

الأكاديمية العربية الدولية



الأكاديمية العربية الدولية
Arab International Academy

الأكاديمية العربية الدولية المقررات الجامعية

تحليل البيانات

إعداد: سامح محمد

ماجستير إدارة أعمال

هذه المقالات تم نشرها في:

موقع الإدارة والهندسة الصناعية

<http://samehar.wordpress.com>

2010

حقوق النشر محفوظة للمؤلف

المحتويات

3	التعامل مع البيانات
6	تخيص البيانات - 2
9	تخيص البيانات باستخدام برنامج إكسل
14	دراسة العلاقة بين متغيرين
19	معامل الارتباط Correlation
25	تحليل الانحدار Regression Analysis
32	تحليل الانحدار الخطي المتعدد Multiple Linear Regression
39	تحليل الانحدار – دراسة البوافي
46	كيف تختار العينة؟ Sampling
53	خرائط المراقبة؟ Control Charts
56	المدرج التكراري؟ Histogram
63	منحنى التوزيع الطبيعي وأهميته Normal Distribution
70	منحنى التوزيع الطبيعي القياسي Standard Normal Distribution
76	منحنى التوزيع الطبيعي القياسي - 2
81	بعض التوزيعات الأخرى
87	نظرية الحد المركزية Central Limit Theorem

التعامل مع البيانات

يونيو 20, 2009

الإدارة السليمة تعتمد على المعلومات السليمة ولذلك فإن المديرين يحتاجون أن يتعاملوا مع البيانات وأن يستبطوا منها معلومات تساعدهم على اتخاذ القرارات. على الجانب الآخر فإن المهندسين الصناعيين يستخدمون البيانات في دراساتهم لتحسين العمليات وهم يستخدمون أدوات متقدمة لتحليل البيانات.

أحاول في هذه المقالة والمقالات التالية تقديم بعض الأساليب البسيطة لتلخيص البيانات واستخراج معلومات مفيدة منها.

تلخيص البيانات

افترض أنك حصلت على معلومات حول درجات الطلاب في الامتحان أو درجات حرارة الجو في كل يوم خلال شهر ما أو حجم المبيعات في كل شهر خلال العام أو عدد العيوب في المنتج خلال فترات دورية أو ما شابه ذلك. كيف يمكننا شرح هذه البيانات باستخدام عدد قليل من الأرقام؟ كيف يمكننا التعليق على هذه البيانات ومقارنتها ببيانات فترة أو فترات سابقة؟ قد يتadar لذهنك المتوسط الحسابي وهو فعلاً أحد وسائل وصف البيانات ولكن هناك أرقاماً أخرى تصف لنا هذه البيانات.

المتوسط الحسابي Average or Mean: المتوسط الحسابي هو من الأرقام التي يشيع استخدامها نظراً لبساطته ولكونه مؤمراً عاماً يالتعامل مع البيانات يمكن أن يعبر عن كل البيانات. ولكن المتوسط الحسابي لا يعبر تماماً عن البيانات ولذلك فإن هناك أرقاماً أخرى لها أهميتها.

المتوسط الحسابي سهل في حسابه فهو حاصل قسمة مجموع البيانات على عددها. على سبيل المثال لو كانت البيانات عبارة عن درجة حرارة الجو خلال ست ساعات وكانت درجة الحرارة في كل ساعة كالتالي:

105، 107، 104، 108، 111، 109

فإن المتوسط الحسابي يكون $(109+104+108+111+107+105) / 6 = 107.3$ درجة مئوية وعندما نستطع أن نقول إن درجة الحرارة المتوسطة خلال ست ساعات هي 107.3 درجة مئوية. هذا الرقم يعطينا انطباعاً عاماً عن درجة الحرارة خلال الفترة كلها. ويمكننا استخدام هذا الرقم لمقارنة درجة حرارة اليوم بالأمس.

الوسيط Median: الوسيط هو رقم غير شائع الاستخدام والكثيرون لا يعرفونه أصلاً مع أن دلالته قوية وفهمه أيسر من فهم المتوسط الحسابي. الوسيط هو رقم مشابه للمتوسط الحسابي ولكن يعني ببساطة: إن 50% من البيانات أقل من هذه القيمة و 50% من البيانات أعلى من هذه القيمة. فمثلاً لو كانت درجات تسعة طلبة هي:

84، 87، 90، 95، 80، 83، 90، 77، 64

فإن الوسيط يكون التعامل مع البيانات لأن أربعة طلبوa حصلوا على درجة أعلى من تلك الدرجة وأربعة طلبة حصلوا على درجة أقل من تلك الدرجة. وبالتالي يكون 50% من الطلبة حصلوا على أعلى من 84 و 50% حصلوا على أقل من 84.

وللوسيط أهمية عندما التعامل مع البيانات هناك قيم قليلة متطرفة أي أن قيمتها تبتعد كثيراً عن باقي البيانات. ففي المثال السابق فإن المتوسط الحسابي هو $9760 / 9 = 84.4$ وهي قيمة تقترب كثيراً من قيمة الوسيط. ولكن ماذا لو كانت درجة الطالب الذي حصل على 64 في المثال السابق هي 25 درجة فقط. في هذه الحالة لا يتغير الوسيط ولكن

المتوسط الحسابي يصبح 80 وهي قيمة تقل عن الوسيط بسبب وجود قيمة واحدة قليلة جداً. في هذه الحالة تجد أن الوسيط يعطيك التعامل مع البيانات عن أفضل. وعلى كل حال فإن المتوسط والوسيط قد يستخدمان معاً.

و عند التعليق على البيانات فإنه لا يلزم أن تستخدم لفظ وسيط فقد لا يفهمك الآخرون ولكن يمكنك أن توضح أن 50% من الطلبة حصلوا على أعلى من 84 درجة. وبالطبع فإن المستمع يفهم أن الـ 50% الآخرين حصلوا على أقل من 84 درجة.

قد يكون عدد البيانات هو رقم زوجي مثل أن يكون عدد الطلبة هو 10 وليس 9. في هذه الحالة فإن الوسيط يكون هو متوسط القيمين المتوسطتين.

مثال: درجات الطلبة هي:

70، 70، 80، 85، 90، 95، 100، 10، 80، 86، 96، 90، 100

نقوم بترتيب البيانات تصاعدياً:

10، 70، 80، 85، 86، 90، 92، 95، 96، 100

فيكون الوسيط هو متوسط الرقمان: 86 و 90 أي 88

لاحظ أن المتوسط هنا هو 80.4 نظراً للوجود قيمة متطرفة جداً هي 10.

المنوال Mode: المنوال هو القيمة الأكثر تكراراً في مجموعة البيانات.

مثال: لدينا سرعة عشر سيارات مرت خلال الخمس دقائق الماضية على طريق ما وهي كالتالي:

100، 90، 80، 85، 95، 90، 82، 77، 90، 76

في هذه الحالة يكون المنوال هو 90 لأن هذه السرعة تكررت ثلاثة مرات. وهذا يعني أن سرعة 90 كم/ساعة هي السرعة الأكثر شيوعاً بين هذه السيارات العشر ومن الطبيعي أن تكون سيارات كثيرة سرعتها تقترب من هذه القيمة.

قد يكون هناك أكثر من منوال عند تساوي تكرار قيمتين وقد لا يكون هناك أي منوال عندما لا تكرر أي قيمة.

الربعين الأدنى والأعلى Quartiles: الربعين الأدنى والأعلى هما مماثلان للوسيط ولكنهما يمثلان نسبة 25% و 75% أي القيمة التي يكون 25% من البيانات أقل منها والقيمة التي يكون 75% من البيانات أقل منها.

هذه القيم تساعدنا كثيراً على فهم البيانات ومقارنتها ببيانات مماثلة لفترة سابقة أو مدرسة أخرى أو شركة أخرى وهكذا.

قيمة الربع الأدنى هي قيمة الملاحظة (البيان) المناظر لـ

$0.25 * (\text{عدد المشاهدات} + \text{البيانات})$

أما قيمة الربع الأعلى فهي القيمة المناظرة لـ

$0.75 * (\text{عدد المشاهدات} + \text{البيانات})$

كما في حساب الوسيط فإن قيمة كلاً من الربع الأدنى والأعلى قد تصادف بياناً بعينه أو تكون قيمة متسطة بين قيمتين.

مثلاً إذا كانت البيانات تمثل درجات 1 طالباً فإن قيمة الربع الأدنى تكون هي القيمة المناظرة لـ

$$3 = (11 + 1) * 0.25$$

وقيمة الربع الأعلى تكون القيمة المناظرة لـ

$$9=(11+1)*0.75$$

فلو كانت الدرجات كالتالي:

96، 95، 90، 86، 82، 76، 66، 40، 35

فإن قيمة الربع الأدنى تكون 66 وقيمة الربع الأعلى تكون 90 أي أن 25% من الطلبة حصلوا على أقل من 66 و 75% من الطلبة حصلوا على أقل من 90.

أما لو كان عدد الطلبة 12 مثلاً فإن القيمة المطلوبة تكون قيمة متوسطة بين قيمتين.

وأسلوب الحساب ليس مهمًا بالنسبة لنا لأننا نستخدم الحاسوب والذي سيقوم بحساب كل هذه القيم لك. من أشهر البرامج التي تقوم بذلك برنامج إكسل الواسع الانتشار وبرنامج SPSS المتخصص في الإحصاء. وهناك موقع على الإنترنت قد توفر ذلك مجاناً كذلك. فالمهم هو فهم معنى كل رقم من هذه الأرقام.

و عند استخدام هذه الأرقام في مجال العمل فإنه من الأفضل أن تشرحها في تعليقك على البيانات بمعنى ان تقول أن 75% من الطلبة حصلوا على درجة أعلى من هذا. فإن القارئ أو المستمع ربما ينزعج عند استخدامك لمصطلح لا يعرفه مثل مصطلح الربع الأدنى أو الأعلى ولكنه يستطيع فهم الجملة السابقة بسهولة.

في المقالات التالية إن شاء الله نستكمم الحديث لنوضح أرقاماً أخرى ونبين كيفية استخدام برنامج إكسل للقيام بكل تلك الحسابات.

[مقالات ذات صلة:](#)

[تحليل البيانات](#)

[من مراجع الموضوع](#)

Lean Six Sigma Pocket ToolBook, M. George et al., McGrawHill, 2005

Applied Statistical Methods, W. Carlson and B. Thorne, Prentice Hall, 1997

Statistics for Managers, Levine et al., Prentice Hall, 1999

[موقع ذات صلة بالموضوع](#)

[الإحصاء](#)

[مقاييس الإحصاء الوصفي](#)

[الإحصاء الوصفي](#)

[Free Statistics and Forecasting Software](#)

تلخيص البيانات-2

يونيو 28, 2009

ناقشت في المقالة السابقة بعض الأرقام الإحصائية التي تستخدم لتلخيص مجموعة من البيانات وللمقارنة بين فترات مختلفة. نستكمل في هذه المقالة الحديث فنناقش بعض الأرقام التي تعطينا بعض المعلومات الإضافية عن البيانات.

11	124	624
	790	
51	2222	690
13		

القيمة القصوى والقيمة الدنيا **Maximum and Minimum**: القيمة القصوى هي أكبر قيمة في مجموعة البيانات والقيمة الدنيا هي أقل قيمة.

مثلاً: تم قياس الوقت الذي تستغرقه عملية ما عشر مرات وكانت البيانات كالتالي:

17، 13، 16، 14، 13، 15، 11، 18، 17 ثانية

في هذه الحالة تكون القيمة القصوى هي 18 والقيمة الدنيا هي 11 ثانية. وهذه المعلومة تجعلنا نتصور التغير الذي يحدث في وقت العملية. فهناك فارق بين أن يكون وقت العملية يتراوح بين 11 و 18 ثانية وأن يكون يتراوح بين 8 و 21 ثانية أو 13 و 16 ثانية.

المدى Range: المدى هو الفارق بين القيمة القصوى والقيمة الدنيا. في المثال السابق يكون المدى هو 7 ثوان. المدى يبيّن مقدار التغير الذي يحدث في البيانات.

منتصف المدى Midrange: هو المتوسط الحسابي للقيمة الدنيا والقصوى. ففي المثال السابق يكون منتصف المدى هو $(18+11)/2 = 14.5$ ثانية

ذكرت في المقالة السابقة أن المتوسط الحسابي يتأثر بوجود قيم متطرفة وأحب أن ألفت الانتباه هنا إلى أن منتصف المدى يتأثر بالقيم المتطرفة بشكل أكبر نتيجة لاعتماده على القيمة الدنيا والقيمة القصوى فقط. مشكلة هذا الأمر أن القيمة الدنيا أو القصوى ربما كانت مجرد قيمة غير حقيقة نتيجة لخطأ في القياس أو التسجيل أو نتيجة لظروف نادرة التكرار.

الانحراف المعياري Standard Deviation

الانحراف المعياري هو رقم شهير جداً حتى أن الكثير من البيانات يتم التعبير عنها عن طريق المتوسط الحسابي والانحراف المعياري. الانحراف المعياري يمكن وصفه بأنه متوسط بعد كل بيان أو كل قيمة عن المتوسط الحسابي. فمثلاً قد يتساوى المتوسط الحسابي لمجموعتين من البيانات ولكن الانحراف المعياري لأحد هما يكون أكبر من الآخر. هذا يعني أن المجموعة الثانية تبتعد فيها القيم عن المتوسط الحسابي أكثر من الأولى. على سبيل المثال: قد يتساوى متوسط درجة الحرارة لبلدين ولكن يكون الانحراف المعياري لإحداهما أكبر من الأخرى. هذا يعني أن درجة الحرارة في البلد ذات الانحراف الأكبر قد تبتعد أكثر عن الدرجة المتوسطة.

مثال: إذا كان متوسط نجاح الطلبة في مادة العلوم هو 77 درجة هذا العام و 77 درجة في العام الماضي ولكن الانحراف المعياري هذا العام هو 5 درجات بينما الانحراف المعياري هو 12 في العام الماضي. في أي الحالتين تكون معظم القيم قريبة من 77 درجة؟ الإجابة هي هذا العام لأن الانحراف المعياري أقل بكثير.

الانحراف المعياري يعبر عن التغير في القيم تعبيراً أدق من القيمة القصوى والدنيا لأنه يأخذ في الاعتبار بعد جميع القيم عن المتوسط الحسابي ولا تتوقف قيمته على قيمتين فقط.

الانحراف المعياري يرمز له بالرمز s إذا تم حسابه لجزء من المجتمع أي إذا كنا نستخدم عينة للتعبير عن المجتمع كله. كلمة مجتمع هنا تعني جميع قيم الشيء الذي ندرس. فلو أخذنا عينة عشوائية من درجات الحرارة خلال اليوم فإننا نتحدث عن عينة. في هذه الحالة فإن الانحراف المعياري يتم حسابه كالتالي:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}}$$

وهو عبارة عن الجذر التربيعي لمجموع (مربع الفارق بين كل قيمة والمتوسط الحسابي) مقسوماً على عدد القيم منقوصاً منها واحد.

أما عند وجود بيانات كاملة للمجتمع كله مثل جميع درجات الطلبة في الاختبار فإن الانحراف المعياري يحسب كالتالي:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{X})^2}{N}}$$

الفارق هو أننا نقسم هنا على عدد البيانات ولا ننقص منها واحداً. ويسمى الانحراف المعياري في هذه الحالة بـ سيجما σ وهي التي ينسب إليها مصطلح Six Sigma. بصفة عامة فالانحراف المعياري له استخدامات كثيرة قد نناقشها في مقالات لاحقة إن شاء الله.

لماذا نستخدم الجذر التربيعي والتربع؟ الفارق بين كل قيمة والمتوسط الحسابي مقسوماً على عدد القيم هو ما نريده. ولكن هذه القيم قد تلاشى بعضها البعض أي أن الفارق قد يكون موجباً أحياناً وسالباً أحياناً. لذلك نلجأ للتربع ليكون الفارق موجباً دائماً. ثم نلجاً للجذر التربيعي ليكون للانحراف المعياري نفس وحدات المتوسط الحسابي ونفس وحدات القيم نفسها.

في الواقع لا يهمني كثيراً أن أوضح تفاصيل الحسابات أكثر من ذلك لأنه من المتوقع أنك ستستخدم الحاسوب للقيام بذلك.

معامل الاختلاف Coefficient of Variation

عندما يكون لدينا حالتان متقاربتان في المتوسط الحسابي فإن المقارنة بين قيم الانحراف المعياري تبين لنا أي الحالتين أكثر تغيراً. ولكن ماذا إذا كان المتوسط الحسابي لإحدى الحالتين أكبر بكثير من المتوسط الحسابي للحالة الأخرى؟ في هذه الحالة فإن مقارنة الانحراف المعياري لا تكون معبرة عن مدى التغير. لذلك نستخدم معامل الاختلاف وهو مجرد نسبة الانحراف المعياري للمتوسط الحسابي.

مثال: إذا كان المتوسط الحسابي لمبيعات شركة ما هو 500 جنيه والانحراف المعياري هو 50 بينما المتوسط لشركة أخرى يساوي 200 والانحراف المعياري يساوي 30 فما هي الحالتين أكثر تغيرا؟

$$\text{معامل الاختلاف للحالة الأولى} = \frac{50}{500} = 0.1$$

$$\text{بينما معامل الاختلاف للحالة الثانية} = \frac{30}{200} = 0.15$$

كما ترى فإن معامل الاختلاف في الحالة الثانية أكثر من الأولى مما يعني أن التغير في القيم في الحالة الثانية أكبر من الأولى.

في المقالات التالية إن شاء الله نستكمم الحديث ونبين كيفية استخدام برنامج إكسيل للقيام بكل تلك الحسابات.

[مقالات ذات صلة بالموضوع:](#)

[تحليل البيانات](#)

[من مراجع الموضوع:](#)

Lean Six Sigma Pocket ToolBook, M. George et al., McGrawHill, 2005

Applied Statistical Methods, W. Carlson and B. Thorne, Prentice Hall, 1997

Statistics for Managers, Levine et al., Prentice Hall, 1999

[موقع ذات صلة بالموضوع:](#)

[الإحصاء](#)

[مقاييس الإحصاء الوصفي](#)

[الإحصاء الوصفي](#)

[Standard Deviation – Wikipedia](#)

تلخيص البيانات باستخدام برنامج إكسل

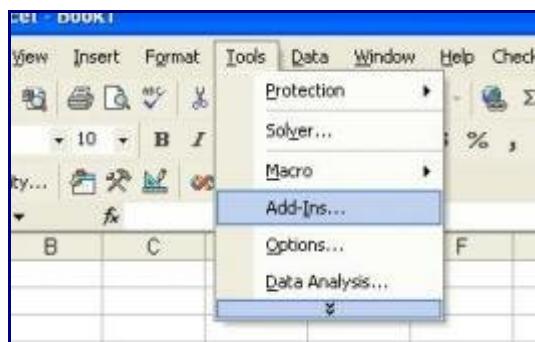
يوليو 6, 2009

ناقشت في المقالتين السابقتين بعض الأرقام التي تستخدم للتعبير عن مجموعة من البيانات مثل المتوسط الحسابي والوسيط والانحراف المعياري والمدى وغيرهم. في هذه المقالة أبين كيفية استخدام برنامج إكسل لحساب تلك الأرقام بشكل مباشر.

