فيمكننا تلخيص خطوات التعامل مع البيانات غير الطبيعية Non-Normal Data كما يلي:

أ- نختبر ما إذا كانت البيانات Data المتاحة لدينا تتبع التوزيع الطبيعى ، أم لا؟.

ب- إذا كانت البيانات لا تتبع التوزيع الطبيعى، نختار طريقة التعامل معها من أحد الطرق السابق الإشارة اليها فى جدول 8-2.

ولتوضيح ما سبق ولمعرفة كيفية استخدام هذه الطرق سنتناول مثالا نقوم فيه (بالرغم من معرفتنا أن البيانات المتاحة لا تتبع التوزيع الطبيعىNon-normal) بدراسة مقدرة العملية Process capability مرة بالطرق التي التقليدية التي تتعامل مع بيانات تتبع التوزيع الطبيعى ، ومرة أخرى بالطرق التي ينبغي أن نتعامل بها مع البيانات غير الطبيعية، مع توضيح الخطأ الذي يصاحب معاملة النوع الأول على أنه من النوع الثانى .

مثال رقم 8- 1

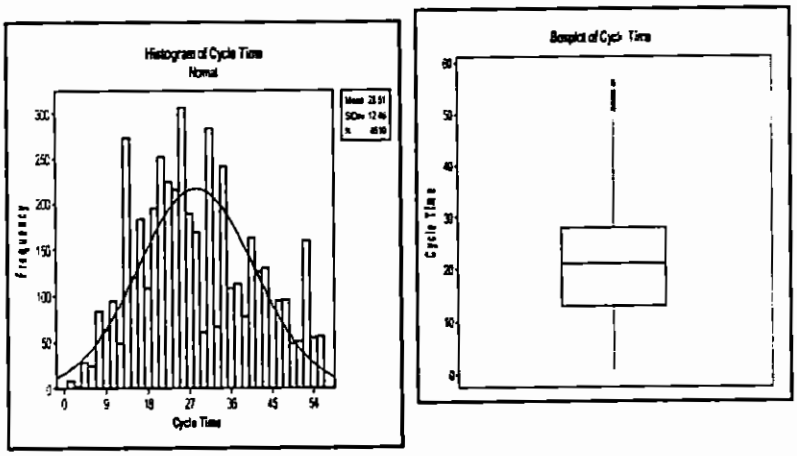
بفرض أن لدينا مجموعة من البيانات كما فى الجدول 8-3 تعبر عن زمن الإنتظار Waiting Time بصورة عشوائية بحيث USL =30 & LSL=0.

جدول رقم8-3 بيانات المثال 8-1

19	11	30	24	24	20	28	27	26	20	9	32	56	48	27	24	21	24	31	33
17	53	32	33	39	26	20	48	21	34	31	15	40	27	24	22	14	13	13	14
36	43	42	43	41	40	35	24	21	23	14	43	37	18	17	47	10	13	14	22
22	20	25	39	26	53	19	13	27	28	8	54	8	25	8	19	18	9	3	32
35	11	42	38	32	27	24	22	18	17	21	16	6	36	36	9	21	7	28	28
17	15	15	9	51	26	25	13	47	37	20	17	25	15	21	10	11	6	4	8
52	17	4	8	9	18	16	5	49	11	21	23	22	5	5	21	15	13	14	6
8	31	14	31	24	19	41	2	25	1	12	34	15	14	7	6	9	14	23	18
8	7	16	34	21	9	14	31	16	14	7	10	14	26	12	28	30	26	34	14
5	2	10	42	21	15	21	11	15	22	25	17	13	18	19	21	27	27	23	13
12	2	2	24	35	12	28													

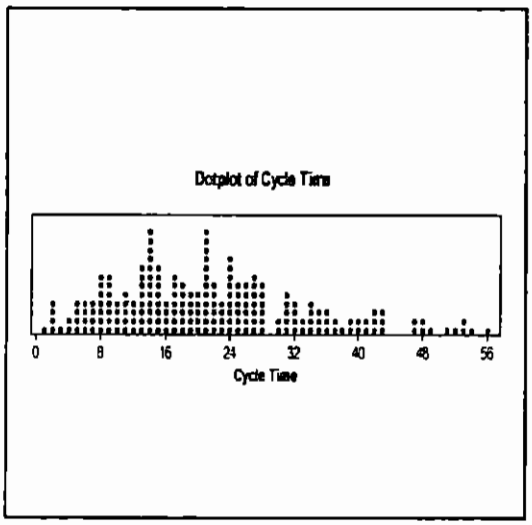
الحل باعتبار أن البيانات تتبع التوزيع الطبيعى:

أولا وللتأكد من كون البيانات المتاحة طبيعية أم لا؟ نرسم لها الأشكال البيانية التالية وهنا يجب أن نتذكر ما قلناه عن ضرورة استخدام أكثر من طريقة لاختبار طبيعية البيانات Normality، فلو نظرنا إلى شكل 8-2 والذى يظهر فيه الهستوجرام Histogram، أومخطط الصندوق مخطط الصندوق Box Plot، لقلنا أن هذه البيانات تتبع التوزيع الطبيعى Normal تقريبا.