هناك أسلوبان لحساب هذه الأرقام باستخدام برنامج إكسل:

الأسلوب الأول: يمكنك الحصول على العديد من الأرقام التي تلخص مجموعة البيانات بخطوة واحدة. ولكن قبل القيام بذلك الخطوة قد تحتاج لتنصيب برنامج إكسل كالتالي:

اضغط على Tools ثم Add-Ins



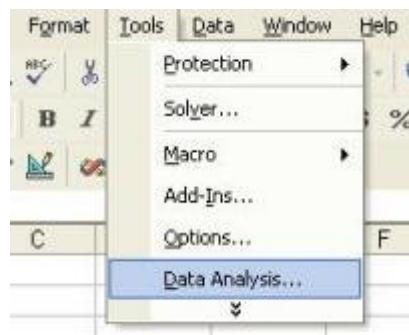
اختر OK ثم اضغط Analysis ToolPak



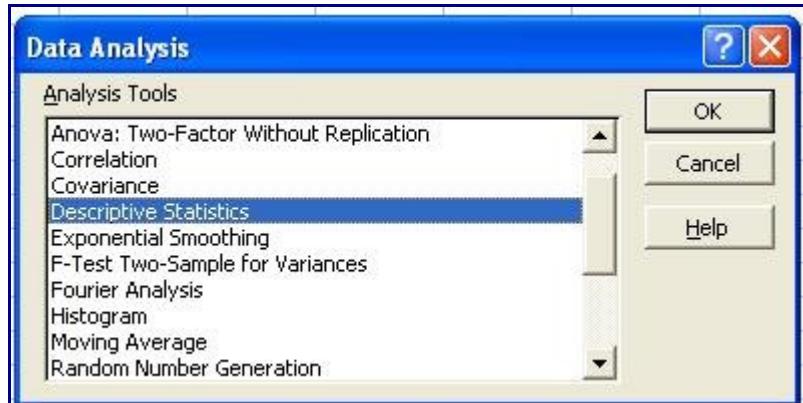
عندما نضغط على Tools الآن نجد أن إحدى الخيارات هو Data Analysis. افترض أن لدينا البيانات الآتية:

	A	B	C
1	112		
2	110		
3	100		
4	120		
5	90		
6	115		
7	97		
8	103		
9	112		
10	119		
11	100		
12	109		
13	98		
14	114		
15	107		
16	113		
17	112		
18	115		
19	119		
20	108		
21			
22			

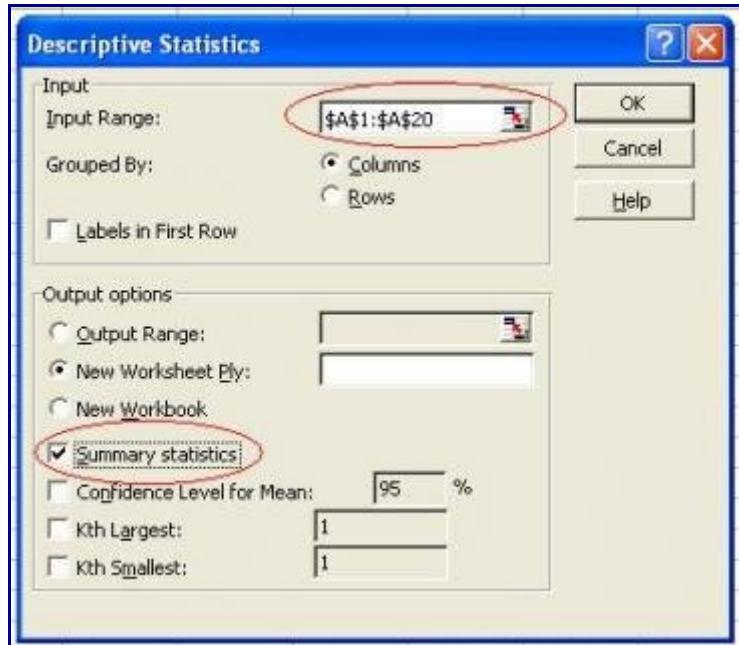
لكي نحل هذه البيانات اضغط على tools ثم اختر Data Analysis



تظهر لك نافذة بها العديد من الخيارات. اختر Descriptive Statistics أي الإحصاء الوصفية



تظهر لك النافذة التالية وفيها يجب تحديد الخلايا التي بها البيانات وهي في حالتنا هذه A1:A20. يمكنك أن تختار باستخدام الفأرة بالطرق المعروفة. اختر كذلك Summary statistics كما بالشكل



تظهر لك البيانات التالية في صفحة منفصلة وقد وضعنا أنا الترجمة العربية لكل رقم بالشكل أدناه.

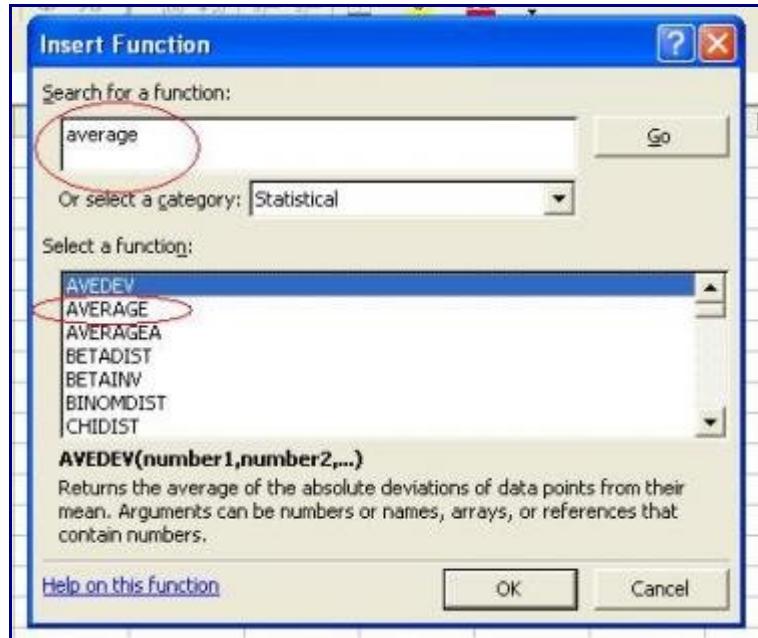
Column1		
Mean	108.65	المتوسط الحسابي
Standard Error	1.841445371	
Median	111	الوسط
Mode	112	المنوال
Standard Deviation	8.235194051	الانحراف المعياري
Sample Variance	67.81842105	انحراف البينة
Kurtosis	-0.258992603	
Skewness	-0.650571846	
Range	30	المدى
Minimum	90	القيمة الدنيا
Maximum	120	القيمة القصوى
Sum	2173	المجموع
Count	20	عدد البيانات

كل هذه البيانات قد ناقشنا معانا في المقالتين السابقتين فيما عدا Kurtosis و Standard Errors و Skewness وهي أرقام ربما نناقشها في المستقبل.

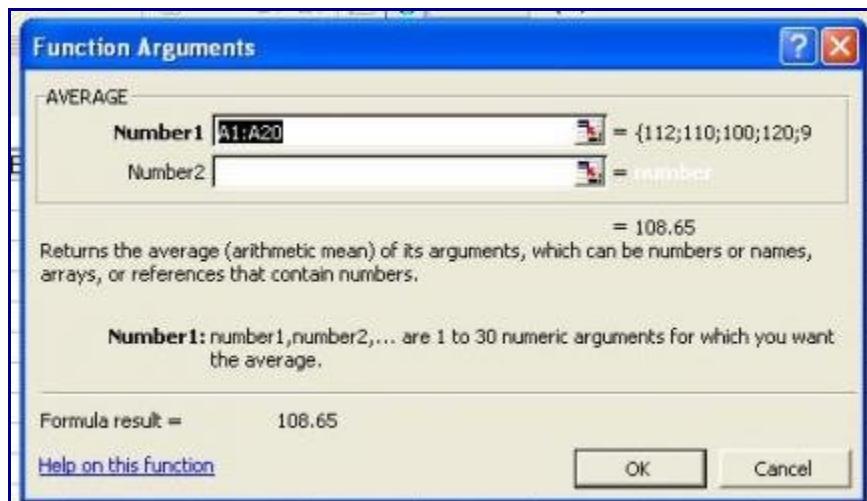
الأسلوب الثاني:

قد تكون لديك الرغبة في استخدام رقم واحد أو عدة أرقام من هذه الأرقام. في هذه الحالة قد تفضل استخدام الأسلوب الثاني.

يمكنا حساب كل رقم من تلك الأرقام باستخدام الدوال المتاحة في إكسيل. هناك عدة طرق للبحث منها كتابة اسم الدالة العلمي في الخانة الأولى. والطريقة الثانية هي اختيار نوع معين من الدوال مثل اختيار الدوال الإحصائية Statistics والبحث فيها.



فمثلاً لحساب المتوسط الحسابي فإننا نختار الدالة Average ثم في النافذة التالية نحدد المدى الذي يحتوي البيانات



فيظهر لنا المتوسط الحسابي في الخلية التي كتبنا فيها هذه الدالة وقيمة في هذا المثال 108.65 وكذلك لكل رقم من تلك الأرقام دالة لحسابه. هذه الدوال موضحة في الشكل أدناه.

=AVERAGE(A1:A20)	Average	المتوسط الحسابي
=MEDIAN(A1:A20)	Median	الموسيط
=QUARTILE(A1:A20,1)	Quartile	الربع الأدنى
=MODE(A1:A20)	Mode	المنوال
=MAX(A1:A20)	Max	المقدمة الضخوى
=MIN(A1:A20)	Min	القيمة الدنيا
=STDEV(A1:A20)	Standard Deviation (Sample)	الانحراف المعياري للعينة
=STDEVP(A1:A20)	Standard Deviation (Population)	الانحراف المعياري للمجتمع

فكتابة أي دالة في أي خلية تحصل على الرقم الذي تريده.

بهذا نستطيع استخدام تلك الأرقام بكل سهولة. في المقالات التالية إن شاء الله نتعرض لمواضيع أكثر تقدما في تحليل البيانات.

دراسة العلاقة بين متغيرين

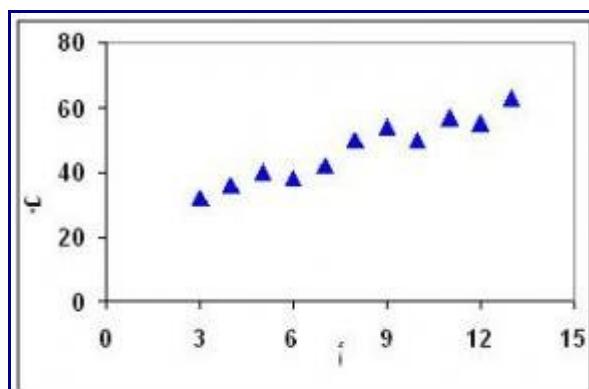
يوليو 25, 2009

من المهارات التي يحتاجها المديرون والمهندسو الصناعيون والكثير من الناس مهارة تحديد العلاقة بين متغيرين أو أكثر. هل هناك علاقة بين حجم المبيعات في كل منطقة وأسلوب التسويق المتبعة؟ هل هناك علاقة بين عدد المشاكل في المعدات ودرجة حرارة الجو؟ هل هناك علاقة بين مرض كذا وجودة مياه الشرب؟ هل هناك علاقة بين رضا الموظفين عن عملهم وجودة الوجبة التي تقدم لهم؟ هذه النوعية من الأسئلة تتعرض لها كثيراً في العمل ونحاول البحث عن إجابة مبنية على أساس صحيح. في هذه المقالة والمقالات التالية -إن شاء الله- أحاول استعراض الأدوات البسيطة التي تمكنا من الإجابة على مثل هذه الأسئلة.

1- الرسم البياني :Scatter Diagram

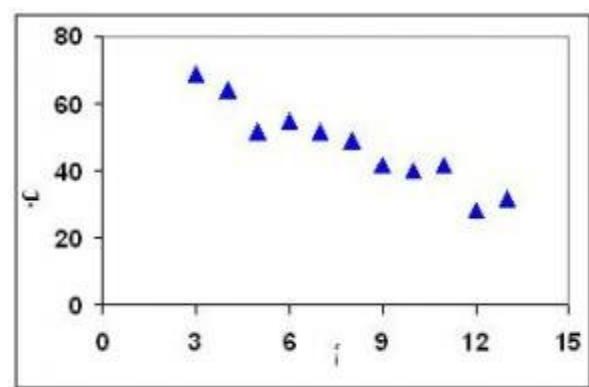
يمكننا أن نرسم العلاقة بين قيم المتغيرين لنرى إن كان الرسم يشير لوجود علاقة أم أنه يبين عدم وجود أي علاقة بين المتغيرين. يسمى هذا المنحنى بـ Scatter Plot أو Scatter Diagram وقد وجدت له عدة ترجمات باللغة العربية مثل: رسم بياني تلقائي، رسم (شكل) الانتشار، رسم بياني مبعثر ورسم (مخطط) التشتت. الرسومات الآتية توضح أمثلة للعلاقة بين متغيرين.

علاقة طردية (إيجابية)



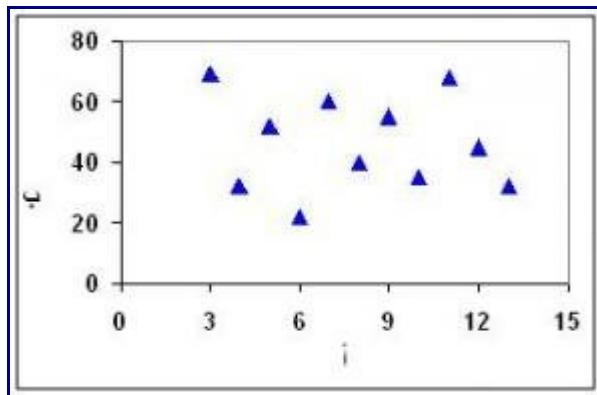
هذا الرسم يظهر أن هناك علاقة طردية بين المتغير ب والمتغير أ فكلما ازداد أ ازداد ب. وإن كنا لا نستطيع التوصيل بخط مستقيم واحد بين كل النقاط ولكننا بمجرد النظر ندرك أن كل النقاط لها اتجاه واحد هو اتجاه العلاقة الطردية.

علاقة عكسية (سلبية)



هذا الرسم يظهر علاقة عكسية فعندما كان A يساوي 3 كان B يساوي 70 وعندما زاد A إلى 9 كان B حوالي 40 وعندما ازداد A إلى 12 كان B حوالي 30. فكلما ازداد A أقل B .

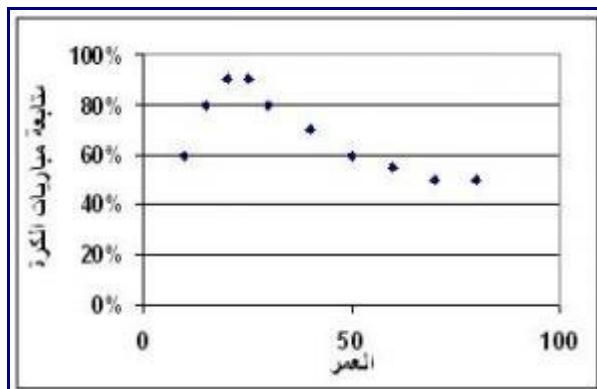
لا علاقة



هذا الرسم يوضح انه لا توجد أي علاقة بين تغير A وتغير B .

علاقة غير خطية

قد يظهر الرسم علاقة غير خطية وهذا من مميزات هذا الرسم. العلاقة الخطية هي العلاقة التي يمكن التعبير عنها بخط مستقيم مثل العلاقة الطردية أو العكسية السابق بيانهما. أما العلاقة غير الخطية فإنها تأخذ شكل منحنى. فمثلاً في الرسم أدناه نلاحظ العلاقة بين العمر ونسبة متابعة مباريات كرة القدم (هذه بيانات افتراضية). فهذا الرسم يوضح علاقة غير خطية ويمكن تحليلها وفهمها فمتباينة المباريات تزداد مع زيادة العمر ما بين 10 و 20 عاماً ثم تقل تدريجياً بداية من عمر 30 عاماً. ويمكن فهم ذلك بأن الأطفال والشباب يبدؤون في حب الكرة تدريجياً فبعضهم لا يفهمها حتى سن الخامسة عشرة. والرجال من سن 30 سنة يبدؤون في الانشغال في أعمالهم فتقل نسبة متابعتهم للمباريات تدريجياً.



لاحظ أن تحليل هذه العلاقات لابد أن ينبع من دراسة وتحليل منطقي يعتمد على فهم الموضوع وتجميع بيانات داعمة لهذا التحليل.

علاقة وليس سببية:

إن وجدت علاقة طردية أو عكسية عن طريق الرسم البياني أو أي وسيلة أخرى فلا يمكنك أن تدعى أن A هو سبب B بل ما تستطيع قوله أن هناك علاقة بين المتغيرين. قد يكون A وب يزيدان بسبب متغير آخر والذي هو سبب A وب. فمثلاً عند ارتفاع درجة حرارة الجو فإن درجة حرارة أي غرفة بالمنزل ستترتفع وإن كانت بعض الغرف قد

تكون درجة حرارتها مختلفة عن الأخرى حسب وضع الغرف من ناحية مواجهة الشمس والتهوية. فلو درسنا العلاقة بين درجة حرارة غرفة النوم ودرجة حرارة غرفة المعيشة فإن سبب علاقة طردية واضح. وهذا أمر طبيعي ولكن لا نستطيع أن نقول أن درجة حرارة غرفة النوم ترتفع نتيجة لارتفاع درجة حرارة غرفة المعيشة ولكن الحقيقة أن كلاهما يزيد كنتيجة لارتفاع درجة حرارة الجوخارجي.

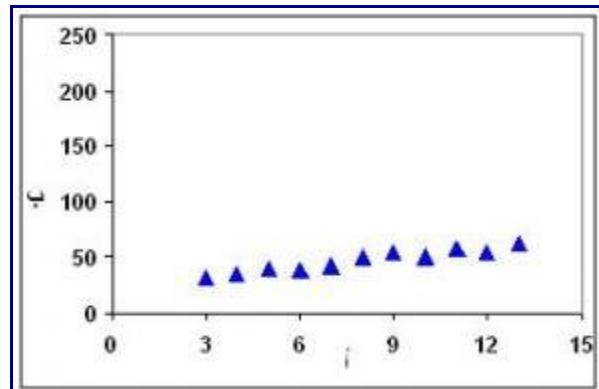
لا ننس أبداً أن كل وسائل تحديد العلاقة بين متغيرين أو أكثر هي وسائل لتحديد وجود علاقة وليس لتحديد وجود سببية.

الأمانة في عرض البيانات:

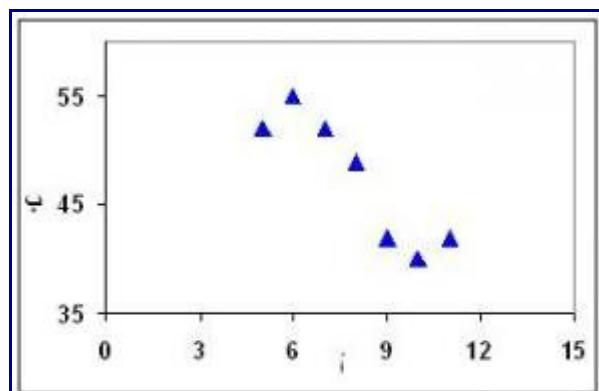
لابد من أن تكون أميناً في عرض البيانات فمثلاً:

- لا تحاول تغيير مقياس الرسم لظهور علاقة غير حقيقية أو تخفي علاقة حقيقية. فعندما تقوم بتكبير مقياس الرسم فقد لا يستطيع القارئ رؤية أي علاقة. وبالعكس فإن علاقة غير حقيقية يمكن أن تخدع القارئ بها إذا قمت بالتركيز على جزء صغير من المحور الرأسي.

الرسم التالي هو لنفس العلاقة الطردية المذكورة في المثال السابق ولكن بعد تكبير محور ب لكي يغطي من صفر إلى 250. إنك لا تكاد ترى العلاقة الطردية الآن.



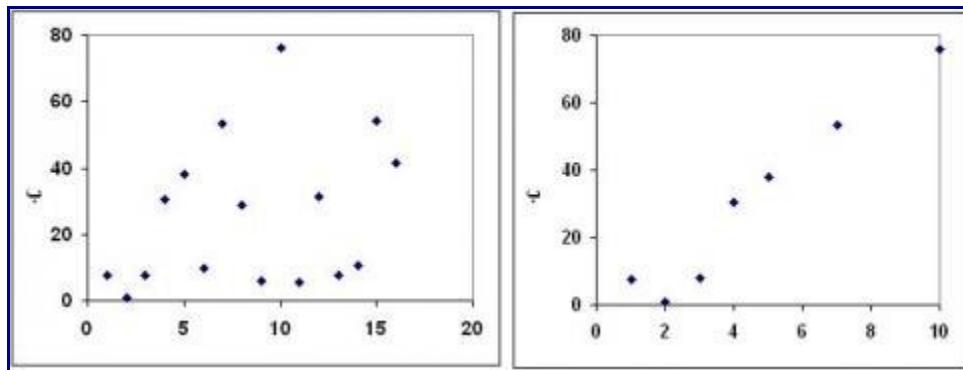
والرسم التالي هو لنفس العلاقة العكسية الواضحة المعروضة في المثال أعلاه ولكن بتحديد المحور ب بين 25 و 60 فقط. من الصعب تحديد علاقة من هذا الرسم.



- لا تحاول إخفاء بعض البيانات لكي تقنع القارئ بما تريده. عليك أن تعرض البيانات كما هي. لا تتنقق النقاط التي تظهر علاقة ما أو تخفي علاقة ما ولكن اعرض كل البيانات بأمانة.

الرسم التالي - على اليسار - يوضح العلاقة الحقيقية بين متغيرين حيث لا توجد أي علاقة. بينما الرسم على اليمين

يظهر علاقة طردية قوية وذلك بعد تحديد المحور الأفقي بحد أقصى عشرة وكذلك إخفاء بعض النقاط.

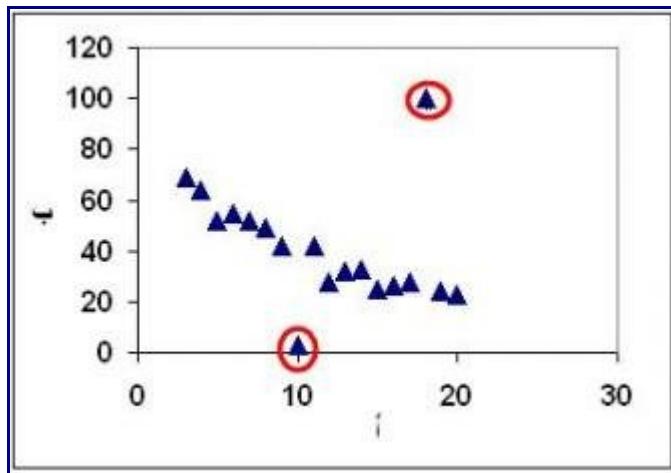


- بالطبع لا تغير البيانات بل اعرضها كما هي فلا تغير أي رقم لكي يبدو الشكل كما نتمنى.

القيم المتطرفة Outliers

من فوائد هذا الرسم (المخطط) أننا نكتشف القيم المتطرفة بمجرد النظر. ففي كثير من الحالات تجد أن البيانات كلها مبعثرة ولكن في منطقة محدودة وهناك قيمة أو أكثر بعيدة كل البعد عن باقي القيم. هذه القيمة أو القيم تسمى نقاط متطرفة أو بيانات متطرفة بمعنى أنها بيانات خارجة عن مجموعة البيانات. هذه القيم المتطرفة قد تكون بسبب:

- 1- خطأ في تسجيل أو تجميع البيانات
- 2- وجود حالات نادرة لها أسبابها في البيانات فمثلا عند رسم العلاقة بين الدخل والسن فقد تجد معظم البيانات تقع في مدى محدود وهناك قيم قليلة متطرفة للأثراء بشكل كبير. وهذه قيم متطرفة ولكنها حقيقة.
- 3- وجود مجموعتين من البيانات مما يعني الحاجة لدراسة كل منها على حدة فمثلا عندما تدرس علاقة وقت تصنيع جزء ما بجودة المواد الخام فقد تجد قيمًا متطرفة إن كان هناك نوعين أو أكثر من المنتجات وفي هذه الحالة سيكون من الأفضل دراسة كل منها بشكل منفصل.



القيم المتطرفة تحتاج دراسة لأنها قد تدلنا على وجود مشكلة في تجميع البيانات فنقوم بتصحيحها أو تدلنا على وجود حالات نادرة فنحاول دراسة سببها أو تدلنا على وجود مجموعات مختلفة فنقوم بدراستها. على أي حال فإننا أحيانا قد نهمل القيم المتطرفة إذا كانت فيما خاطئة أو خارجة عن نطاق دراستنا ولكن هذا لا يعني إهمال أي قيم لمجرد عدم رغبتنا في عرضها. ولذلك فمن الأفضل توضيح هذه القيم التي استبعدت وسبب استبعادها.

في المقالات التالية إن شاء الله نستعرض وسائل تحديد قوة العلاقة بين متغيرين وكذلك تحديد معادلة رياضية تصف

هذه العلاقة.

مقالات ذات صلة:

تحليل البيانات

من مراجع الموضوع:

Lean Six Sigma Pocket ToolBook, M. George et al., McGrawHill, 2005

Applied Statistical Methods, W. Carlson and B. Thorne, Prentice Hall, 1997

Statistics for Managers, Levine et al., Prentice Hall, 1999

موقع ذات صلة:

[Scatter Plot](#)

[Scatter Plot](#)

[Scatter Diagram](#)

معامل الارتباط Correlation

اغسطس 2, 2009 4 عدد التقييمات

ناقشت في المقالة السابقة كيفية دراسة العلاقة بين متغيرين باستخدام الرسم البياني (المنحنى التقاطعي) Scatter Diagram والذي يساعدنا على اكتشاف أي علاقة خطية أو غير خطية بين المتغيرين أو اكتشاف عدم وجود أي علاقة. وفي هذه المقالة نتعرض لطرق أخرى لدراسة العلاقة بين متغيرين. هذا الموضوع هو من الأساسيات التي يحتاجها المدير وأي شخص يحتاج لتحليل بيانات ودراسة علاقتها ببعضها.

معامل الارتباط :Correlation

معامل الارتباط هو رقم يتراوح بين -1 و 1 وهو يبين وجود علاقة خطية بين متغيرين واتجاه تلك العلاقة كما يلي:

- +1 تعني علاقة طردية بمعنى أنه كلما زاد أ زاد ب وكلما قل أ فإن ب يقل
- 1 تعني علاقة عكسية بمعنى أنه كلما زاد أ فإن ب يقل وكلما قل أ فإن ب يزيد
- صفر يعني عدم وجود أي علاقة بين المتغيرين

عندما يقترب معامل الارتباط من إحدى هذه القيم فإنه يدل على ما تدل عليه هذه القيم ولكن بدرجة أقل. فمثلاً 0.9+ تدل على وجود علاقة طردية قوية بين المتغيرين ولكنها ليست مطلقة مثل تلك التي تتوقعها عندما يكون معامل الارتباط يساوي +1.

يسمي معامل الارتباط بمعامل الارتباط لبيرسون Pearson Correlation Coefficient ويُشيع تسميته بمعامل الارتباط . ولمعامل الارتباط تطبيقات عديدة فمثلاً في مجال التسويق قد تحب أن تدرس إن كان هناك علاقة بين زيادة مبيعات منتجك وزيادة مبيعات سلعة أخرى أو تحسن درجة الحرارة أو تخفيض السعر. وقد تكون مهندساً ي يريد أن يعرف ما الذي يؤثر على جودة الغاز المنتج هل هو تغيير الضغط أم الحرارة أم جودة أي غاز من الغازات الداخلة في العملية الإنتاجية.

طريقة الحساب :

معامل الارتباط يتم حسابه بسهولة عن طريق الحاسوب ولذلك فلسنا بحاجة للدخول في حسابات مملة ولكن من الضروري أن نلقي نظرة على طريقة الحساب لنفهم معنى معامل الارتباط. يتم حساب معامل الارتباط كالتالي

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_x S_y}$$

والبسط في هذه المعادلة هو مجموع حاصل ضرب الفارق بين كل قيمة للمتغير الأول ومتوسطه الحسابي في الفارق بين كل قيمة للمتغير الثاني ومتوسطه الحسابي. والمقام هو حاصل ضرب الاتحراف المعياري لكل من المتغيرين في عدد البيانات منقوصاً منها واحد. هذا في حال أن لدينا عينة من البيانات وأن نأخذ عينة عشوائية من مجموعة كبيرة (المجتمع) وندرس ظاهرة معينة على هذه العينة. أما عند دراسة المجتمع كله فإن طريقة الحساب تختلف اختلافاً طفيفاً وتكون كالتالي

$$\rho = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n) \sigma_x \sigma_y}$$

في هذه الحالة فإن المقام يكون حاصل ضرب الانحراف المعياري للمجتمع لكل من المتغيرين مضرباً في عدد البيانات.

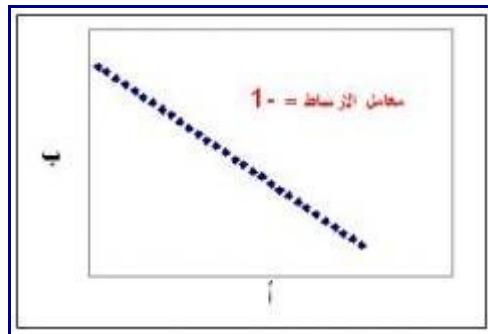
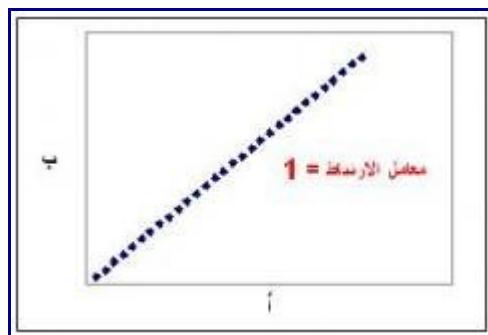
ماذا نفهم من هذه المعادلة المعقّدة؟

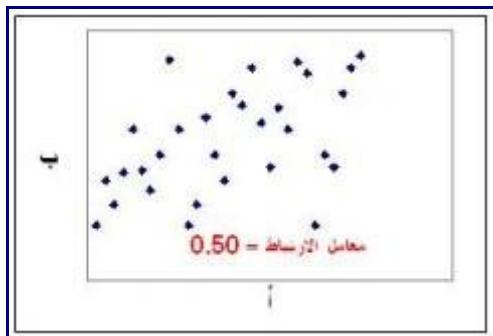
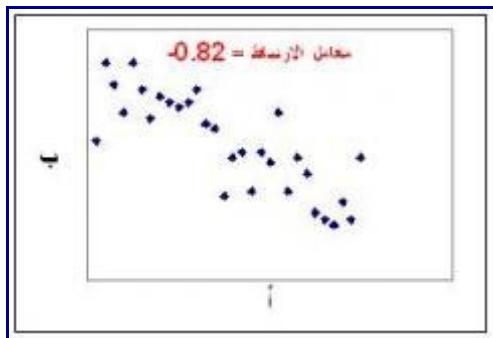
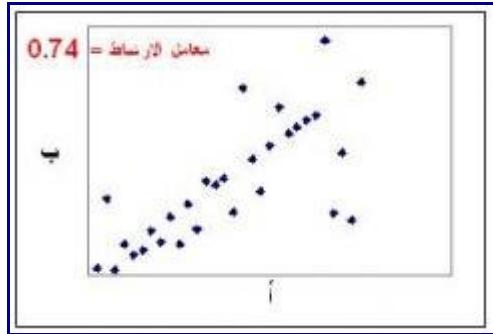
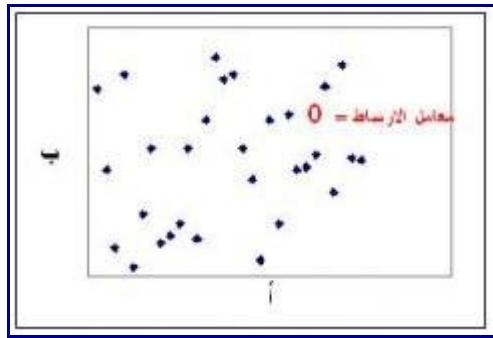
أولاً المقام هو حاصل ضرب أرقام موجبة (أكبر من الصفر) فالانحراف المعياري هو دائماً رقماً موجباً وكذلك عدد البيانات. فمتى يكون معامل الارتباط موجباً ومتى يكون سالباً؟ الأمر يتوقف على البسط. فإذا كان الفارق بين قيمة ما للمتغير الأول ومتوسطه الحسابي موجباً وكان الفارق بين القيمة المقابلة والمتوسط الحسابي للمتغير الثاني موجباً كانت النتيجة موجبة لأن حاصل ضرب قيمة موجبة في قيمة موجبة يساوي قيمة موجبة. وإذا كان كل منها سالباً فإن الناتج يكون موجياً لأن حاصل ضرب قيمة سالبة في قيمة سالبة يساوي قيمة موجبة. ومعنى ذلك (في الحالة الأولى) أنه عند زيادة المتغير الأول عن متوسطه الحسابي فإن المتغير الثاني يزيد عن متوسطه الحسابي هو الآخر وكذلك (في الحالة الثانية) عند نقصان المتغير الأول عن متوسطه الحسابي فإن نفس الأمر يحدث للمتغير الثاني.

وبالتالي فإنه عندما تكون العلاقة عكسية فإن الناتج يكون سالباً لأن أحد الفارقين سيكون موجباً والآخر سالباً. وهذا يجعلنا نفهم القاعدة بأن معامل الارتباط كلما كان أقرب للواحد الصحيح فإن ذلك يعني وجود علاقة طردية قوية وكلما اقترب من -1 فإن ذلك يعني وجود علاقة عكسية قوية. وكلما اقترب من الصفر فإن ذلك يعني عدم وجود علاقة خطية.

شكل العلاقة:

لنتنظر إلى بعض الرسومات البيانية المرادفة لقيم مختلفة لمعامل الارتباط لنفهم ما يعنيه هذا الرقم.





كيف نستخدم إكسل لحساب معامل الارتباط:

هناك طريقتان يمكننا استخدامهما.

افتراض أن لدينا البيانات التالية:

<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>
22	3	25	1
40	37	39	2
30	3	35	3
43	15	30	4
60	10	39	5
44	13	27	6
36	22	29	7
50	16	33	8
88	28	32	9
60	15	31	10
22	34	32	11
30	22	34	12
65	46	28	13
50	43	27	14
40	47	15	15

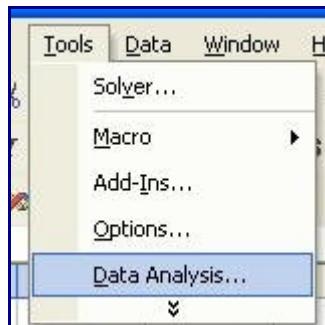
ونريد حساب معامل الارتباط بين المتغير *a* و *b*. **الطريقة الأولى** هي أن نستخدم الدالة المتاحة في إكسل لحساب معامل الارتباط فنكتب ما يلي في أي خلية:

=CORREL(D2:D16,C2:C16)

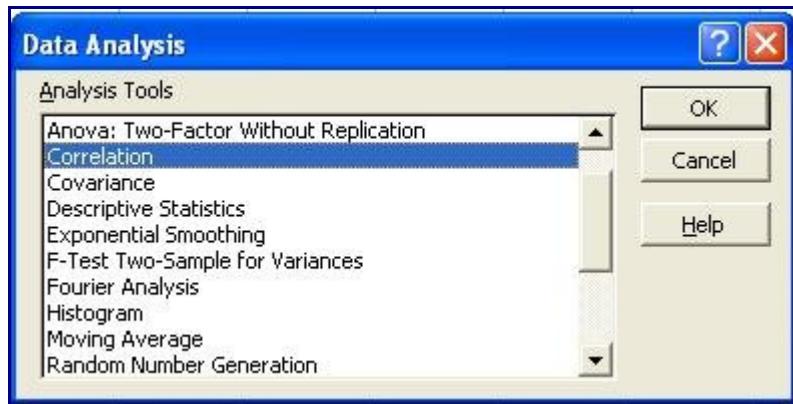
وبالتالي نحصل على معامل الارتباط بين *a* و *b* وهو -0.46 . هذه القيمة تعني وجود علاقة عكسية ضعيفة لأن القيمة لا تقترب من -1 بل هي أقرب قليلا إلى الصفر.

الطريقة الثانية تساعدنا في الحصول على معامل الارتباط بين متغيرين أو عدة متغيرات مرة واحدة. هذه الطريقة تتم كالتالي:

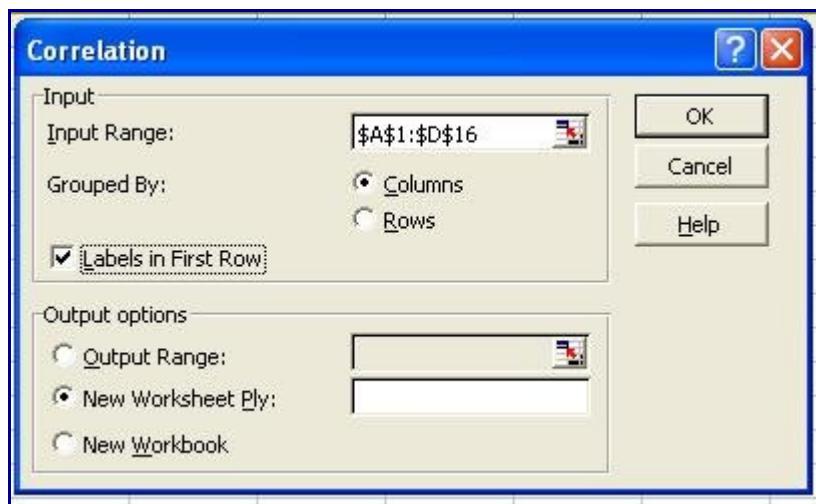
اضغط على Tools ثم Data Analysis (أو أوضحت من قبل كيفية اظهار Data Analysis



ثم اختر Correlation



تظهر لك النافذة التالية وعليك ملء Input Range بأسماء الخلايا التي مسجل بها البيانات. وقد علمت على أي أن أسماء الأعمدة في الصف الأول (أي أ ب و ت و ث) Labels in First Row



نضغط OK فنحصل على النتيجة كالتالي:

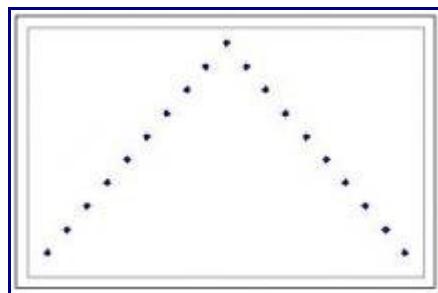
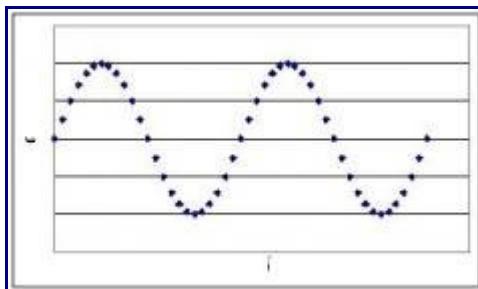
	أ	ب	ث	ت
أ	1			
ب	0.21	1		
ث	0.12	-0.35	1	
ت	0.25	0.72	-0.46	1

هذا الجدول (بالأعلى) يبين أن معامل الارتباط بين ث و ت مثلا هي 0.21 ومعامل الارتباط بين ب و ت هي -0.35 وهكذا. بالطبع فإن العلاقة بين المتغير ونفسه هي 1 فترى في الجدول معامل ارتباط بـ ث هو 1 وهي قيمة لا تعنينا في شيء. هذه الطريقة سريعة جدا عندما يكون لدينا أكثر من متغيرين. من هذه النتيجة نرى أن العلاقة الخطية بين أ و ت هي الوحيدة التي يمكن أخذها في الاعتبار لأنها تساوي 0.72 أما باقي القيم فهي صغيرة جدا.