شكل رقم 8-2 الهستوجرام و مخطط الصندوق

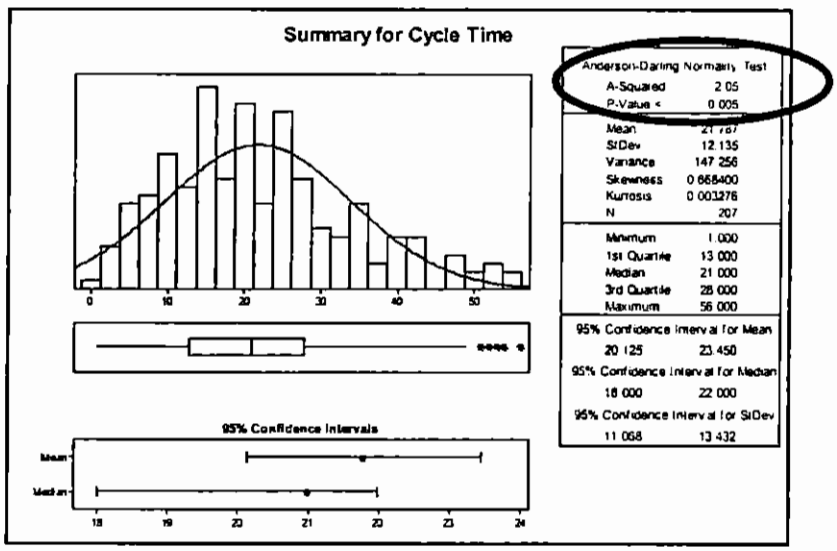
ولكن برسم مخطط الساق والورقة Stem and leaf وكذلك مخطط النقطةDot plot كما فى شكل 8-3،وايضا مخطط الملخص البيانىGraphical Summary كما فى شكل 8-4، و يظهر لنا أنها لا تتبع التوزيع الطبيعى Non normal، وأيضا لو نظرنا إلى مخطط اختبار الطبيعية Normality plot فى شكل 8-5، و إلى قيمة بى P value وهى هنا أقل من 5 % لظهر لنا أنها لا تتبع التوزيع الطبيعى Non normal.



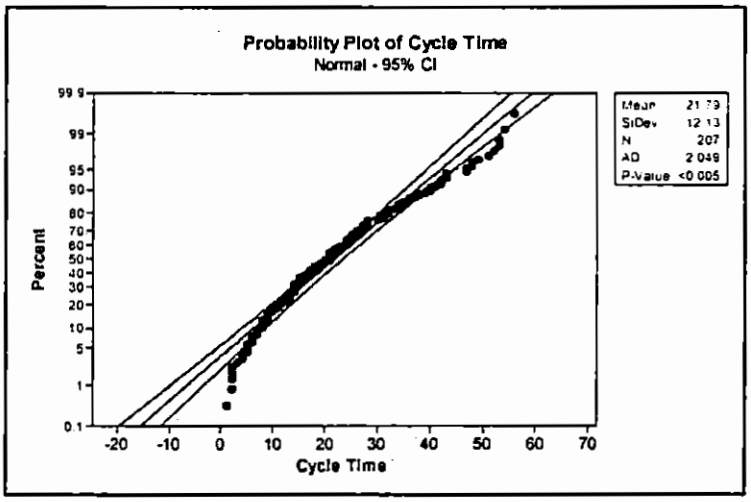
Stem-and-Leaf Display: Cycle Time
Stem-and-leaf of Cycle Time N = 207
Leaf Unit = 1.0

8	0	12222344
34	0	55556666777788888889999999
67	1	000011111222233333333444444444444
97	1	55555555666677777788888899999
(36)	2	000001111111111112222223333444444444
74	2	55555566666677777777888888
49	3	00111112222334444
32	3	55566677899
21	4	0011222333
11	4	77889
6	5	12334
1	5	6

شكل رقم 8-3 الساق والورقة و مخطط النقطة



شكل رقم8-4 مخطط الملخص البيانى للمثال 8-1



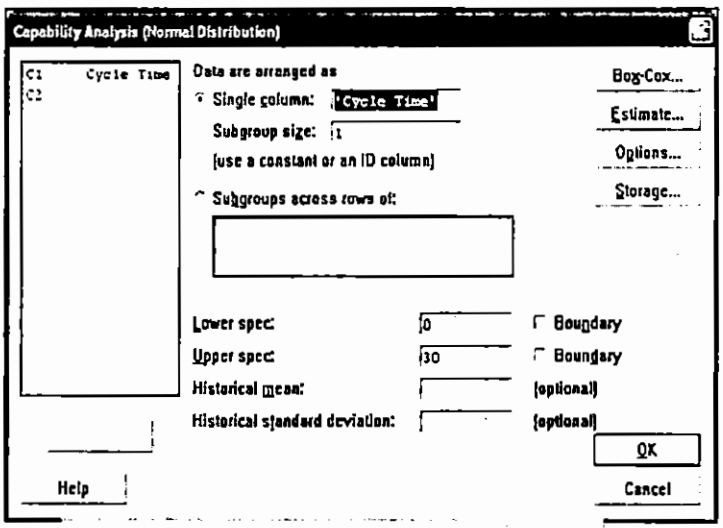
شكل رقم 8-5 مخطط اختبار الطبيعية للمثال 8-1

الحل باعتبار أن البيانات غير طبيعية:

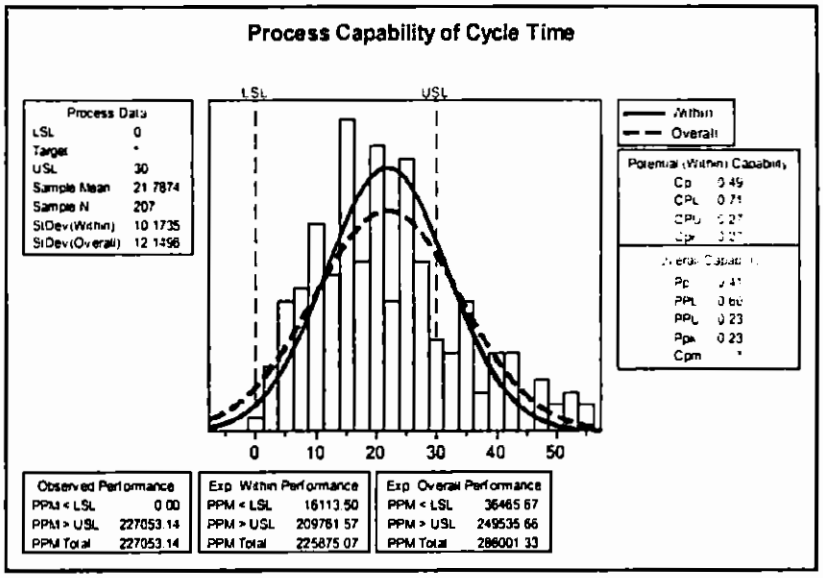
فى هذه الحالة، نقوم بدراسة مقدرة العملية Process Capability باعتبار البيانات المتاحة تتبع التوزيع الطبيعى Normal، ومن برنامج المينيتاب نختار القوائم التالية :

Minitab>Stat> Quality Tools> Capability Analysis > Normal

يظهر الشكل 8-6 وفيه نختار Cycle Time ونضع Subgroup=1 ثم ندخل كلا من USL=30 & LSL=0، فيظهر الشكل 8-7.