هل لا توجد علاقة؟

ليس معنى أن يكون معامل الارتباط صفرًا أو قريباً من الصفر أنه لا توجد أي علاقة بين المتغيرين. فمعامل الارتباط يبين قوة العلاقة الخطية. والعلاقة الخطية هي علاقة في شكل خط مستقيم فهي علاقة ليس بها منحنيات أو

طلع ونزو. فالعلاقة الخطية تكون طردية أو عكسية فقط. وبالتالي فقد يكون معامل الارتباط يساوي صفرًا ولكن توجد علاقة قوية بين المتغيرين ولكنها غير خطية أي أنها ليست على شكل خط مستقيم كما في الأمثلة التالية:



في هذين الشكلين نرى علاقة واضحة بين المتغيرين ولكنها ليست مجرد علاقة طردية أو عكسية ولا يمكن تمثيلها بخط مستقيم. ففي الحالة الأولى نلاحظ تغير المتغير الثاني بشكل دوري مع المتغير الأول. وفي الحالة الثانية نجد علاقة طردية حتى نقطة ما ثم تحول العلاقة إلى علاقة عكسية. هذه العلاقات هي علاقات غير خطية ولا يمكن التنبؤ بها بمعامل الارتباط.

بهذا نكون قد استطعنا دراسة شكل العلاقة عن طريق منحنى الانتشار (المنحنى التتفطي) ومعرفة قوة العلاقة الخطية عن طريق معامل الارتباط. في المقالة التالية إن شاء الله نناقش كيفية الوصول لعلاقة رياضية بين متغير وكل المتغيرات التي تؤثر فيه.

[مقالات ذات صلة:](#)

[تحليل البيانات](#)

[من مراجع الموضوع](#)

Lean Six Sigma Pocket ToolBook, M. George et al., McGrawHill, 2005
Applied Statistical Methods, W. Carlson and B. Thorne, Prentice Hall, 1997

[موقع ذات صلة:](#)

[الارتباط والانحدار الخطى](#)

[What is a Correlation](#)

[Pearson's Correlation](#)

تحليل الانحدار ... Regression Analysis....

اغسطس 13, 2009

ناقشت في المقالتين السابقتين كيفية دراسة العلاقة بين متغيرين باستخدام الرسم البياني (المنحنى التلقائي) Scatter Diagram وكيفية دراسة قوة العلاقة الخطية بين متغيرين باستخدام معامل الارتباط Correlation. في هذه المقالة نناقش كيفية تحديد العلاقة بين متغيرين أو أكثر بشكل أكثر تحديداً وبحيث نستطيع تكوين نموذجاً رياضياً لتلك العلاقة. هذا الموضوع هو من الأدوات الرائعة لتحليل البيانات والتي يحتاجها الكثير من الناس.

انحدار :Regression

تحليل الانحدار Regression Analysis هو تحليل يمكننا من إيجاد معادلة رياضية تربط بين متغير تابع ومتغير أو متغيرات مستقلة. فمثلاً يمكننا باستخدام تحليل الانحدار دراسة العوامل التي تؤثر في زيادة الطلب على المنتج وتحديد نموذجاً (معادلة) رياضياً لهذه العلاقة. هذا النموذج يجعلنا قادرين ليس فقط على فهم طبيعة العلاقة وتحديد العوامل المؤثرة فعلاً بل إنه يجعلنا قادرين على توقع تأثير تغيير أي متغير من هذه المتغيرات المستقلة على المتغير التابع.

الحاجة لاستخدام هذا الانحدار كثيرة ومتنوعة. فالمهندس يحتاج لدراسة العوامل التي تؤثر في ارتفاع درجة حرارة الغازات المستخدمة في عملية ما وقد يكون لديه العديد من العوامل التي يريد أن يعرف تأثيرها الحقيقي. باستخدام الانحدار فإن هذا المهندس يستطيع تحديد العوامل المؤثرة وإهمال تلك غير المؤثرة ويمكنه توقع التغير الذي يحدث في درجة حرارة الغازات نتيجة لتغير محدد في أي من تلك المتغيرات المؤثرة. ومدير الموارد البشرية يريد تحديد العوامل التي تؤثر على أداء العاملين الجدد من بين عدة عوامل مثل السن وتقدير التخرج وجامعة الدراسة وغيرها. فيمكنه باستخدام تحليل الانحدار معرفة ما هي العوامل التي لا تؤثر ولا ترتبط بأداء العاملين الجدد وتلك المؤثرة ويمكنه الحصول على نموذجاً رياضياً يمكنه من توقع وفهم حجم تأثير تلك العوامل على الأداء.

ما هو الحل البديل لتحليل الانحدار؟ إنه محاولة تغيير أحد العوامل مع ثبات العوامل الأخرى ثم إجراء ذلك مع كل عامل من العوامل الأخرى وهذا غير متاح في الواقع العملي. فلا يمكنك أن تطلب من درجة حرارة الجو أن تثبت حتى تدرس تأثير تأثير نسبة الألترة في الجو على صحة البشر. وكذلك لا يمكنك أن تقوم بثبات سرعة الماكينة في العمل لمدة أسبوع لتدرس تأثير نوع الزيت المستخدم بشكل مستقل عن تغير السرعة. تحليل الانحدار يجعلنا لا نلجأ لهذه الطرق شبه المستحيلة فبمجرد وجود عينة من البيانات للمتغيرات المختلفة يمكننا تحديد العوامل المؤثرة وطبيعة تأثيرها بشكل محدد واضح.

أنواع تحليل الانحدار: هناك نوعان من تحليل الانحدار أولهما هو الانحدار الخطى وهو الأكثر انتشاراً. الانحدار الخطى يعني أننا ندرس العلاقة الخطية. أما النوع الثانى فهو الانحدار غير الخطى والذي نحتاجه عند دراسة علاقات على شكل منحنى وليس خطًا مستقيماً. الانحدار الخطى هو الأكثر شيوعاً وهو الذي نناقشه هنا. والانحدار الخطى له نوعان بسيط ومتعدد فالبسيط يحاول التنبؤ بالعلاقة بين متغير ما وعامل واحد يؤثر فيه والمتعدد يحاول التنبؤ بالعلاقة بين متغير ما وعدة عوامل تؤثر فيه. في هذه المقالة نناقش النوع الأول وهو الانحدار الخطى البسيط.

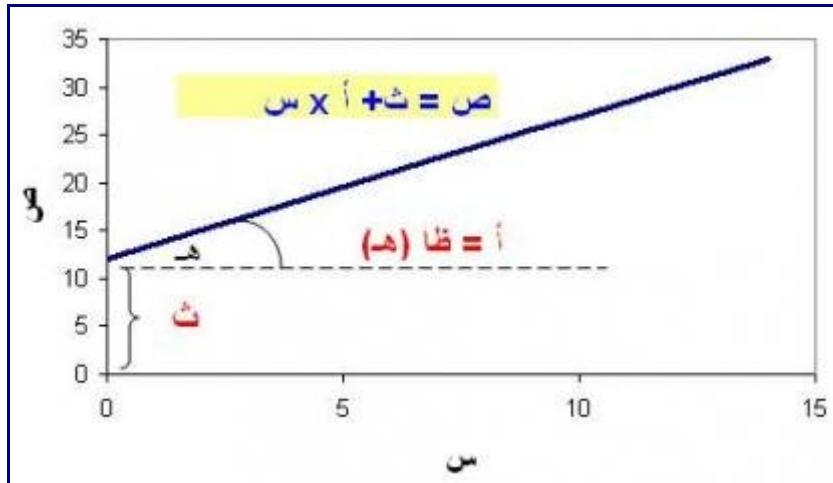
انحدار الخطى البسيط Simple Linear Regression

الانحدار الخطى البسيط هو دراسة العلاقة بين متغيرين فقط بحيث نحاول الوصول إلى علاقة خطية (أي معادلة خط مستقيم) بين هذن المتغيرين في صورة:

$$ص = \theta + \alpha x$$

حيث $ص$ و x هما متغيران و θ و α هما ثابتان. هذه المعادلة هي المعادلة التي ترسم خطًا مستقى بين $ص$ و x .

لاحظ أن s هنا يسمى متغيراً تابعاً أي أن تغييره يتبع تغيير s وأما t فليس متغيراً مستقلاً أي أن تغييره هو تغيير مستقل وهذه تسميات رياضية فقط.



فمثلاً نريد أن نعرف طبيعة العلاقة بين متوسط ساعات الدراسة ودرجة الطلبة في الامتحانات أو نريد دراسة العلاقة بين سعر المنتج وحجم المبيعات أو نريد دراسة العلاقة بين عدد المنتجات المعبأة ومعدل التحميل أو نريد دراسة العلاقة بين استهلاك الكهرباء في الساعة ودرجة حرارة الجو.

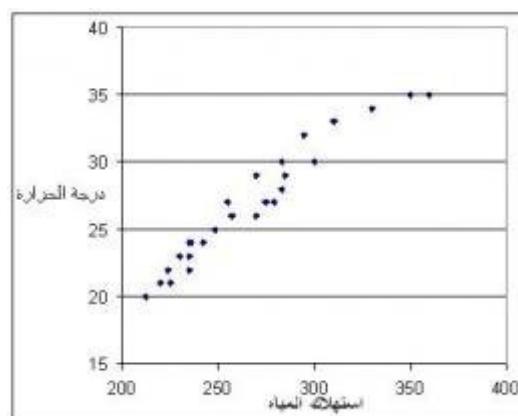
مثال: قبل أن نستفيض في شرح تحليل الانحدار دعنا نرى كيف يمكننا إجراء مثل هذا التحليل.

افترض أننا سجلنا درجة الحرارة المتوسطة في كل يوم واستهلاك المدينة للمياه في تلك الأيام فكانت البيانات كالتالي:

استهلاك المياه	درجة الحرارة
212	20
220	21
230	23
235	24
255	27
270	29
283	30
295	32
310	33
330	34
360	35
350	35
225	21
235	22
235	23
242	24
248	25
257	26
279	27
283	28
285	29
300	30
224	22
236	24
248	25
270	26
275	27

بإمكاننا أن نلاحظ ارتفاعاً في استهلاك المياه عند ارتفاع درجة الحرارة وهذا أمر متوقع. ولكننا نريد أن نتأكد من ذلك إحصائياً وأن نصل إلى نموذجاً رياضياً يمكننا من توقع حجم الاستهلاك عند أي درجة حرارة فمثلاً قد نتساءل ما هو حجم الاستهلاك لو وصلت درجة الحرارة إلى 39 درجة مئوية.

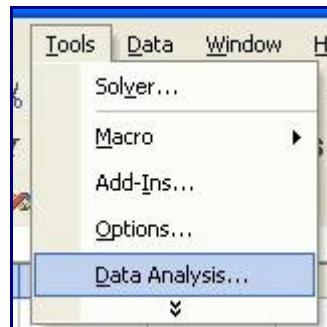
برسم العلاقة بين المتغيرين قد تبين لنا هذه العلاقة والتي يمكننا أن نتأكد من قوتها باستخدام معامل الارتباط.



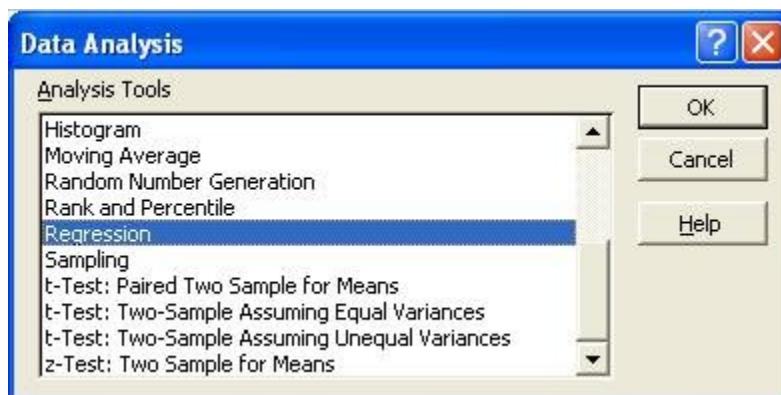
$$\text{معامل الارتباط} = 0.97$$

كل هذا يشير إلى علاقة طردية قوية. ولكن حتى الآن ليس لدينا معادلة رياضية تبين لنا حجم الاستهلاك عند درجة حرارة 20 أو درجة حرارة 40 درجة مئوية. نحن بحاجة لإجراء تحليل الانحدار.

نستخدم في ذلك برنامج إكسيل. اختر قائمة الأدوات Tools ثم اختر تحليل البيانات Data Analysis



بعد ذلك يظهر لنا نافذة بها عدة اختيارات فنختار منها الانحدار Regression



تظهر لنا النافذة الآتية والتي نحدد فيها مكان تسجيل بيانات المتغير التابع (استهلاك المياه) والمتغير المستقل (درجة الحرارة) وقد نضع علامة بجوار Labels إذا كنا قد كتبنا عنوانا لكل عمود متلما فعلنا في البيانات أعلى فكتورنا درجة الحرارة واستهلاك المياه في أعلى العمودين. بعد ذلك يمكننا أن نضغط OK



وبذلك نحصل على النتيجة في صفحة منفصلة كالتالي:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.971398157							
R Square	0.943616323							
Adjusted R Square	0.941360976							
Standard Error	9.512370655							
Observations	27							
ANOVA								
	dF	SS	MS	F	Significance F			
Regression	1	37858.16646	37858.16646	418.3907329	3.97825E-17			
Residual	25	2262.12984	90.48519358					
Total	26	40120.2963						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	34.42308259	11.48643323	2.996848966	0.006084088	10.76634708	58.07981809	10.76634708	58.07981809
درجة الحرارة	8.673929045	0.424057586	20.45460175	3.97825E-17	7.800566709	9.54729138	7.800566709	9.54729138

نتيجة تحليل الانحدار في إكسيل تحتوي على أرقام كثيرة ولكن ذلك لا ينبعني أن يصييك بالذعر من هذا التحليل فهناك أرقام محدودة نريد التعرف عليها وهي التي قد غطت بخلفية صفراء. ولمزيد من الوضوح فإن هذه الأرقام نعرضها منفصلة كالتالي

R Square	0.943616323	
	Coefficients	P-value
Intercept	34.42308259	0.006084088
درجة الحرارة	8.673929045	3.97825E-17

ما معنى هذه البيانات؟

وهي رقم يتراوح بين صفر والواحد الصحيح وهذا الرقم يبين نسبة التغير في استهلاك المياه التي يمكننا توضيحه من خلال هذه المعادلة (يأتي بيانها لاحقاً). أي ببساطة هو مؤشر لمدى توضيح تحليل الانحدار لقيمة المتغير الذي نحاول التنبؤ به وهو في هذه الحالة استهلاك المياه. في حالتنا هذه فإن هذا الرقم R Square يساوي

0.94 وهو رقم يقترب جداً من الواحد الصحيح مما يعني أن نموذج تحليل الانحدار قوي جداً ويمكننا من حساب قيمة استهلاك المياه بشكل يقترب كثيراً من الصحة. أما لو كانت قيمة هذا الرقم هي 0.3 فإننا يجب علينا إهمال المعادلة التي حصلنا من هذا التحليل والبحث مرة أخرى عن العوامل المؤثرة في استهلاك المياه.

Intercept هو قيمة الثابت β في المعادلة $y = \beta + \alpha x$ والتي تعني في حالتنا استهلاك المياه $= \beta + \alpha \times \text{درجة الحرارة}$

وبالتالي فإن $\beta = 34.42$ ولذلك فإن استهلاك المياه $= 34.42 + \alpha \times \text{درجة الحرارة}$ **Coefficient** المعاملات والتي تبين قيمة التقطاع مع المحور الرأسي ومعاملات المتغيرات فمثلاً قيمة التقطاع مع المحور الرأسي هي 34.42 وقيمة معامل درجة الحرارة هو 8.67

وبالتالي فإن النموذج الرياضي (المعادلة الرياضية) التي حصلنا عليها هي كالتالي:

$$\text{استهلاك المياه} = 34.42 + 8.67 \times \text{درجة الحرارة}$$

P value هي قيمة تظهر إن كان العامل المقابل لها في نفس الصف هو عامل مؤثر فعلاً أم لا. في هذه الحالة لدينا عامل واحد وهو درجة الحرارة والتي يقابلها **P Value** بقيمة 3.98E-17

وهو يعني 0.000000000000000398

هل هذه القيمة مقبولة أم لا؟ إذا كان **P Value** أقل من 0.05 فإن العامل المقابل لها (درجة الحرارة في هذا المثال) هو عامل مؤثر في المتغير الذي نحاول دراسة تغييره (استهلاك المياه) وبالتالي فمن الواضح أن درجة الحرارة هي عامل مؤثر فعلاً في استهلاك المياه. وقد تعتبر العامل مؤثراً حتى قيمة **P Value** تساوي 0.1 ولكن إن زادت عن 0.1 فإن هذا العامل يجب استبعاده من النموذج فهو غير مؤثر.

هل درجة الحرارة تؤثر في استهلاك المياه؟

ليس معنى أننا توصلنا إلى علاقة مقبولة إحصائياً بعد دراسة **R Square** و **P Value** أن سبب تغيير y ص ولكننا توصلنا إلى علاقة المصاحبة بين x و y أي أن تغير استهلاك المياه يصاحب تغير درجة الحرارة. أي أنه كما قلنا في دراسة العلاقة بين متغيرين عن طريق منحنى الانتشار scatter diagram ومعامل الارتباط Correlation فإننا ندرس وجود مصاحبة بين المتغيرين أو العلاقة بينهما ولكننا لا نستطيع لمجرد وجود علاقة أن نقول أن هذا يتسبب في حدوث ذاك. فمثلاً لو درسنا العلاقة بين استهلاك الكهرباء واستهلاك المياه عن طريق تحليل الانحدار فقد نصل لنموذجاً مقبولاً للعلاقة بينهما ولكن ذلك لا يعني اطلاقاً أن زيادة استهلاك الكهرباء تؤدي لزيادة استهلاك المياه ولكننا قد نقول أن كلاً منهما يزيد مع زيادة درجة الحرارة نتيجة لتشغيل مكيفات الهواء وزيادة الحاجة للاستحمام بالماء.

والخلاصة أننا لكي نقول أن شيئاً ما يتسبب في حدوث آخر فلابد لنا من فهم طبيعة هذه المتغيرات ودعم قولنا بدراسة أو تحليل منطقي. ففي مثلكنا هذا يمكن أن نتوقع علاقة السببية هذه لأن ارتفاع درجة الحرارة فعلاً يؤدي إلى زيادة استخدام الناس للمياه في الاستحمام والشرب. ولكن لابد من دعم ذلك بعمل استبيان مثلاً لأسباب زيادة استهلاك المياه عند ارتفاع درجة الحرارة.

ما معنى هذه المعادلة؟

توصلنا إلى هذه المعادلة فما معناها؟

$$\text{استهلاك المياه} = 34.42 + 8.67 \times \text{درجة الحرارة}$$

إن هذه المعادلة تعني أن زيادة درجة الحرارة بدرجة واحدة مؤوية تعني زيادة استهلاك المياه بـ 8.67. وتعني كذلك

أن استهلاك المياه عند درجة حرارة صفر هي 34.42 ويمكننا استخدام المعادلة للتنبؤ بقيمة استهلاك المياه عند درجة حرارة 40 درجة مئوية
استهلاك المياه = $8.67 + 34.42 \times \text{درجة مئوية}$ 40 = 381 ويمكننا توقع استهلاك المياه عند درجة حرارة 18 درجة مئوية بـ 190

هذه مقدمة لتحليل الانحدار الخطي البسيط. في المقالات التالية إن شاء الله نتعرف بشكل أكثر عمقاً وأكثر دقة على هذا التحليل ونتعرف كذلك على تحليل الانحدار المتعدد والذي يربط متغيراً بعوامل عديدة.

[مقالات ذات صلة:](#)

[تحليل البيانات](#)

[من مراجع الموضوع](#)

Lean Six Sigma Pocket ToolBook, M. George et al., McGrawHill, 2005

Applied Statistical Methods, W. Carlson and B. Thorne, Prentice Hall, 1997

Statistics for Managers, Levine et al., Prentice Hall, 1999

[موقع ذات صلة:](#)

[الانحدار الخطي البسيط](#)

[Regression Analysis 1 – Video](#)

[Regression Analysis 2 – Video](#)

الانحدار الخطي المتعدد Multiple Regression

سبتمبر 7, 2009

ناقشت في المقالة السابقة تحليل الانحدار الخطي البسيط Regression Analysis والذي يربط متغير بمتغير واحد فقط. في هذه المقالة نناقش كيفية تحليل علاقة متغير بعدة متغيرات أخرى باستخدام تحليل الانحدار الخطي المتعدد.

لماذا؟

ما هي الحاجة للانحدار الخطي المتعدد؟ إننا كثيراً ما نحتاج تحديد العوامل التي تؤثر في متغير ما مثل تحديد العوامل التي تؤثر في حجم المبيعات أو التي تؤثر في عدد الأعطال أو عدد عيوب المنتج. وفي معظم الأحيان يكون لدينا عوامل عدة تؤثر في ذلك المتغير الذي نريد دراسته فلا يمكننا الاعتماد على تحليل الانحدار الخطي البسيط. الانحدار الخطي المتعدد يساعدنا على دراسة تأثير هذه العوامل على المتغير محل الدراسة مثل دراسة تأثير كل من جودة المنتج وسعر البيع وزمن التسليم وعدد منافذ التوزيع وسعر المنافسين على حجم المبيعات.

تحليل الانحدار الخطي المتعدد Multiple Linear Regression

تحليل الانحدار المتعدد يختلف قليلاً عن تحليل الانحدار البسيط في حالة الانحدار البسيط فإننا ندرس العلاقة بين المتغير محل الدراسة ومتغير آخر نتصور أنه يؤثر فيه. ولكن في حالة الانحدار المتعدد فإنه لدينا عدداً كبيراً من المتغيرات التي قد تكون مرتبطة بالمتغير محل الدراسة علينا استخلاص تلك التي لها علاقة حقيقية بهذا المتغير واستبعاد الباقي ثم إن علينا تحديد العلاقة بين هذه المتغيرات وهذه المتغيرات المؤثرة. وفي هذه الحالة فإننا نهدف إلى الوصول إلى معادلة شبيهة بالمعادلة التالية:

$$ص = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

حيث ص هي المتغير محل الدراسة وس 1 وس 2 وس 3 وس 4 هم المتغيرات أو العوامل المؤثرة في المتغير ص. فمثلاً ص هي حجم المبيعات وس 1 هي عدد منافذ البيع وس 2 هي جودة المنتج وس 3 هي سعر المنتج وس 4 هي حجم الإنفاق على الدعايا.

تقييم نموذج الانحدار المتعدد:

في حالة الانحدار البسيط فإننا بینا أهمية معامل التحديد R^2 ولكن في حالة الانحدار المتعدد فإننا نهتم بمعامل التحديد المعدل $R^2_{Adjusted}$ لأن قيمة معامل التحديد R^2 تزداد بشكل طبيعي كلما أضفنا متغيراً بمعنى أن قيمتها عندما ندرس علاقة المتغير بمتغيرين ستكون أكبر منها عند استبعاد أحدهما. وهذا لا يساعدنا على معرفة ما إذا كان هذا المتغير الإضافي قد أفاد في التحليل أم لا. أما مع معامل التحديد المعدل $R^2_{Adjusted}$ فإن هذا لا يحدث لأن طريقة حسابه تأخذ في الاعتبار عدد المتغيرات الدالة في التحليل. لذلك فإننا لكي نعرف إن كان إضافة متغير لها تأثير إيجابي على النموذج الرياضي (المعادلة التي تربط المتغير التابع بالمتغيرات المستقلة) فإننا ننظر إلى معامل التحديد المعدل $R^2_{Adjusted}$.

وهناك مقياس آخر أكثر دقة من معامل التحديد المعدل F Test $R^2_{Adjusted}$ وهو F Test $R^2_{Adjusted}$ وبدون الدخول في تفصيلات إحصائية فإن قيمة F test تزداد كلما تحسن النموذج أي أننا لو أضفنا متغيراً له علاقة إحصائية مؤثرة بالمتغير محل الدراسة فإن قيمة F test تزداد. وهذا مشابه لما ذكرناه في F test $R^2_{Adjusted}$ غير أن قيمة F test لا تتراوح بين صفر وواحد بل تأخذ أي قيمة. والأمر المهم أننا نستطيع معرفة مدى دقة F test significance لقيمة F test كما كان نعرف تأثير أي متغير عن طريق P value. فإذا كان أقل من أو يساوي 0.05 فإن هذا يعني أن النموذج مقبول إحصائياً وأما إذا زاد عن ذلك فإن

النموذج يكون غير مقبول.

كيف يتم تقييم عدة متغيرات؟

هناك طريقتان لذلك. الأولى وهي الطريقة التدريجية. أن نبدأ باعتبار متغير واحد ونسجل قيمة معامل التحديد R^2 ثم نضيف متغيراً آخر ونسجل قيمة معامل التحديد المعدل $R^2_{Adjusted}$ ونقارنها بالسابقة فإن كانت قيمتها قد زادت فإننا نبني على هذا المتغير وإن كانت قيمتها قد نقصت فإن هذا يعني أن هذا المتغير غير مرتبط بالمتغير محل الدراسة. ثم نضيف متغيراً آخر وهكذا. فمثلاً لو كان درس علاقة انتشار مرض ما بعوامل مثل نقاط مياه الشرب والمستوى التعليمي وتتوفر مراكز صحية والحال الاجتماعية فإننا نبدأ بدراسة علاقة انتشار المرض بنقاط مياه الشرب ثم نضيف المستوى التعليمي فإن زادت قيمة معامل التحديد المعدل R^2 فإن هذا يعني أن المستوى التعليمي هو عامل مؤثر ولكن إن قلت فإن هذا يعني أنه غير مؤثر. ثم نضيف توفر المراكز الصحية وهكذا. وفي نفس الوقت فإننا ننظر إلى قيمة F test قيمة F significance وكلما زادت الأولى فالنموذج يتحسن وعندما تكون الثانية أقل من 0.05 فإن النموذج يكون مقبولاً.

الطريقة الثانية هي أن نأخذ في الاعتبار كل المتغيرات ثم نبدأ في استبعاد واحداً تلو الآخر ونقارن قيمة معامل التحديد المعدل $R^2_{Adjusted}$ بنفس الطريقة. فمثلاً نحن نريد تحديد العوامل المؤثرة في حجم المبيعات وتحديد نموذج رياضي لعلاقة حجم المبيعات بهذه العوامل. فنبدأ بدراسة تحليل الانحدار بين حجم المبيعات وكل هذه العوامل مثل السعر والجودة وحجم الدعايا وعدد منافذ البيع وسعر المنتج المنافس. ثم نسجل قيمة معامل التحديد المعدل $R^2_{Adjusted}$ وبعد ذلك نستبعد أحد هذه المتغيرات ونرى تأثير ذلك على قيمة معامل التحديد المعدل $R^2_{Adjusted}$. ويتم مراعاة F test وF Significance كما ذكرنا أعلاه.

ولا يوجد ما يمنع من اتباع أسلوب وسط وهو أن نأخذ في الاعتبار بعض العوامل التي لدينا قناعة قوية بتأثيرها ثم بعد ذلك نبدأ في إضافة المتغيرات الأخرى تباعاً. عموماً الاختيار بين هذه الطرق لا يمثل مشكلة فكلها تؤدي في النهاية لنفس النتيجة.

كيف سنحدد المتغير الذي نستبعده؟ كما تذكر فإننا نهتم بقيمة P Value لأنها تعني ما إذا كان هذا المتغير مؤثر أم لا. في حالة الانحدار المتعدد فإننا نأخذ في الاعتبار قيمة P Value فنبدأ بحذف المتغير الذي له قيمة P Value كبيرة وخاصة تلك التي تتجاوز 0.05.

تحليل الارتباط بين المتغيرات Multicollinearity

قبل القيام بتحليل الانحدار الخطي المتعدد فإن علينا التخلص من بعض المتغيرات المرتبطة بعضها. فمثلاً لا يمكن حساب العلاقة بين مستوى الطالب اعتماداً على نسبة الحضور ونسبة الغياب ونتيجة العام السابق. لماذا؟ لأن نسبة الحضور ونسبة الغياب هما شيئاً يقيسان نفس الشيء فهذه هي واحدة منقوصاً منه الأخرى بمعنى أنه لو كانت نسبة الحضور هي 80% فإن نسبة الغياب ستكون 20% وهكذا. وهذا أمر منطقي وهو يؤدي لمشاكل في نموذج تحليل الانحدار. لذلك يجب أن نقوم بدراسة معامل الارتباط بين كل المتغيرات قبل إدخالها في تحليل الانحدار.

وفي حالة وجود ارتباط قوي بين متغيرين فإنه يجب استبعاد أحدهما. وهنا يكون قرارنا في اختيار المتغير الذي نستبعده بناءً على فهمنا لطبيعة الموضوع الذي ندرس. وكقاعدة عامة فإن الارتباط القوي الذي يثير الفرق في تحليل الانحدار المتعدد يمكن تحديده بقيمة معامل الارتباط أكبر من 0.9 ويجب أيضاً التفكير فيما له معامل ارتباط بين 0.8 و 0.9.

ويجب أن ندرس المتغيرات أيضاً بحيث لا تكون هناك مجموعة متغيرات في محصلتها مرتبطة بمتغير آخر. وهذا يجب تحليله بناءً على فهمنا للمتغيرات وعلاقتها بعضها. وعلى سبيل المثال فإنه لا يمكن أن نبني نموذج انحدار يقيس العلاقة بين مستوى أداء الطالب وعدد أيام الحضور وعدد أيام الغياب وعدد أيام الأجزاء المرضية وذلك لأن عدد أيام الأجزاء المرضية وعدد أيام الغياب هما مقاييس مباشر لعدد أيام الحضور. وكذلك فإنه لا يمكن أن ندرس

تأثير كل عنصر من العناصر الكيميائية المضافة لمعدن ما على المتانة ونأخذ في الاعتبار أيضاً عامل يكفي كل هذه العناصر لأن هذا العامل هو محصلة كل العناصر.

حجم العينة Sample Size

لابد أن تكون عدد الحالات التي يتم استخدامها كافية للقيام بهذا التحليل وكقاعدة عامة فلابد من توفر عدد من الحالات يزيد عن عشرة أمثال إلى عشرين مثل عدد المتغيرات المستقلة. أي أننا لو كنا نريد قياس حجم المبيعات بناء على عدد منافذ التوزيع والإتفاق على الدعايا وسعر البيع فإننا نحتاج بين 30 إلى 60 بيان للمبيعات لكي يمكننا إجراء تحليل الانحدار. وهذه القاعدة فيها كلام كثير بين الإحصائيين وهناك من قال إن خمسة أمثال عدد المتغيرات يكفي وهناك من اعترض على القاعدة من أساسها. وتتجذر الإشارة إلا أنه في حالة استخدام التحليل التدريجي فإننا نحتاج إلى عدد أكبر من الحالات أي إلى عينة أكبر يصل فيها عدد الحالات إلى أكبر من 40 مرة عدد المتغيرات أي 120 في المثال السابق.

والخلاصة لنا كتطبيقين لا نقوم بدراسة عدد كبير من المتغيرات من عينة صغيرة فلا تستخدم عينة حجمها عشرين حالة لدراسة تأثير خمس متغيرات. وفي الواقع العملي قد يكون من السهل جداً زيادة حجم العينة وفي هذه الحالة فحاول الالتزام بالحدود العليا وفي بعض الحالات قد تكون هناك صعوبة بالغة في زيادة حجم العينة فحاول الالتزام بحد 5 أو 10 أو 20 مثل عدد المتغيرات. ويمكنك أن تركز على المتغيرات الأكثر أهمية في حالة صغر حجم العينة وعدم القدرة على زيتها.

مثال:

افتراض أننا نريد معرفة العوامل المؤثرة على عيوب المنتج ولدينا تصور أن هناك ثلاثة متغيرات قد تكون مؤثرة في جودة المنتج وهي: خبرة العامل، سرعة الماكينة ودرجة حرارة سائل التبريد.

سرعة الماكينة	درجة حرارة سائل التبريد	خربة العامل	العوب
45	34	10	33
55	36	12	35
60	37	11	38
50	31	10	29
60	32	10	38
60	35	5	39
65	36	4	44
65	32	5	43
50	31	11	34
60	34	7	38
55	35	6	39
55	33	11	31
50	31	10	34
50	31	12	32
50	32	9	34
65	36	5	39
55	35	11	35
60	31	4	40
50	30	12	28
60	37	11	35
60	32	13	30
55	36	9	33
50	34	11	33
45	37	12	26
60	30	4	40
65	31	5	41
50	32	6	33
50	33	7	31
50	34	9	33
60	35	6	38

نبدأ بتجمیع البيانات وتسجيلها وافترض أن البيانات كانت كالتالي:

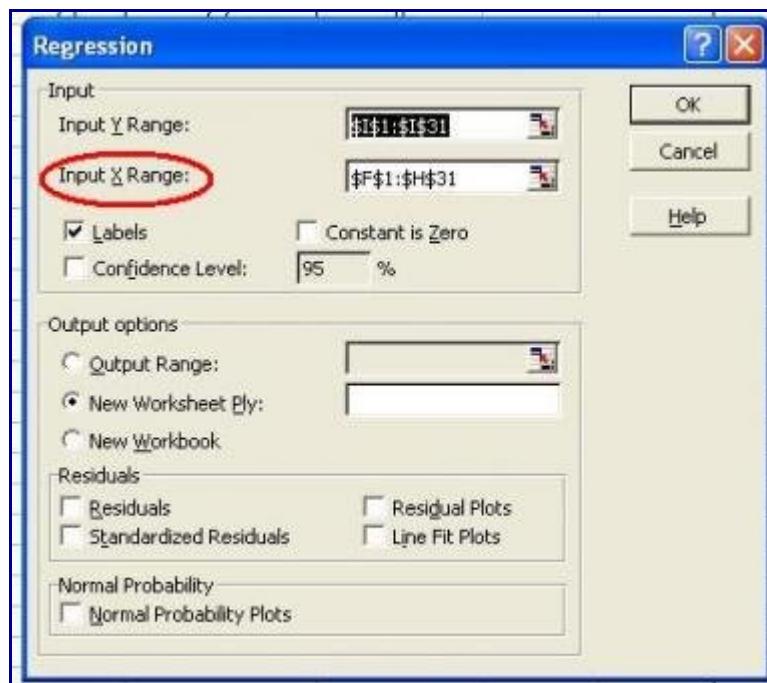
علينا الآن أن نستخدم برنامج إکسل لكي نقوم بتحليل الانحدار الخطى المتعدد. ولكن قبل ذلك علينا التأكد من الأمور التي ناقشناها في هذه المقالة. هل العينة مناسبة؟ لدينا 30 حالة مسجلة ونحن نريد دراسة ثلاثة عوامل فكما وأننا قد حققنا نسبة 10 أمثل تقریبا. ونعتبر هذا مقبولا ولا مانع من استخدام عينات أكبر في الواقع.

وهناك برامج أخرى تقوم بتحليل الانحدار مثل SPSS ولكننا نفضل التطبيق على برنامج إکسل لسعة استخدامه. الأمر التالي هو دراسة ارتباط العوامل بعضها. نقوم باستخدام تحليل الارتباط Correlation من إکسل. ونحصل على النتيجة التالية

سرعة الماكينة	درجة حرارة سائل التبريد	خربة العامل	العوب
سرعة الماكينة	1		
درجة حرارة سائل التبريد	0.149	1	
خربة العامل	-0.560	0.064	1
العوب	0.794	0.109	-0.753

هذا أمر جيد فلا يوجد ارتباط كبير بين المتغيرات الثلاث ولذلك فلا مانع من استخدامهم كلهم في التحليل.

سنستخدم نفس الطريقة التي اتبناها في الانحدار الخطي البسيط. نختار Data Analysis Tools ثم Regression . الفارق هو أننا عندما نكتب الخلايا التي تمثل المتغيرات المؤثرة في العيوب فإننا نختار كل الخلايا المكتوب فيها المتغيرات الثلاث: خبرة العامل، درجة حرارة سائل التبريد، سرعة الماكينة. وتجدر الإشارة إلا أنها كلها يجب أن تكون متلاصقة. لاحظ أنني كتبت بيانات العيوب في العمود I والمتغيرات الثلاث في الأعمدة: G, H



سنبدأ بتحليل الانحدار مستخدمنا الثلاثة عوامل كلها. فنحصل على النتيجة التالية (سوف أحاوِل الاقتصار على عرض الأرقام التي سنستخدمها في التحليل):

R Square	0.7726		
Adjusted R Square	0.7463		
		F	Significance F
Regression	29.4395	1.61899E-08	
	Coefficients	P-value	
Intercept	15.9661	0.0409	
سرعة الماكينة	0.3808	0.0001	
درجة حرارة سائل التبريد	0.1186	0.5343	
خبرة العامل	-0.6903	0.0004	

تعتبر هذه النتيجة طيبة لأن قيمة معامل التحديد المعدل Adjusted R Square هي 0.74 ونلاحظ أن Significance F هي تقريباً صفر فهي تحديداً 0.000000016. عندما نستعرض P Value نلاحظ أن درجة حرارة سائل التبريد تبدو غير مؤثرة حيث أن P Value هي 0.1186 أي أكبر من 0.05. لنجرِب حذف هذا المتغير.

R Square	0.7691		
Adjusted R Square	0.7520		
		F	Significance F
Regression	44.9656	2.5489E-09	
		Coefficients	P-value
Intercept	19.1112	0.0016	
سرعة الماكينة	0.3926	0.0000	
خبرة العامل	-0.6708	0.0004	

هل النتيجة الآن أفضل أم أسوأ؟ لقد زاد معامل التحديد المعدل Adjusted R Square من 0.746 إلى 0.752. لقد زادت قيمة F من 29.4 إلى 44.9 وقلت قيمة F Significance. وكل هذه مؤشرات تحسن كما أوضحتنا في بداية المقالة. نعيد النظر في P Value فنجد أنها أقل من 0.05. فنحن توصلنا فعلاً إلى أنه توجد علاقة خطية بين عدد العيوب وخبرة العامل وسرعة الماكينة وأنه لا توجد علاقة خطية بين عدد العيوب ودرجة حرارة سائل التبريد.