شكل رقم8-6 خطوة لحل المثال 8-1 بالمينيتاب



شكل رقم 8-7 خطوة لحساب مقدرة العملية بفرض ان البيانات طبيعية

ومن الشكل 8-7 يتضح أن العملية غير قادرة على الوفاء بمتطلبات العملاء لوجود قيم كثيرة أكبر من USL، كما سنجد قيما غير منطقية أقل من الصفر إذ أنه لا يمكن قياس الوقت بقيم سالبة، كذلك العدد الذي يقع خارج النطاق المطلوب سيكون286001.33 منتج معيب، وكذلك قيمة كل من قدرة العملية Cpk وأدانها Ppk الأقل من الواحد الصحيح.

ويلاحظ ان الشكل 8-7 يمثل النموذج العام الذى يظهر عند حساب مقدرة أى عملية، وهو يتكون من لوحة رئيسية يظهر فيها منحنى الهستوجرام، يحيط به ستة مربعات وهى:

1. مربع بيانات العملية Process Data وفيه بيانات عامة عن العملية، مثل الحد الأعلى والأدنى للمواصفات والمتوسط وحجم العينة والانحراف المعيارى للمجموعات والانحراف المعيارى للعينة المسحوبة ككل.
2. مربع الأداء الملاحظ Observed Performance وفيه يظهر عدد الوحدات المعيبة جهتى الحد الأدنى والأقصى للعملية.
3. مربيعين لأداء العملية المتوقع Expetecd Performance، أحدهما على المدى القصير Short term أو داخل المجموعات Within، والآخر على المدى الطويل Long term أو لإجمالى المجموعات Overall، وفيه يظهر عدد الوحدات المعيبة جهتى الحد الأدنى والأقصى للعملية.

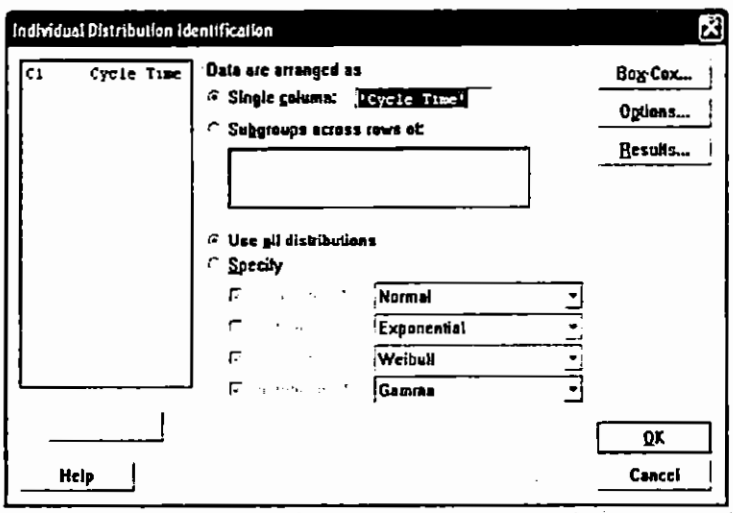
4. مربيعين لقدرة العملية أحدهما على المدى القصير Short term أو داخل المجموعات Within، والآخر على المدى الطويل Long term أو لإجمالى المجموعات Overall، وفيه يظهر قيمة القدرة CpK أو Ppk.

5. مربع الأداء المتوقع على المدى الطويل Expetecd Performance أو الأداءالكلى المتوقع Expetecd Overall Performance

والآن سنقوم بدراسة مقدرة العملية Process Capability باعتبار البيانات المتاحة لا تتبع التوزيع الطبيعىNon normal ، وسنختار الطريقة الرابعة من طرق التعامل مع هذا النوع والتى تم الإشارة إليها فى بداية هذا الفصل، وهى إستخدام أكثر من توزيع Using different distributions وهى اختيار أفضل توزيع يعبرعن البيانات بدقة أكثر من غيره (و ستكون هى التوزيعات التى لها أكبرقيمة للمعامل بى P value) ثم نعتمد على هذا التوزيع distribution في حساب مقدرة العملية Process Capability كما يلي:

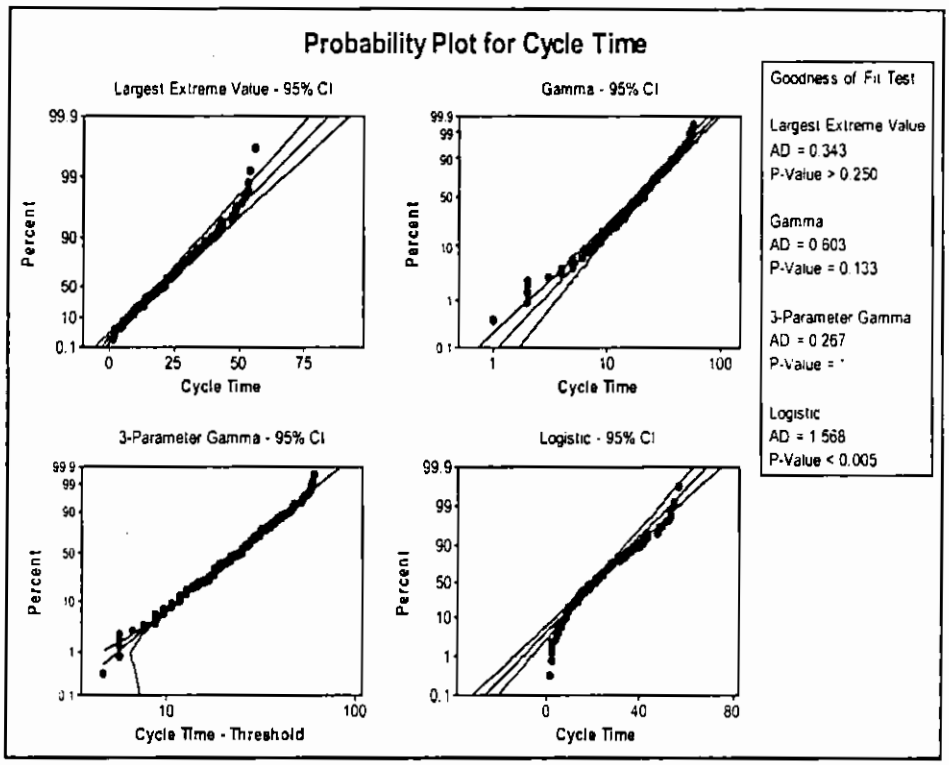
لتحديد نوع التوزيع Distributions التي تعبر عن البيانات بدقة أكثر من غيرها

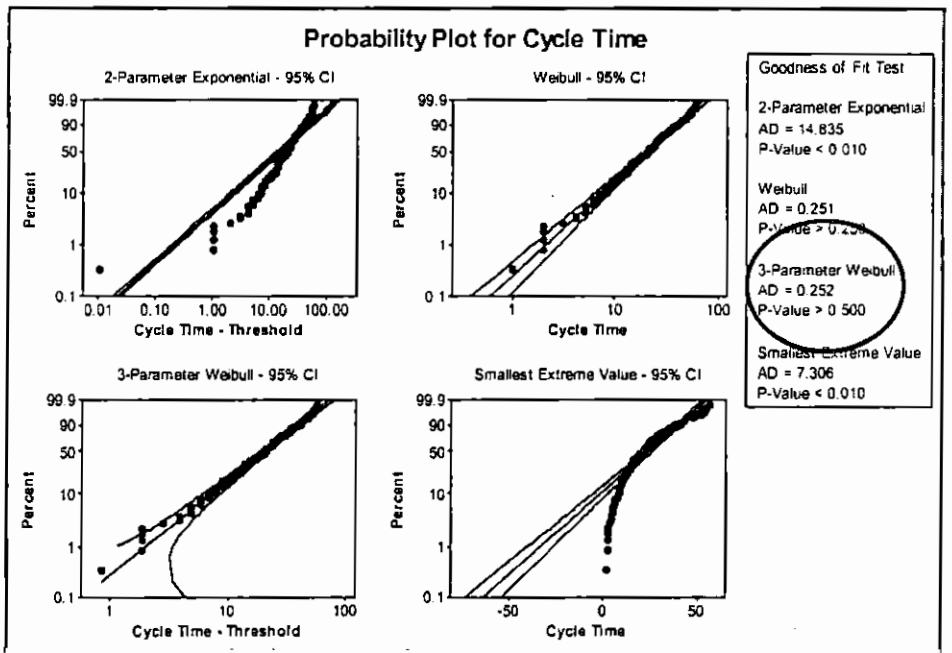
Minitab>Stat> Quality Tools> Individual Distribution Identification



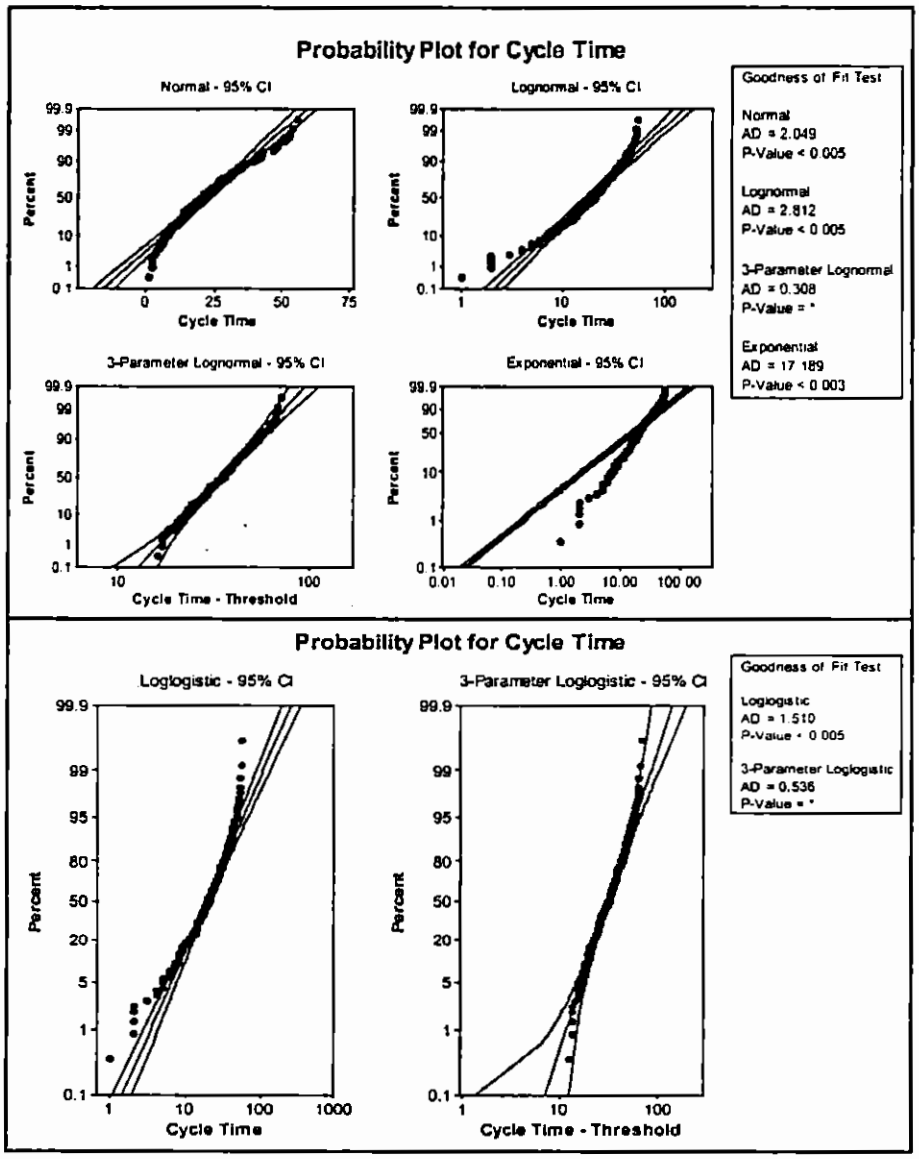
شكل رقم 8-8 خطوة لحل المثال 8-1 بالمينيتاب