يمكننا أن نكتب معادلة عدد العيوب كالتالي:

$$\text{عدد العيوب} = 19.11 + 0.392 \times \text{سرعة الماكينة} - 0.67 \times \text{خبرة العامل}$$

نحن كمدربين أو كمحاضرين أو كمهندسين توصلنا لأمور مهمة فسنعمل تأثير درجة حرارة سائل التبريد في سعينا لتقليل العيوب. من الواضح أن زيادة سرعة الماكينة تؤدي إلى زيادة العيوب بينما زيادة خبرة العامل تؤدي إلى نقصان العيوب. هذا واضح من المعاملات Coefficients فمعامل سرعة الماكينة موجب بينما معامل خبرة العامل سالب. وهذا أمر يقبله العقل فالافتراض أنه كلما زادت خبرة العامل زادت مهارته وقلت أخطاؤه.

يمكننا معالجة الأمور بأن نحاول تدريب العاملين محدودي الخبرة وتحديد أخطائهم المتكررة وتوضيحها لهم ونبين لهم كيفية تجنبها. يمكننا كذلك أن ننصح العمال محدودي الخبرة باستخدام سرعة متوسطة لأن اجتماع السرعة العالية مع قلة الخبرة تؤدي إلى زيادة العيوب كثيراً.

هكذا ترى قيمة تحليل الانحدار المتعدد وكيفية استخدامه. في مقالات تالية إن شاء الله نستعرض المزيد من الأمثلة وننظر إلى بعض المحاذير عند استخدام تحليل الانحدار.

[مقالات ذات صلة:](#)

[تحليل البيانات](#)

[من مراجع الموضوع:](#)

Lean Six Sigma Pocket ToolBook, M. George et al., McGrawHill, 2005

Discovering Statistics using SPSS for Windows, A. Field, Sage, 2003

Statistics for Managers, Levine et al., Prentice Hall, 1999

[موقع ذات صلة:](#)

[الانحدار المتعدد](#)

[الانحدار المتعدد](#)

[Multiple Regression](#)

Multiple Regression

LISREL approaches to interactive effects in multiple regression

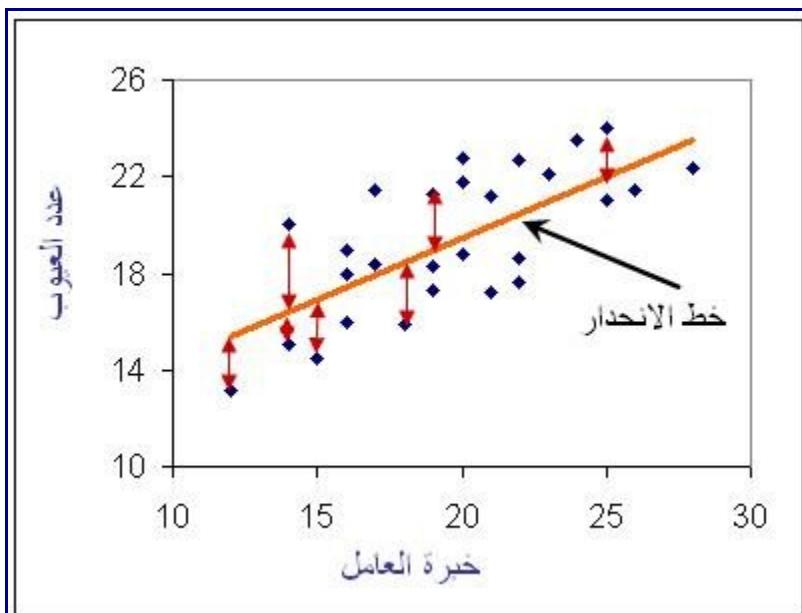
تحليل الانحدار – دراسة الباقي

نوفمبر 11, 2009

نستكمل في هذه المقالة موضوع تحليل الانحدار الخطي Linear Regression فنستعرض فرضيات تحليل الانحدار وكيفية التحقق منها. والأسلوب المستخدم للتأكد من تحقق هذه الفرضيات هو تحليل الباقي.

الباقي Residuals

ما هي الباقي؟ الباقي Residuals هي الفرق بين القيمة التي نحسبها من نموذج الانحدار والقيمة الحقيقية. فمثلاً إذا قمنا بتحليل الانحدار لحجم المبيعات بناء على سعر البيع والجودة وعدد منافذ البيع فالباقي هو الخطأ في النموذج. فعند مقارنة حجم المبيعات لإحدى الحالات المعلومة لدينا بنتيجة النموذج فإننا نجد فارقاً بينهما وهذا الفارق هو الخطأ في النموذج أو الباقي. يمكن أن نقول أن وجود هذا الفارق أو الخطأ هو من طبيعة تحليل الانحدار فمن النادر أن يكون تحليل الانحدار صحيحاً بنسبة مائة بالمائة.



هذا الشكل يوضح الباقي عند القيام بتحديد خط الانحدار أي تحديد العلاقة الخطية بين خبرة العامل وعدد العيوب في المنتج. البيانات الأساسية التي سجلناها هي عبارة عن النقاط الزرقاء المبعثرة. وعندما رسمنا خط الانحدار وهو الخط المستقيم باللون البرتقالي فإنه لا ينطبق بطبيعة الحال على كل النقاط. ولذلك هناك فارق بين البيانات المسجلة وعدد العيوب التي ستنتجها من خط الانحدار أو معادلته. هذا الفارق هو الفارق بين كل نقطة من البيانات الأساسية وقيمة خط الانحدار. فعلى سبيل المثال فإنه إذا كانت خبرة العامل هي 25 عاماً فإن البيانات المسجلة تبين أن عدد العيوب في المنتج كانت 24 تقريباً ولكننا لو استخدمنا خط الانحدار لوجدناه يعطينا قيمة مختلفة وهي 21 تقريباً. الفارق بين القيمة الحقيقة وتلك التي نحصل عليها من معادلة الانحدار أو خط الانحدار هو الباقي وهو بالنسبة لهذه النقطة يساوي $24 - 21 = 3$.

عندما قررنا أن نستخدم تحليل الانحدار الخطي فإننا افترضنا أن العلاقة بين عدد العيوب وخبرة العامل هي علاقة خطية أي أن $\text{عدد العيوب} = \text{ثابت} + \text{معامل} * \text{خبرة العامل بالسنين} + \text{خطأ}$. فنحن نفترض علاقة خط مستقيم. لذلك فإن أي شيء يبين أن العلاقة ليست خطية فهو ببساطة يهدم فرضتنا الأساسي وبالتالي يجعلنا نبحث عن طريقة أخرى لدراسة العلاقة بين المتغيرين.

فرضيات تحليل الانحدار:

يبني تحليل الانحدار على عدة فرضيات لابد أن نضمن صحتها عند إجراء هذا التحليل. هذه الفرضيات هي:

1- علاقة خطية Linearity بمعنى أن العلاقة هي علاقة خط كستقيم وليس خطًا منحنياً

2- التجانس Homoscedasticity ومعنى ثبات التغيير (التباین) في قيمة الباقي. عندما يكون هناك تجانس فإن الباقي ستكون متساوية إلى حد ما عند جميع القيم أو بمعنى آخر لنلاحظ اتجاه لزيادة أو نقصان الباقي مع تغير قيمة المتغير المستقل. فمثلاً لو حاولنا دراسة العلاقة بين حجم المبيعات وسعر البيع فإننا لنلاحظ أن الباقي تتوجه للزيادة مع زيادة سعر البيع.

3- استقلالية الباقي Independence of Residuals بمعنى أن الباقي لأي نقطة لا يعتمد على الباقي في النقطة أو النقاط السابقة. عندما تكون الباقي غير مستقلة فإننا نحتاج أن نستخدم نموذجاً آخر يأخذ في الاعتبار هذه العلاقة.

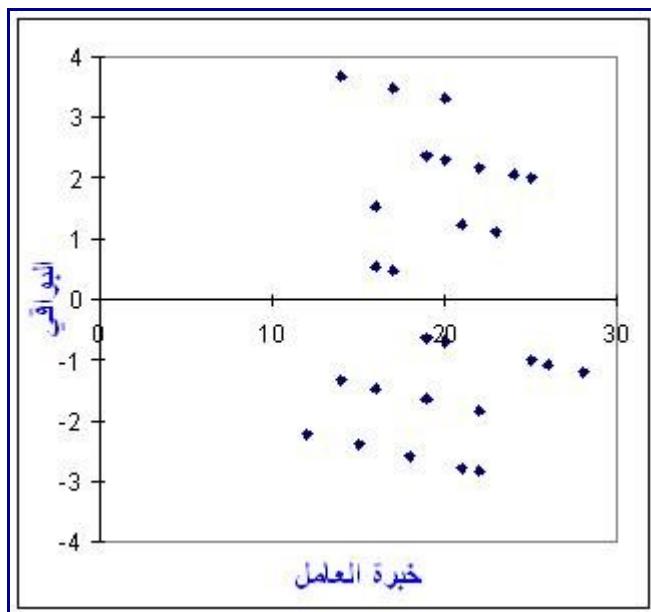
4- التوزيع الطبيعي للباقي Normality of Residuals يبني تحليل الانحدار على أن الباقي موزعة توزيعاً طبيعياً عند كل النقاط للمتغير المستقل مثل سعر البيع. وهذا يعني أنها تتغير من سالب لموجب حول قيمة الصفر بشكل توزيع طبيعي وبحيث يكون مجموعها صفرًا.

هذه هي الفرضيات باختصار وسوف نزيد الأمر وضوحاً في الأقسام التالية.

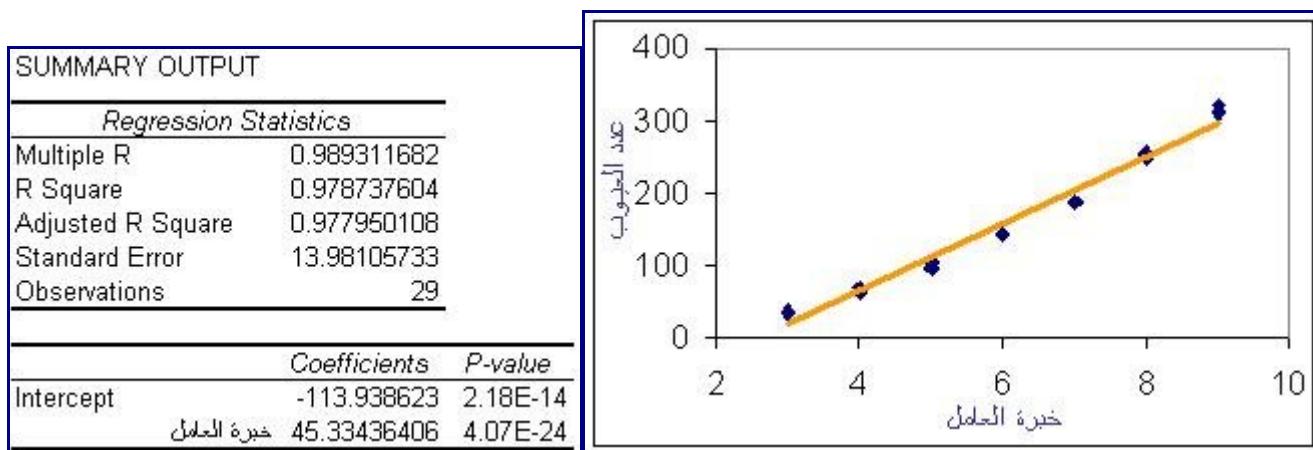
التأكد من تحقق فرضيات تحليل الانحدار:

لكي نتأكد من أن البيانات التي ندرسها تخضع لفرضيات التي نفترضها في تحليل الانحدار الخطى فإننا نلجأ لدراسة الباقي فنرسم مجموعة من الرسومات البيانية التي تبين تتحقق هذه الفرضيات من عدمه.

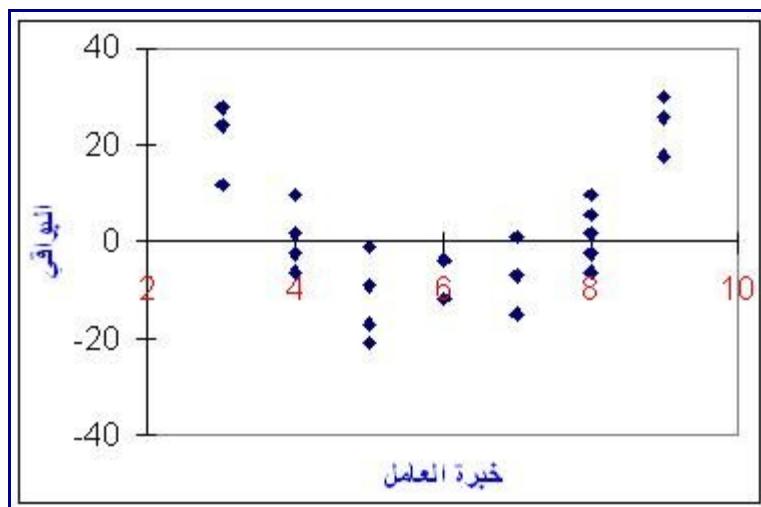
العلاقة بين الباقي وكل متغير مستقل: لابد أن تظهر هذه العلاقة كنقط مبعثرة بشكل عشوائي في الاتجاهين السالب والموجب بدون وجود أي شكل أو منحنى. الشكل أدناه يبين هذه العلاقة للمثال السابق. لاحظ أن النقاط لا تأخذ شكلاً محدداً وهو ما يعني أن الخطأ هو خطأ عشوائي.



انظر إلى الشكل أدناه والذي يبيّن علاقة أخرى بين عدد العيوب وخبرة العامل. إن العلاقة هنا تأخذ شكلاً مختلفاً ولكن ربما لا تلاحظ شيئاً مزعاً في خط الانحدار. وعندما ندرس نتائج تحليل الانحدار نجد أن النتائج مرضية فنسبة R square كبيرة وقيمة P صغيرة جداً ولكن دعنا ننظر لشكل الباقي.



الشكل أدناه يبين العلاقة بين الفوائد وخبرة العامل. المحور س يبين خبرة العامل والمحور ص يبين الباقي المناظر لها. ماذا نلاحظ في هذا الشكل؟ إن هناك اتجاه واضح للفوائد فهي كانت موجبة ثم سالبة ثم موجبة مرة أخرى. هذا يعني أن العلاقة هي علاقة غير خطية. فلو كانت العلاقة خطية لما وجدنا هذا الاتجاه الواضح للباقي ولو جدنا الباقي مبعثرة بشكل عشوائي. ففي مثالنا هذا يكون الأخذ بالنموذج الخطي هو أمر غير صحيح لأن العلاقة في الحقيقة غير خطية.



ماذا لو أخذنا بهذه العلاقة الخطية وأهملنا مشكلة الباقي؟ في هذه الحالة تكون قد استخدمنا معادلة غير مناسبة وهذا يعني أننا لو استخدمنا هذا النموذج لتوقع عدد العيوب المناظرة لخبرة عامل ما فإن هناك خطأ في التقدير. ربما سنتقول لي ولكن العلاقة البيانية بين خط الانحدار والنقط الأصلية لا يبدو كبيرا في الشكل الأول؟ هذا صحيح ولكن مقاييس الرسم يتدخل في هذا الأمر. لو نظرت إلى رسم الباقي لوجدت أن الخطأ في تقدير عدد العيوب المناظرة لخبرة عامل 3 سنوات يصل إلى أكثر من عشرين. هل هذا خطأ بسيط؟ لو نظرت إلى الرسم لاحظت أن عدد العيوب الحقيقي المناظر لخبرة عامل 3 سنوات يتراوح بين 35 و 45 تقريريا. فنسبة الخطأ هنا هي حوالي $40 / 20 = 50\%$. هل هذه نسبة مقبولة.

دعك من هذا. لنرى كيف يمكننا استخدام النموذج لتوقع نسبة الخطأ المناظرة لعامل لديه خبرة قدرها سنة واحدة. إن النموذج الرياضي الذي استنتاجناه هو

$$\text{عدد العيوب} = 45.33 + 113.9 * \text{خبرة العامل} + \text{باقي (الخطأ)}$$

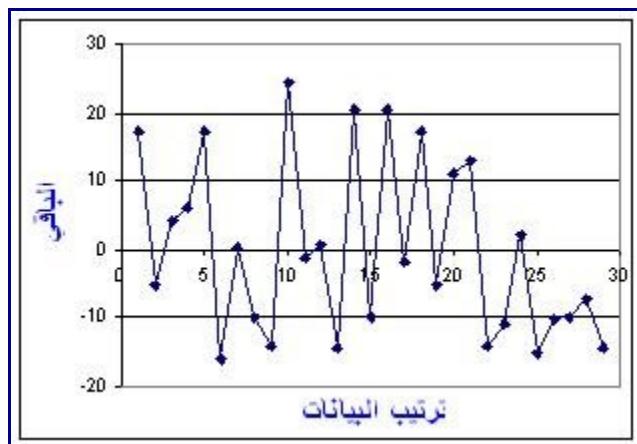
عند التعويض بخبرة عامل قدرها سنة واحدة نحصل على عدد العيوب = 68-

بالطبع لا توجد عيوب أقل من الصفر. فالمعادلة هنا غير معبرة بالمرة. ماذا لو قدرنا الخطأ لعامل لديه خبرة قدرها عشرين سنة؟ إن النتيجة تكون 792. هل هذه نتيجة صحيحة؟ باستخدام العلاقة الحقيقية التي أنشأنا بها هذه البيانات لاستخدامها في هذا المثال وهي:

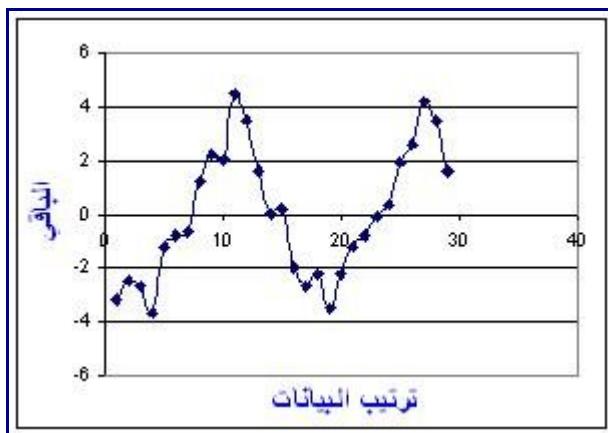
$$\text{عدد العيوب} = 3.85 + 2.5 * \text{عدد العيوب}^2$$

فإننا نجد أن نسبة العيوب المناظرة لخبرة عامل عشرين سنة هي 1542. فنسبة الخطأ هنا تقارب 50%. ربما في بعض النقاط نجد نسبة الخطأ قليلة جداً وتقترب من الصفر ولكن هذا لا يعني صحة العلاقة بشكل عام. بالطبع هذا المثال هو مثال توضيحي ولا يقصد بعلاقة الحقيقة بين خبرة العامل وعدد العيوب.

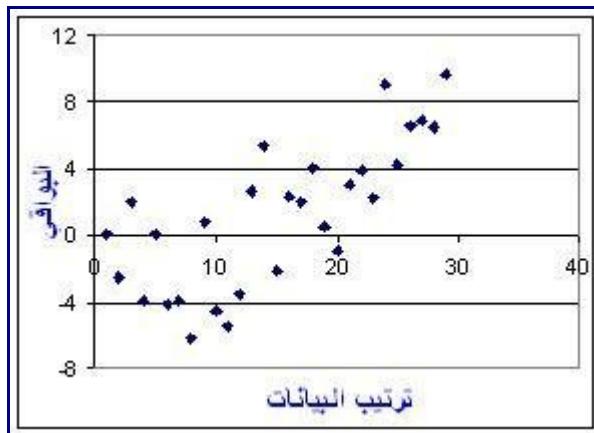
العلاقة بين الباقي وترتيب البيانات: يمكننا كذلك رسم العلاقة بين الباقي وترتيب تسجيل البيانات والذي ينبغي إلا يظهر اتجاهها متزايداً أو شكل دوري متكرر. إن أحد فرضيات تحليل الانحدار هو استقلالية الباقي أي أن الباقي عند أي نقطة لا يعتمد على قيمة الباقي عند النقطة السابقة (أو النقاط السابقة) أي أن الباقي عشوائية.



هذا الشكل لا يظهر أي تزايد أو أي تغير دوري للباقي. ولكن انظر إلى المثال التالي

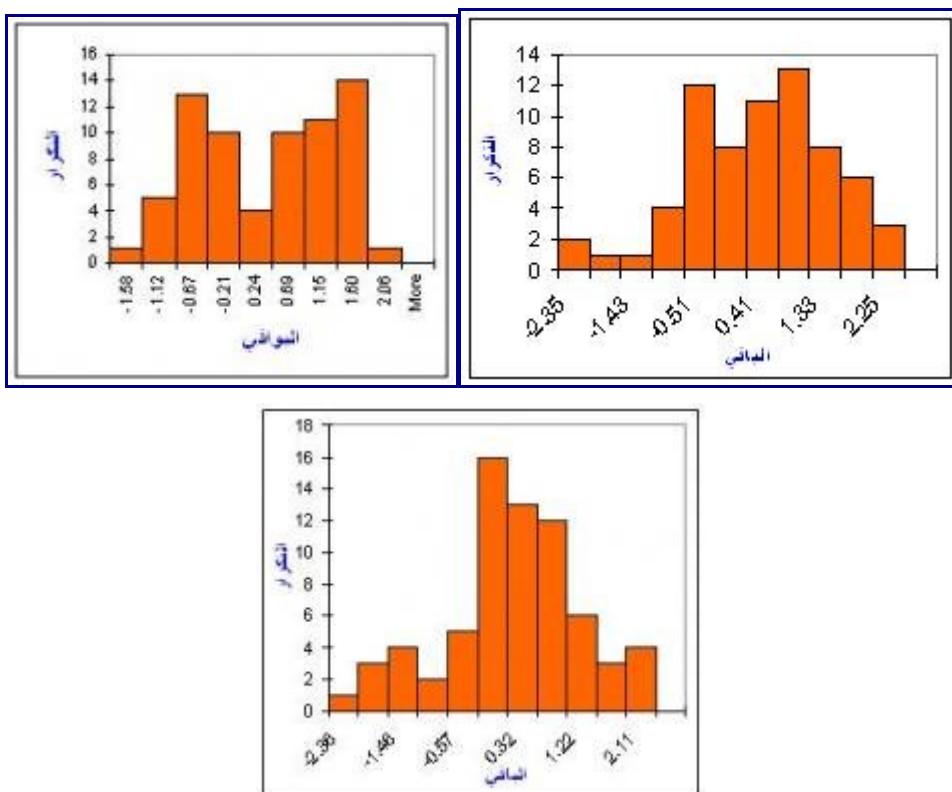


إن هذا الشكل يظهر تغير دوري شبه متكرر للباقي. إذن فالباقي غير مستقلة بل لها اتجاه محدد. في هذه الحالة مثلاً يبدو أن هناك تغير موسمي seasonal في المتغير الذي نقيسه مع الزمن. فمثلاً لو افترضنا أن هذا المتغير هو درجة حرارة سائل ما فإنه من الواضح تأثير الليل والنهار على درجة حرارة هذا السائل. في هذه الحالة فإن استخدامنا لنموذج تحليل الانحدار الخطي ليس هو الحل السليم بل يمكننا استخدام أسلوب التنبؤ باستخدام نموذج موسمي أي نموذج يأخذ في اعتباره هذا التغير الدوري في قيمة المتغير.



ماذا تلاحظ في هذا الشكل (أعلاه)؟ إن هناك تزايداً في قيمة الباقي مع الوقت وبالتالي فهي غير مستقلة. معنى ذلك أن الخطأ في النموذج يتزايد مع مرور الوقت فهو عند القراءات الأولى صغير وعند القراءات الأخيرة يأخذ قيمة أكبر.

توزيع الباقي: هناك طريقتان تستخدمان للتأكد من توزيع الباقي توزيعاً طبيعياً. الأولى هو رسم التوزيع التكراري Histogram. فإذا كانت الباقي تبع التوزيع الطبيعي فإن الفرض يكون قد تحقق. أما الثاني فهو منحنى الاحتمال الطبيعي Normal Probability Plot وهو أسلوب يستخدم للتحقق من أن مجموعة بيانات تبع التوزيع الطبيعي. وإن كانت البيانات تبع توزيعاً طبيعياً فإنها تأخذ شكل خط مستقيم تقريباً وإن كانت غير ذلك فإنها تأخذ اتجاهات مختلفة حول هذا الخط المستقيم. لا يهمنا الاستنتاج الرياضي لمنحنى الاحتمال الطبيعي ولكن يهمنا التعرف عليه واستخدامه.

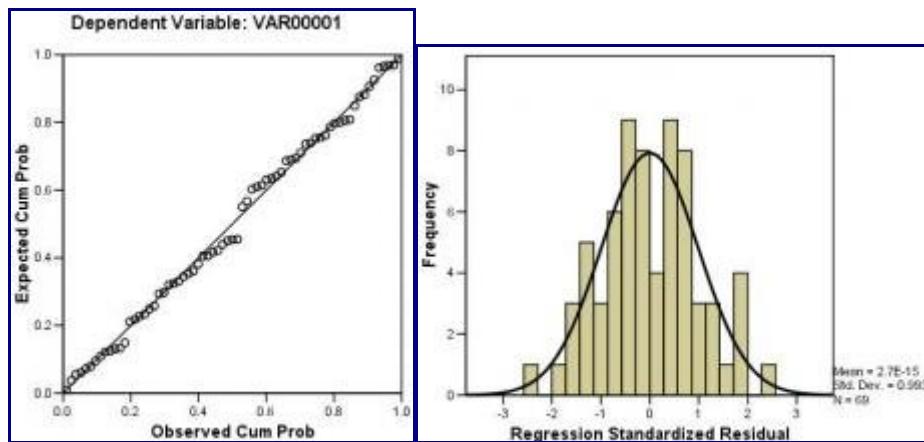


قد تجد أن التوزيع التكراري للأخطاء يشبه التوزيع الطبيعي مثل الأشكال الثلاثة أعلاه وقد تجده مختلفاً كثيراً. في

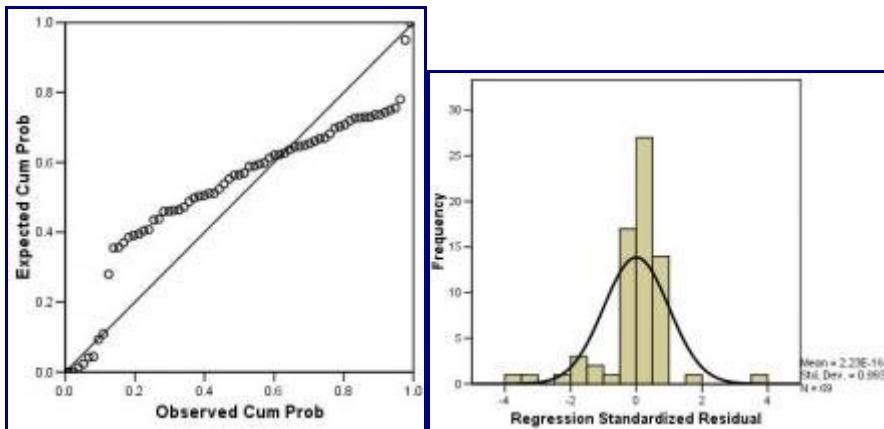
حالة أن التوزيع لا يتبع التوزيع الطبيعي بالمرة فإننا نبحث عن وسيلة أخرى غير تحليل الانحدار الخطي. الأمر المزعج هنا هو أن الحكم على الشكل يخضع للتقيير بشكل كبير وقد يختلف الرأي من شخص لآخر. بالإضافة لذلك فإن صغر حجم العينة قد يجعل الحكم على التوزيع الطبيعي أمراً غير دقيق.

التوزيع التكراري للبواقي لا يظهر مع النتيجة بشكل تلقائي في برنامج إكسل ولكن يمكن رسمه باستخدام Tools....Data Analysis...Histogram مثل Minitab, SPSS فإنك تحصل عليه مباشرة من النتيجة.

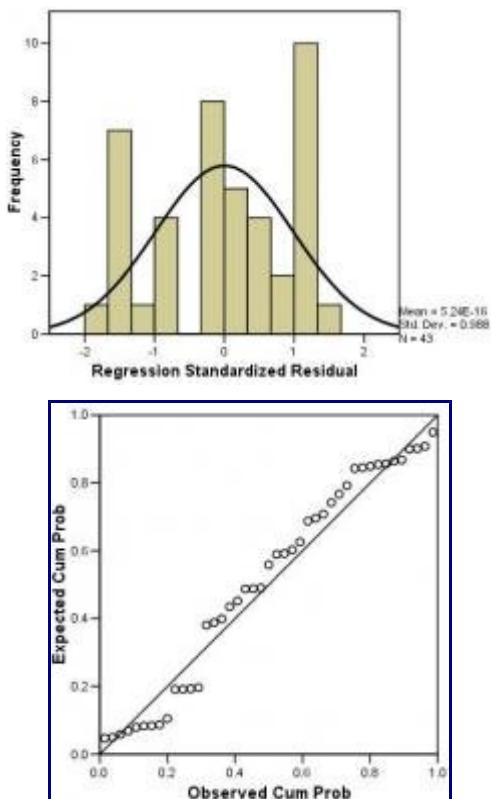
أما منحنى الاحتمال الطبيعي Normal Probability Plot فتحصل عليه من البرامج المتخصصة ولا تحصل عليه مباشرة من إكسل ويمكنك رسمه بإجراء بعض الحسابات. وهذا الرابط يقدم ملفاً لرسم هذا المنحنى: [ملف يشرح كيفية رسم منحنى الاحتمال الطبيعي](#). وإن شاء الله أتناول هذه التقاصيل في المستقبل. والذي يهمنا الآن هو نتائج هذا الاختبار.



لاحظ في الشكلين أعلاه كيف أن منحنى التوزيع التكراري يشبه إلى حد كبير منحنى التوزيع الطبيعي. والشكل على اليمين هو منحنى الاحتمال الطبيعي وتلاحظ فيه أن النقاط تتطابق كثيراً على الخط المائل والذي يمثل منحنى التوزيع الطبيعي. ولذلك فإن الشكل يبين أن البيانات والتي هي البواقي القياسية في هذه الحالة تتبع توزيعاً طبيعياً.



أما الشكلان أعلاه فيظهران اختلافاً كبيراً عن التوزيع الطبيعي. انظر كيف تبعد النقاط في الشكل على اليمين عن الخط المائل الذي يمثل التوزيع الطبيعي.



والشكلان أعلاه يبينان بعدا عن التوزيع الطبيعي كذلك. بهذا تستطيع أن تحكم على نتائج منحنى الاحتمال الطبيعي. موضوع دراسة البوachi ربما بدا معقدا بعض الشيء ولكن بالممارسة تعتاد عليه وتقهمه. وكما عرفنا فهو أمر مهم للتأكد من صحة استخدامنا لتحليل الانحدار.

[موقع ذات صلة بالموضوع:](#)

Regression Assumptions

Are the Model Residuals Well Behaved?

[من مراجع الموضوع:](#)

Discovering Statistics using SPSS for Windows, A. Field, Sage, 2003

Statistics for Managers, Levine et al., Prentice Hall, 1999

Sampling كيف تختار العينة؟

أكتوبر 24, 2009

عندما تقوم بإجراء استبيان فإنك تحدد من سيجيب عليه وهنا تواجه سؤالاً مهما وهو: هل ستسأل كل من له علاقة بالموضوع (مجتمع الدراسة) أم ستسأل بعضاً منهم (عينة)؟ فمثلاً إذا كان سنجري دراسة عن مميزات وعيوب الخدمة التي نقدمها فهل سنسأل كل المستهلكين أم بعضاً منهم؟ ربما يبدو سؤال كل المستهلكين كما لو كان الحل الدقيق والواجب ولكن الأمر ليس بهذه البساطة. هل تتصور صعوبة سؤال كل المستهلكين؟ هل تقدر الوقت والتكلفة اللازمين لسؤال كل المستهلكين؟ ما هو تأثير بقاء جمع المعلومات على قدرتنا على المنافسة؟ إن سؤال كل المستهلكين هو عملية صعبة تحتاج وقت طويل وتكلفة عالية وتجعل عملية تحليل البيانات أكثر صعوبة. وفي نفس الوقت فإننا إن سلنا عشر المستهلكين أو أقل مما يدرينا أن رأيهم يمثل رأي كل المستهلكين.



يبدو لنا من ذلك أن طرح الاستبيان على عينة محددة أمر سريع وأيسر من سؤال عدد هائل من الناس ولكن لابد من أن نبحث عن الطرق التي تجعل رأي العينة ممثلاً لرأي كل المستهلكين وإلا فإن البيانات التي سنحصل عليها ستقودنا إلى استنتاجات خاطئة.

وعملية أخذ العينات ليست مقتصرة على طرح الاستبيانات بل هي مستخدمة كذلك في أي عملية مسح عن Survey طريق المقابلات الشخصية أو المقابلات عن طريق التليفون وهي مستخدمة عند أخذ عينات من المنتج للتحليل أو الفحص وهي مستخدمة عند ملاحظة عينات من عملية ما لتقدير وقتها وعند قياس عدد الناس المنتظرين في الطابور في أوقات مختلفة (عينات من الوقت). فالتطبيقات متعددة جداً فمنها تطبيقات في مجال الصحة ومنها تطبيقات في مجال الصناعة ومنها تطبيقات في مجال التعليم ومنها تطبيقات في مجال السياسة ومنها تطبيقات في مجال التسويق وهكذا. فالكثير من وسائل الإعلام الأجنبية تقوم بعمل اقتراع لمعرفة رأي السكان أو المشاهدين أو القراء ومن الطبيعي أن بعض القراء أو المشاهدين أو السكان – وليس كلهم – سيشاركون في الاقتراع ومع ذلك فإن نتيجة الاقتراع تعتبر معتبرة عن رأي المجتمع كله. وإذا أرادت هيئة معرفة العادات الصحية لسكان بلد ما فإنك تسأل عينة من الناس وتعتبر أنها تمثل المجتمع كله.

وأحب توضيح بعض المصطلحات المستخدمة في هذه المقالة. فالمجتمع أو مجتمع الدراسة يقصد الأشخاص أو الأشياء التي ندرسها مثل السكان أو العملاء أو المرضى أو الطلبة أو المنتجات أو المواد الخام أو البهائم أو الدواجن وهذا. وهذه الأشياء تسمى مفردات المجتمع وقد تستخدم أحياناً الأفراد بدلاً من مفردات لأن الكثير من هذه الدراسات تتم على البشر.

عينات احتمالية (عشوانية) وعينات غير احتمالية (غير عشوائية):

هناك نوعان رئيسيان من العينات. النوع الأول هو العينات الاحتمالية (العشوانية) والتي تعتمد على وجود فرصة

علوم (احتمال يمكن حسابه) لكل فرد من مجتمع الدراسة لكي يتم اختياره في العينة. أي أن عملية اختيار العينة تتبع أسلوب عشوائي بطرق مختلفة. أما العينات غير الاحتمالية فهي عينة يتم اختيارها بطرق غير عشوائية ولا يمكن تحديد احتمالية اختيار كل فرد من مجتمع الدراسة. ففي العينة غير الاحتمالية يتم اختيار العينة بناء على قواعد محددة مثل سهولة الوصول للأفراد. اختيار عينة غير احتمالية يحتاج مجهودا أقل في اختيار العينة وقد يساعد على تخفيض الكلفة والوقت في الاتصال بأفراد العينة لجمع المعلومات.

فيمكنك تصور اختيار العينة الاحتمالية (العشوائية) كما لو كنا سنجري قرعة ونختار من تخرج أسماؤهم في القرعة. أما العينة غير الاحتمالية فيمكنك أن تتصور أننا نختار أفراداً محددين ليشكلوا العينة. وهناك أنواع من العينات العشوائية وهناك أنواع من العينات غير الاحتمالية وسوف نستعرضهم بمشيئة الله في هذه المقالة.

العينات الاحتمالية (العشوائية) أفضل من ناحية اختيار عينة معايرة عن مجتمع الدراسة فطريقة الاختيار ليس فيها تعمد لاختيار أفراداً بعينهم أو أجزاء بعينها. والعينة الاحتمالية تمكنا من حساب نسبة التغير (الاختلاف) المتوقع بين القيم التي حصلنا عليها من العينة وبين تلك الحقيقة لمجتمع الدراسة. أما في حالة العينات غير الاحتمالية فلا يمكننا أن نستخدم أي أساليب إحصائية لتقدير قيمة الخطأ أو نسبة. بالإضافة لذلك فإن هناك احتمالات لوجود تحيز عند اختيار عينة غير احتمالية وبالتالي هناك شك في أن العينة تمثل المجتمع.

قد تتصور أنه لا بدile عن استخدام العينات الاحتمالية. ولكن في الواقع فإن العينات غير الاحتمالية تستخدم كثيراً. نعم العينات الاحتمالية تعطي نتائج أدق ولكن في كثير من الأحيان يكون من الصعب اختيار عينة عشوائية. قد لا تسمح لك الميزانية أو طبيعة المكان بالوصول لأي فرد من مجتمع الدراسة فتقرر اختيار عينة من أماكن محددة، قد تكون الدراسة مبنية على فكرة عن الموضوع ثم يستتبعها دراسة شاملة فتقرر الإبقاء بعينة غير احتمالية في الدراسة المبنية، قد يكون ضيق الوقت سبباً في اختيار عينة غير احتمالية وهذا. وفي بعض الأحيان قد تكون العينة غير العشوائية أفضل من العينة العشوائية كما سنبين في العينات الاجتهادية.

أنواع العينات العشوائية (الاحتمالية):

1- العينة العشوائية البسيطة Simple Random Sample

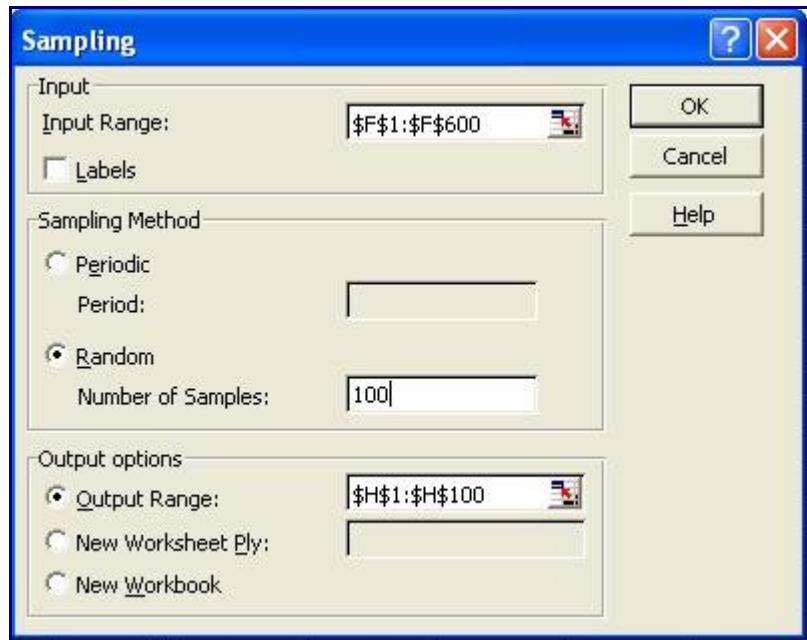
يتم اختيار العينة العشوائية البسيطة بطريق بسيطة وهي القرعة. فمثلاً إذا أردنا اختيار عينة من طلبة جامعة ما فإننا نكتب رقم كل طالب أو اسمه في ورقة ثم نضع الأوراق في إناء كروي ويتم تقليل الأوراق داخل الدورق ثمبدأ عملية سحب عشوائي. في هذه الحالة فإن كل طالب يتم سحب رقمه يكون أحد أفراد العينة ولا يمكن تغيير طالب مكان آخر أو إهمال أي طالب.