فيظهر الشكلان 8-9 و 8-10 .





شكل رقم 8-9 توزيعات المثال 8-1

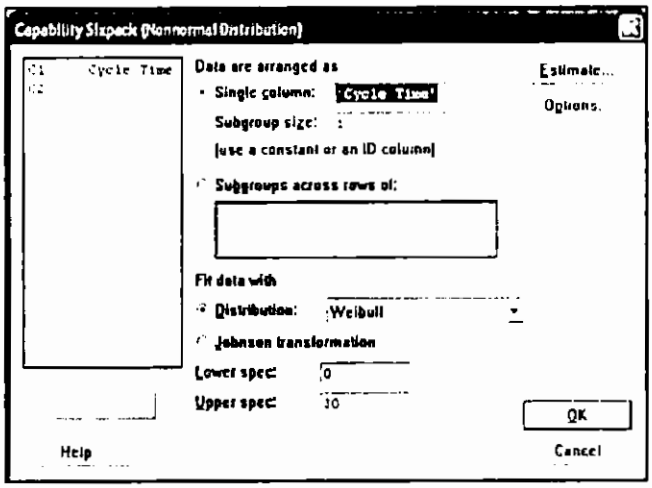


شكل رقم 8-10 باقى توزيعات المثال 8-1

وفى هذه الخطوة قمنا برسم البيانات بكل التوزيعات المتاحة في Minitab وعددها حوالي 14 توزيعا ، لنختار منها أكثرها تعبيرا عن البيانات، وهو التوزيع الذي له أعلى P-value وهو في هذه الحالة توزيع وايبول Weibull.

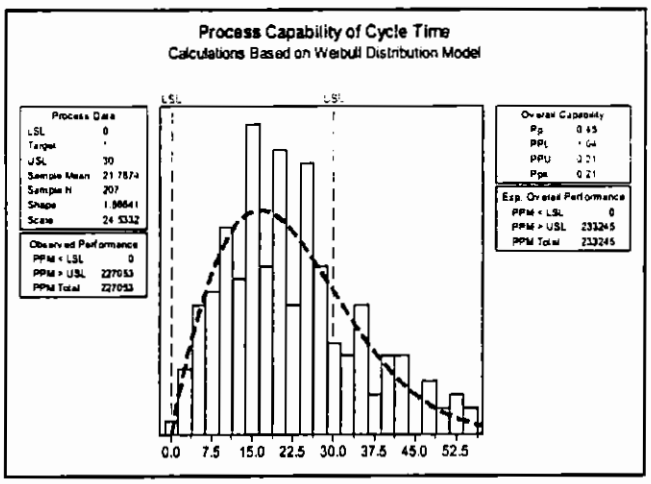
ثم نقوم بدراسة مقدرة العملية Process Capability اعتمادا على أن البيانات تتبع التوزيع وايبول Weibull، ومن قوائم مينيتاب نختار القوائم التالية :

Minitab>Quality Tools> Capability analysis>Non normal



شكل رقم 8-11 خطوة لحل المثال 8-1

فيظهر شكل 8-11 ، ومنه نختار Cycle Time ونضع Subgroup=1 ونختار Weibull ثم ندخل كلا من LSL=0 & USL=30



شكل رقم 8-12 مقدرة العملية

فيظهر الشكل 8-12، وبالنظرإلى قيمة العدد الذي يقع خارج النطاق المطلوب نجد أنه يساوى 233245 وهو أقل من ذلك الرقم الذي حصلنا عليه عندما اعتبرنا البيانات المتاحة تتبع التوزيع الطبيعى Normal وهو 286001.33. أضف إلى ذلك أن الأرقام الأقل من الصفر في هذه الحالة (وان وجدت) هي أقل من الطريقة السابقة.

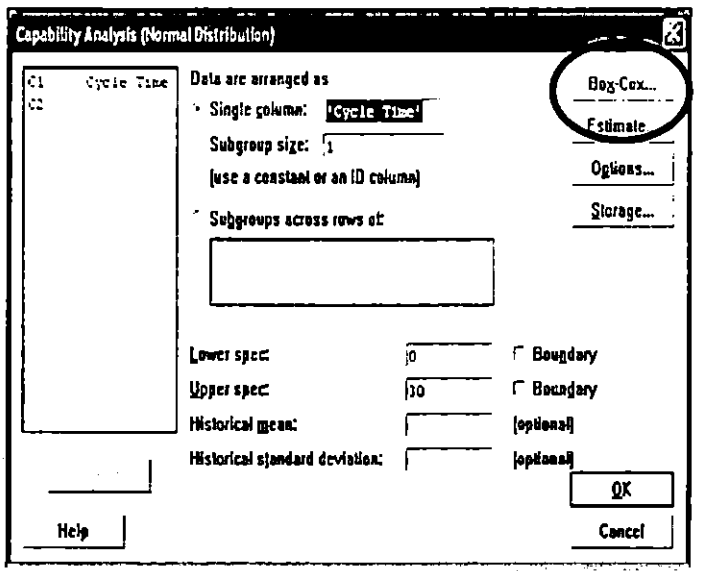
حل آخر للمثال 8-1 باستخدام بوكس كوكس:

مرة أخرى سنتناول فيها نفس المثال، نقوم فيه (بعد أن نكون قد أثبتنا أن البيانات المتاحة هي Non-normal) بدراسة مقدرة العملية Process Capability بعد تحويل البيانات غير الطبيعية Non-Normal Data باستخدام الطريقة الثالثة وهى التحويل باستخدام بوكس كوكس Transformation by using Box-Cox plot، وتعتمد على حساب واختيار أفضل قيمة لأس معادلة التحويل التالية:

$$(Y)=x^{\lambda}$$

وفيها نقوم باختيار أفضل قيمة للمعامل Lambda، وأفضل قيمة له هى القيمة التى يكون عندها الميل مساويا للصفر، أى عند نقطة التماس مع الخط الأفقى، ونقوم بدراسة مقدرة العملية Process capability عند هذه القيمة كما يلي:

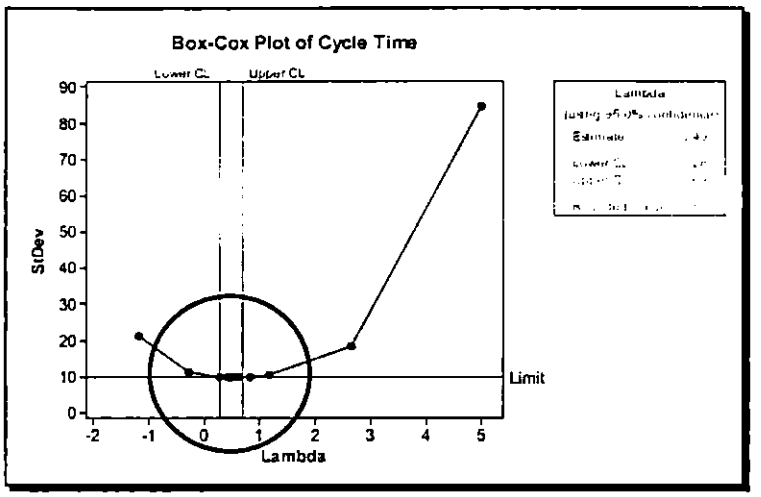
Minitab>Stat> Control Charts> Box-Cox transformation



شكل رقم 8-13 استخدام بوكس كوكس لحل المثال 8-1

، و حيث إن البيانات المتاحة هي individuals فإن subgroup size = 1 ، فيظهر الشكل 8-13 .

وفى الشكل 8- 14 نجد فيه أن أفضل قيمة لمعامل Lambda هي 0.5 ، أي أن أفضل معادلة تحويل Transformation Equation هي $(Y)=X^{0.5}$.

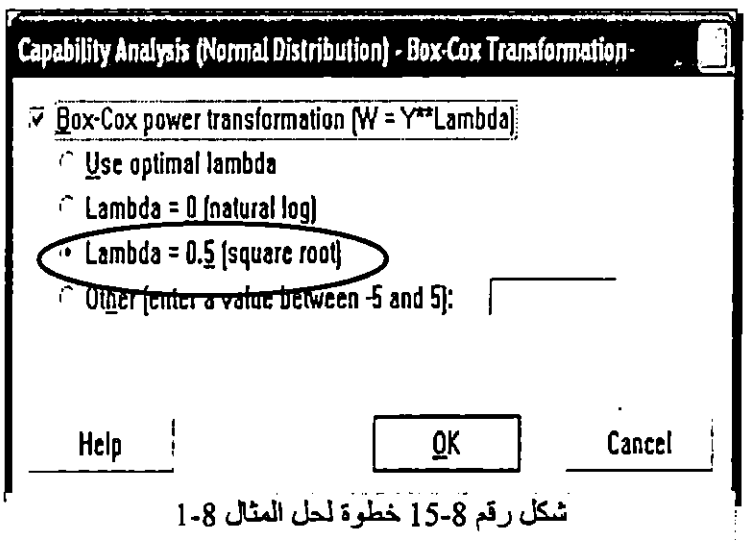


شكل رقم 8-14 تحديد أفضل قيمة للمعامل لامدا

ثم ندرس العملية Process capability عندما Lambda=0.5

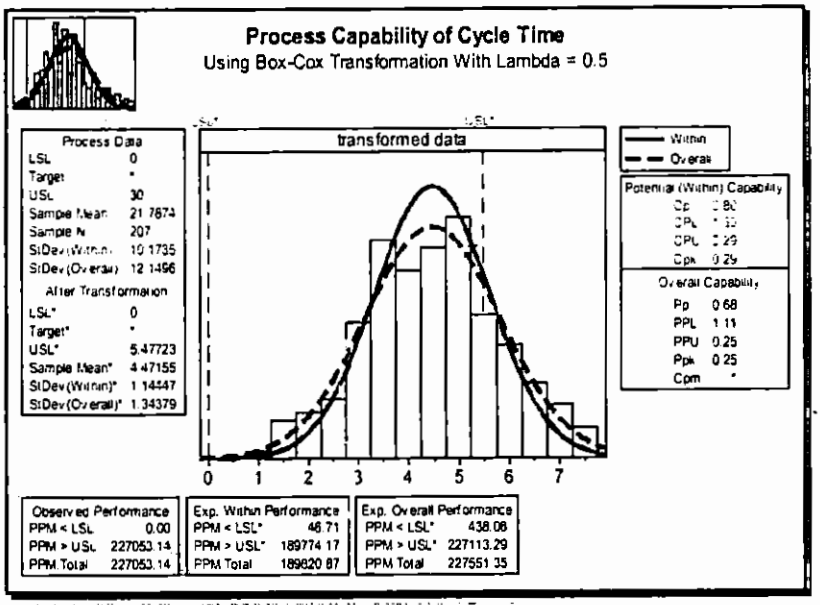
Minitab>Stat > Quality Tools> Capability Analysis> Normal.

نختار C1، و Subgroup=1، ثم ندخل كلا من LSL=0 & USL=30، ثم نختار Box-Cox



شكل رقم 8-15 خطوة لحل المثال 8-1

فيظهر الشكل 8-15، نختار فيه قيمة Lambda=0.5، فتظهر النتيجة كما بالشكل8-16



شكل رقم 8-16 نتيجة حل المثال بطريقة بوكس كوكس

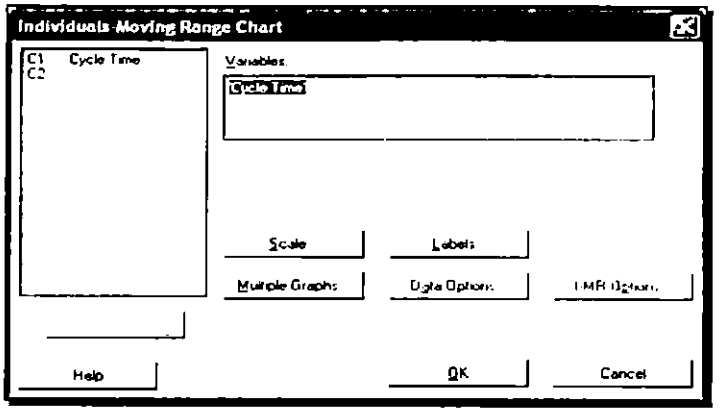
وفى هذا الشكل نجدد أن قيمة العدد الذي يقع خارج النطاق المطلوب تساوى 227551.35، وهو قريب جدا من الرقم الذي حصلنا عليه عندما استخدمنا Weibull Distribution وهو 233245، وكلاهما أقل من ذلك الرقم الذي حصلنا عليه عندما اعتبرنا البيانات المتاحة Normal وهو 286001.33.

حل ثالث للمثال 8-1 بخريطة المدى المتحرك:

للمرة الثالثة سنتناول نفس المثال، في محاولة لإثبات خطأ معاملة البيانات غير الطبيعية على أنها بيانات ذات توزيع طبيعى ، وسنقوم فيه (بعد أن نكون قد أثبتنا أن البيانات المتاحة لا تتبع التوزيع الطبيعى) برسم بخريطة المدى المتحرك I moving range chart مرتين، مرة بالبيانات التي لا تتبع التوزيع الطبيعى ونرى شكل الخرج، ومرة أخرى بعد تحويل البيانات عن طريق التحويل بوكس كوكس Box-Cox Transformation وتخليق بيانات جديدة تماما، ثم نتعامل مع هذه البيانات الجديدة، ونرسم I moving range Chart مرة أخرى.

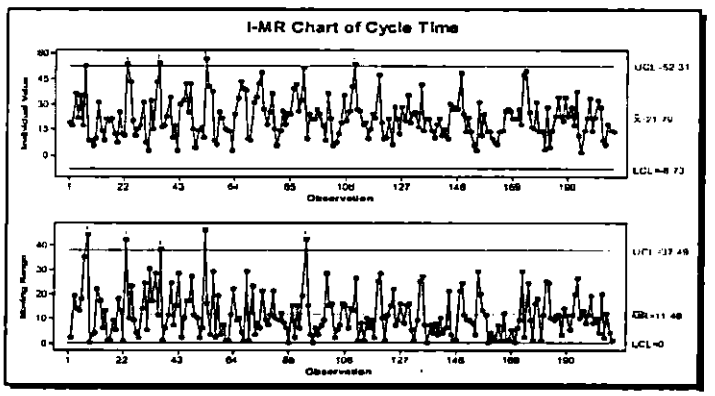
أولا نرسم I-MR Chart للبيانات المتاحة كما يلي :

Minitab> Stat> Control Charts> Variable charts for individuals> I-MR



شكل رقم 8-17 خطوة لحل المثال 8-1

يظهر الشكل 8-17 فنختار Cycle Time ثم OK، فيظهر الشكل 8-18، ونلاحظ أنه ممتلئ بالأسباب الخاصة للتباين Special Causes، أي أن العملية تكون خارج السيطرة Process is Out of control، وفى هذا دعوة رسمية، ومبررة، للتدخل في Process، لمحاولة كشف النقاب عن مصادر هذه الأسباب الخاصة Special Causes.



شكل رقم 8-18 خريطة المدى المتحرك

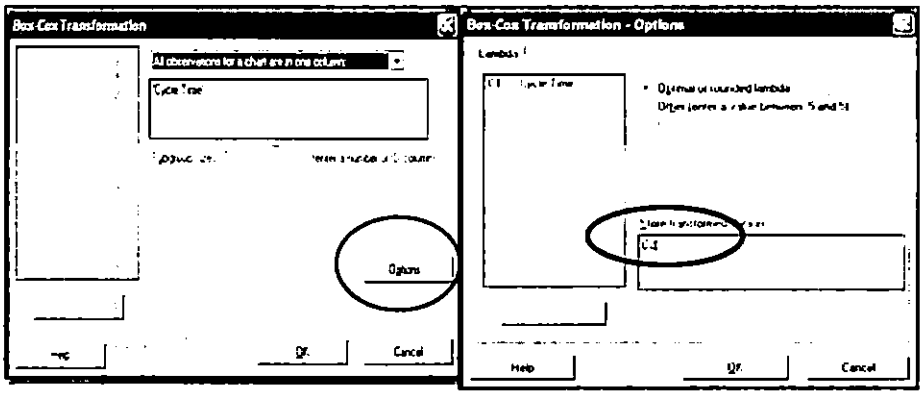
وسيظهر لنا فى نهاية حل المثال أن ذلك ليس صحيحا، وان هذا التدخل، وهو ما نسميه العبث أو التدخل غير المبرر Tempering، قد يؤدى إلى:

1- إذا كانت العملية متزنة، فقد يؤدى هذا التدخل الغير مطلوب، إلى الجنوح بالعملية حتى تفقد اتزانها وتكون خارج السيطرة ، أو

2- إذا كانت العملية غير متزنة ، فحتما سيؤدى هذا التدخل إلى أن يزيدها سوءا إلى سوء(يرجى الرجوع إلى الفصل الرابع تحت عنوان ظاهرة التباين الطبيعى فى هذا الكتاب).

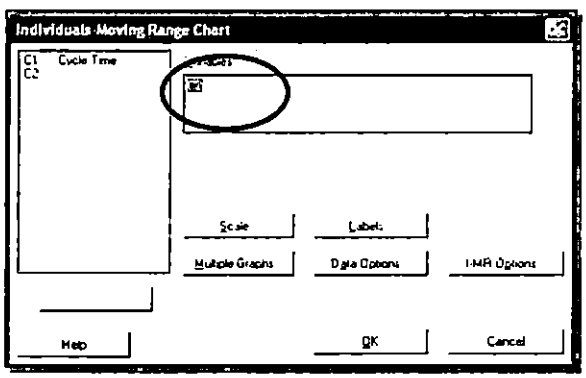
والآن سنقوم بتحويل البيانات المتاحة إلى بيانات جديدة كما يلي

Minitab>Stat> Control Charts> Box-Cox transformation



شكل رقم 8-19 تحويل البيانات باستخدام بوكس كوكس

و حيث إن البيانات المتاحة فردية Individuals فإن حجم المجموعات يساوى واحداً Subgroup Size = 1 ثم نختار Options ، فيظهر الشكل 8-20، نطلب فيه تخزين البيانات المحولة في مكان ما فى مينيتاب، وليكن C2 .

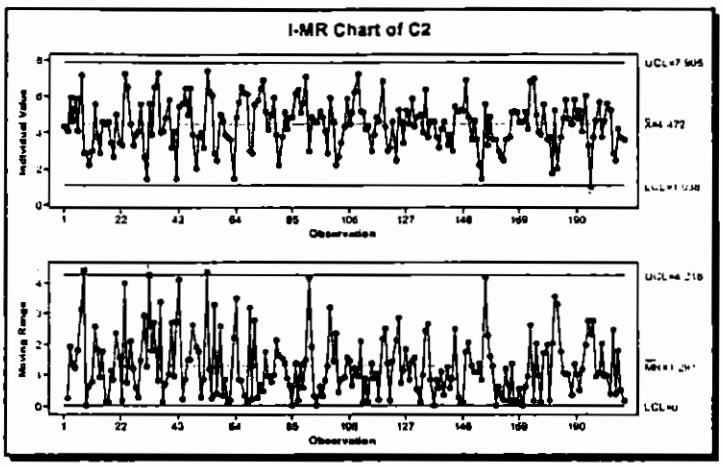


شكل رقم 8-20 خطوة لحل المثال 8-1

ثم نقوم برسم خريطة المدى المتحرك I-MR Chart للبيانات المحولة والمخزنة في C2 كما يلي:

Minitab> Stat> Control Charts> Variable charts for individuals> I-MR

نختار C2، فيظهر الشكل 8-22، ونرى فيه بوضوح أن حالة العملية أفضل بكثير، مما أظهرته لنا نفس خريطة التحكم Control Chart بدون تحويل البيانات، وبالتالي سنحتاج لتدخل أقل، وإجراءات تصحيحية أقل.

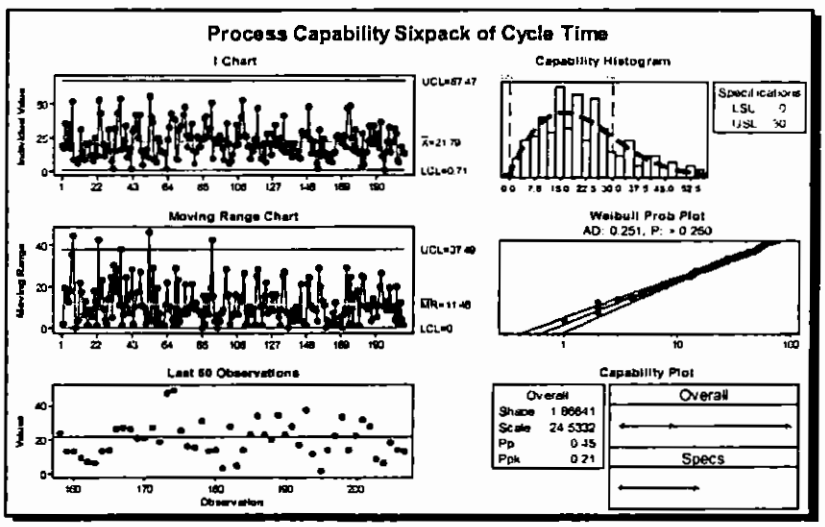


شكل رقم 8-21 خريطة المدى المتحرك للبيانات المعدلة

في الجزء البسيط السابق، تعمدنا الشرح بالتفصيل حتى نيسر فهم الموضوع، والآن وبعد أن استوعبنا ذلك فإنه يمكننا دمج ما سبق في خطوة واحدة، وحينئذ يمكن الحصول على قيم قريبة جدا لما حصلنا عليه سابقا، فبعد أن نكون حددنا نوع التوزيع الذي تتبعه البيانات كما سبق وأوضحنا، نقوم بعمل الاتى :

Minitab>Stat> Quality Tools> Capability Six-pack>Non normal

ثم نختار Cycle Time ونختار Weibull ثم ندخل كلا من LSL=0 & USL=30، فيظهر الشكل 8-22 ، الذي يضم أكثر من مقياس في لوحة واحدة



شكل رقم 8-22 خريطة المدى المتحرك للبيانات المعدلة