وقد أصبح الأمر أيس من ذلك في زمننا هذا حيث يمكننا اختيار العينة العشوائية البسيطة باستخدام الحاسوب. فمثلاً يمكن أن نستخدم برنامج إكسل لتخليق عدد من الأرقام العشوائية بين رقمين. فلو كانت أرقام الطلبة تتراوح بين 1000 و 1600 فإننا نكتب في أي خلية

$=RANDBETWEEN(1000,1600)$

وبنسخ هذه الخلية في 100 خلية مثلاً نحصل على 100 رقم بين 1000 و 1600. وتكون هذه الأرقام هي عينة عشوائية بسيطة من أرقام الطلبة أي من الطلبة (مجتمع الدراسة).

وهناك طريقة أخرى وهي أن تكتب الأرقام التي ستختار منها في عمود ثم تضغط على Tools ثم Data Analysis ثم Sampling.

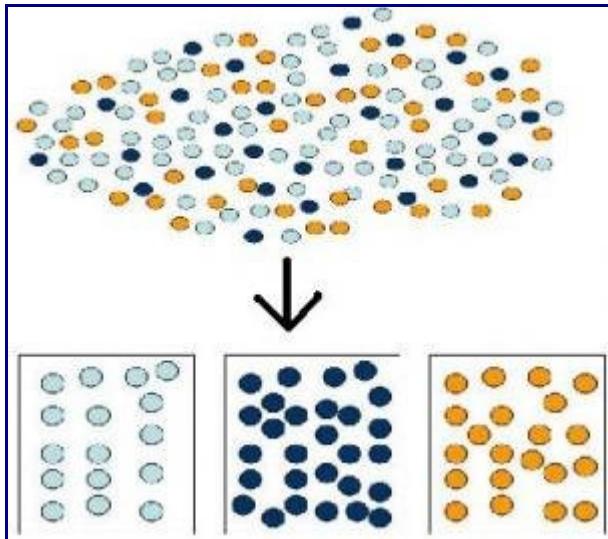


لاحظ أن Input Range هو الخلايا التي كتبت فيها الأرقام وهي في المثال الذي استخدمته F1 إلى F600. وهناك اختياراتان هما Random و Periodic. في هذه الحالة نختار Random أي اختيار عشوائي. ويتم كتابة حجم العينة في خانة Number of Samples. وأما Output Range فهو الخلايا التي تريد أن يكتب فيها إكسل الأرقام التي اختارها وهي في هذا المثال H1 إلى H100. وبهذا نحصل على عينة عشوائية بسيطة من الطلبة.

العينة العشوائية البسطة هي عينة خالية من التحيز ولكنها لا تخلو من المشاكل. قد تفاجأ بأن بعض من تم اختيارهم يصعب أن يجيبوا على الاستبيان أو يصعب عليك الذهاب لهم لإجراء مقابلة شخصية. قد يكون مجتمع الدراسة مكوناً من مجموعات لها سمات مميزة وقد تجد أن العينة العشوائية البسطة لم تحتوي على عدد كافٍ من بعض هذه المجموعات وبالتالي فلا يمكنك تحليل آراء أو بيانات كل مجموعة ومقارنتها بالأخرى.

2. العينة الطبقية Stratified Sample

في هذه الحالة يتم تقسيم مجتمع الدراسة (البحث) إلى مجموعات غير متداخلة ثم يتم اختيار عينة عشوائية بسيطة من كل مجموعة. فمثلاً لو كنا ندرس طلبة الجامعة فقد نقسمهم إلى تخصصات مختلفة ولو كنا ندرس المرضى فقد نقسمهم إلى نوعيات مختلفة من المرض ولو كنا ندرس العمال فقد نقسمهم حسب حجم تعاملهم معنا أو إلى رجال ونساء أو إلى عائلات وأفراد وهكذا. بعد ذلك نختار عينة عشوائية بسيطة من كل مجموعة.



ولكن هناك عدة خيارات في الحجم النسبي للعينات فقد نجعل حجم العينات يتناسب مع حجم كل مجموعة وقد نجعل حجم العينات متساوٍ بغض النظر عن حجم المجموعات. وقد يصل الأمر أن تأخذ عينات لا يتناسب حجمها مع حجم المجموعة التي أخذت منها وذلك لوجود تباين كبير داخل المجموعة. فمثلاً قد يكون مجتمع الدراسة هو ألف طالب وهو لاء الطلبة ينقسمون إلى طلبة محليين (600 طالب) وأجانب (400 طالب). ونحن نعلم أن آراء ومتطلبات الطلبة الأجانب تتتنوع كثيراً بتنوع بلادهم التي نشأوا فيها. لذلك فإننا قد نأخذ عينة أكبر من الطلبة الأجانب لكي تكون عينة مماثلة فعلاً لهذه المجموعة، وأما بالنسبة للطلبة المحليين فربما كانت مجموعة أصغر كافية لوجود تجانس في أفكارهم وآرائهم إلى حد ما.

بهذه الطريقة نستطيع تحليل نتائج كل مجموعة وأن نقول هذه المجموعة تقضي كذا وهذه تقضي كذا أو هذه المجموعة تميز بـكذا وهذه تميز بـكذا. ويبقى أن نقوم بتجميع ذلك لنعبر عن مجتمع الدراسة كله. يستخدم في ذلك المتوسط الحسابي المرجح (الموزون) Weighted Average. فمثلاً لو كان لدينا ثلاثة مجموعات من العملاء وقمنا بقياس رضا كل مجموعة عن الخدمة التي نقدمها ونريد تحديد رضا العملاء كلهم عن الخدمة. افترض أن المجموعات عددها هو 300، 500، 200 وأن مستوى الرضا عن الخدمة هو 3، 4، 3.6 على التوالي. علينا أن نحسب الوزن النسبي لكل مجموعة بقسمة حجم المجموعة على حجم المجتمع كله كالتالي:

$$\text{الوزن النسبي للمجموعة الأولى} = \frac{300}{1000} = 0.3$$

$$\text{الوزن النسبي للمجموعة الأولى} = \frac{500}{1000} = 0.5$$

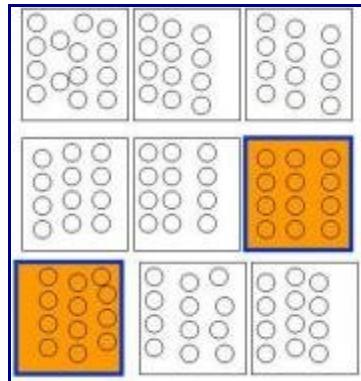
$$\text{الوزن النسبي للمجموعة الأولى} = \frac{200}{1000} = 0.2$$

والآن نحسب المؤشر العام لرضا العملاء عن الخدمة بضرب نتيجة كل مجموعة في وزنها النسبي

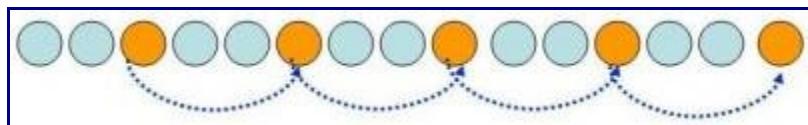
$$\text{المؤشر العام للرضا عن الخدمة} = 4.7 = 3.6 * 0.3 + 4 * 0.5 + 0.3 * 0.2$$

3. العينة العنقودية Cluster Sample

عندما يكون مجتمع الدراسة كبيراً وموزعاً بين مناطق متعددة بحيث يصعب الوصول إليها كلها فإنه يتم اختيار عينة من المناطق بشكل عشوائي وتعتبر العينة مكونة من كل مفردات المناطق المختارة. وفي حالة عدم القدرة على الحصول على آراء أو بيانات كل العينة فنحاول على الأقل أن نحصل على آراء أو بيانات معظمها. بهذا الأسلوب تكون قد قلنا التكلفة والوقت اللازمين لعملية المسح سواءً كانت عن طريق استبيان أو مقابلات شخصية. ولكن المخاطرة تكمن في أن بعض المناطق قد لا تكون معبأة عن مناطق أخرى.



4. العينة النظامية Systematic Sample هذه الطريقة هي طريقة عملية لاختيار عينة شبيهة بالعينة العشوائية البسيطة. في هذه الحالة يتم اختيار العينة بنظام محدد فمثلا لو أردنا أن نختار عينة مكونة من 100 موظف من أصل 1000 موظف فإننا نختار موظف من كل عشرة بنظام ثابت أي أننا نختار الموظف رقم 10 ثم عشرين ثم ثلاثين وهكذا. ومثلا لو كنا نختبر منتجات مرتبة في المخزن أو في المصنع فإننا ببساطة نختار منتج من كل عدد ثابت منها مثل أن نختار أول منتج ثم السادس ثم الحادي عشر. وكذلك لو كنا نريد سؤال العمال عن خدمة ما فقد نختار أول عميل يدخل ثم الحادي عشر ثم العشرين وهكذا. وإذا كنا نريد قياس طول طابور العملاء المنتظرين فإننا قد نقيس طوله كل فترة ثابتة مثل عشرين دقيقة فنقيس الساعة التاسعة وعشرين دقيقة ثم التاسعة وأربعين دقيقة ثم العاشرة ثم العاشرة وعشرين دقيقة وهكذا.



كما ترى فهي طريقة عملية جدا ويمكن تطبيقها في بعض الأحيان بدون استخدام الكمبيوتر أو غيره كما في عملية اختبار المنتج النهائي أو سؤال العمال الزائرين لمركز الخدمة. وتتجدر الإشارة إلى أن اختيار نقطة البداية هي عملية اختيارية فقد نختار أحد الموظفين في أول عشرة موظفين مثل أن نختار الموظف الثالث ثم الثالث عشر ثم الثالث والعشرين وهكذا على فرض أننا نختار موظف من كل عشرة موظفين.

وعلى الرغم من أن هذه الطريقة تشبه كثيرا العينة العشوائية البسيطة فإنه ينبغي التفكير في وجود تسلسل ما لمجتمع الدراسة فمثلا عندما نقيس طول الطابور كل عشرين دقيقة فإننا لن نقيس الطابور أبدا في منتصف الساعة أي الساعة العاشرة والنصف أو الحادية عشرة والنصف. فإن كانت هناك فترات ازدحام قصيرة تحدث في تلك الأوقات فلن نستطيع الإحساس بها من خلال هذه العينة. ومثلا لو كان لدينا منتج مخزن في أكواخ أو صناديق وقررنا اختبار منتج من كل سادس صندوق أي السادس ثم الحادي عشر ثم الثامن عشر فقد لا تكون العملية عشوائية لو كان ترتيب الصناديق يتبع أسلوبا محددا مثل أن يكون كل عشرة صناديق تمثل إنتاج يوم محدد وبالتالي فإننا نقيس جودة المنتج في منتصف اليوم فقط. فينبغي العناية بهذه النقطة عند استخدام العينة النظامية.

5. الاختيار متعدد المراحل Multistage Sampling قد تحتاج لاختيار عينة على مراحل متعددة. فمثلا قد نختار عينة عشوائية من كل محافظات مصر ثم عينة عشوائية من أحياء المحافظات التي تم اختيارها ثم عينة طبقية من الأحياء التي تم اختيارها أو عينة نظامية من بيوت تلك الأحياء. والسبب في تعدد المراحل هو الحاجة للوصول لعينة صغيرة نسبيا. وينبغي مراعاة تناسب كل طريقة اختيار لكل مرحلة.

أنواع العينات غير العشوائية: على الرغم من أفضلية العينات العشوائية فإنه في كثير من الأحيان يتم استخدام عينات غير عشوائية نتيجة لصعوبة أو تكلفة العينة العشوائية. نستعرض هنا بعض هذه الطرق.

1 عينة مريحة Convenience Sampling

هذه العينة تعني أن تختار عينة مريحة مثل أن تسأل بعض السكان من المناطق القريبة أو تسأل بعض الموظفين الذين تعرفهم أو وهكذا. هذه الطريقة تعتبر غير دقيقة ولكنها تستخدم في حالة الرغبة في اتخاذ قرارات سريعة وغير مهمة. فمثلاً قد تستخدم هذه الطريقة لمجرد اختبار الاستبيان قبل إرساله لمجموعة عشوائية وقد تستخدم لاستطلاع رأي مبدئي وهكذا.

وتعتبر هذه الطريقة مناسبة لو كان مجتمع الدراسة متشابهاً تماماً في ما يتعلق بالموضوع الذي ندرس له ويعتبر غير دقيق في حالة وجود اختلافات كبيرة. وكبسط للموضوع فإن اختبار جودة الطبخ لشيء متباين تماماً مثل الملوخية قد يصلح فيه عينة مريحة وأما اختبار جودة شيء غير متباين مثل شواء اللحم فإن العينة المريحة قد لا تكون معبرة بدقة.

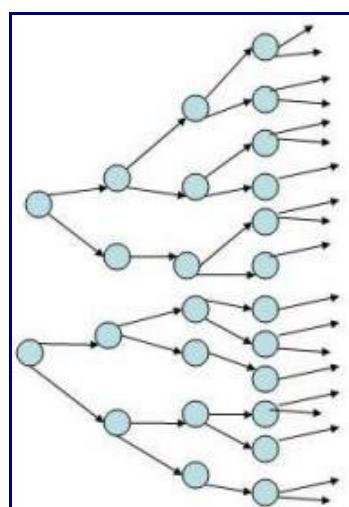
2- عينة اجتهادية :Judgmental Sampling

في هذه الطريقة يقوم شخص خبير بالموضوع وبمجتمع الدراسة بتحديد أسلوب اختيار العينة. فمثلاً قد يحدد مدناً بعينها لدراستها بدلاً من دراسة كل المدن أو اختيار عينة عشوائية. وفي هذه الحالة فإن هذا الشخص الخبير يختار مدناً تعبيراً فعلاً عن النوع الموجود في المدن كلها. وكذلك فإنه في حالة اختبار منتج فإن العينة الاجتهادية قد تستخدم لأن يتم اختيار عينات أكثر من المناطق التي يتحمل وجود العيوب بها أو من ظروف العمل التي تنتج عيوباً أكثر.

العينة الاجتهادية قد تحتمل بعض الانحياز فهي عينة غير احتمالية ولكنها في بعض الأحيان قد تكون أفضل من العينة العشوائية. ففي حالة اختيار بعض المدن عشوائياً فإننا بعض المدن ذات الصفات الخاصة قد لا يقع عليها الاختيار. وفي حالة فحص منتجات مصنعة فإن اختيار عينة عشوائية بسيطة سوف يجعلنا نفحص عدداً أقل من المنطقة التي نتوقع منها العيوب. فالأمر يتوقف على طبيعة الدراسة وجدية الاجتهداد في اختيار العينة.

3- عينة مرجعية (كرة الثلج) :Snowball Sampling

هذه العينة تستخدم حين لا يكون مجتمع الدراسة معلوماً لدينا على مستوى الأشخاص. فمثلاً لو كنا نريد أن نطرح استبياناً على المتخصصين في دراسة تأثير الاحتباس الحراري على سلوك الإنسان أو أردنا أن ندرس تأثير تناول الكحوليات على صحة الإنسان أو أردنا أن ندرس احتياجات الأطباء الذين يستخدمون أسلوباً محدداً في إجراء جراحة ما فإننا في هذه الحالات كلها لا يمكننا تحديد هؤلاء الأشخاص ومن ثم اختيار عينة عشوائية منهم. ماذا نفعل؟ إننا نحاول الوصول إلى واحد أو اثنين أو ثلاثة ثم نسألهم عن ما نريد ثم نطلب منهم ترشيح أشخاص لديهم نفس المواصفات المطلوبة. وبهذا فإن كل شخص نقابلة يرشح لنا شخصاً أو اثنين أو أكثر من تتطبق عليهم شروط الدراسة.



هذه الطريقة لا تخلو من الانحياز فهي عينة غير عشوائية ولكن استخدامها قد يكون هو الحل الوحيد في بعض الحالات مثل الأمثلة المذكورة أعلاه.

4 عينة حصصية :Quota Sampling

هذه العينة شبيهة جداً بالعينة الطبقية حيث يتم تقسيم المجتمع إلى عدة طبقات (شرائح) ثم يتم الاختيار من بين هذه الطبقات. ولكن الاختلاف هنا أن الاختيار من داخل الطبقات لا يتم بشكل عشوائي. وبهذا تكون العينة قد حافظت على المجموعات الموجودة في المجتمع وفي نفس الوقت فهي أبسط من العينة الطبقية في طريقة اختيار مفردات العينة. ولا يخفى عليك عيوب أنها عينة غير عشوائية.

كما ترى فهناك طرق مختلفة لاختيار العينات وعليك أن تتقى منها – بعنایة – ما يناسب الدراسة التي تقوم بها وطبيعة مجتمع الدراسة وقدراتك المادية والوقت المتاح للدراسة.

من مراجع الموضوع:

Essentials for Marketing Research, Kume, Akker and Day, Wiley, 2nd Edition, 2002

Statistics for Managers, Levine et al., second edition, Prentice Hall, 1999

sampling -ChangingMinds.org

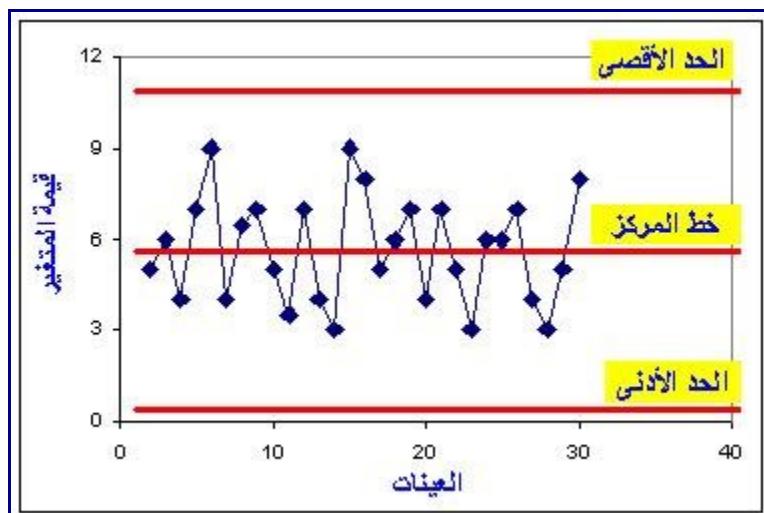
خرائط المراقبة Control Charts

يناير 22, 2010

ما هي خرائط المراقبة؟

خرائط المراقبة (الضبط) Control Charts هي وسيلة أساسية لضبط العمليات إحصائياً Control. فباستخدام خرائط المراقبة يمكننا متابعة سير العمليات واستخدام علم الإحصاء لمعرفة ما إذا كان هناك تغير غير طبيعي في العملية. فهي تمكننا من التدخل المبكر جداً لتصحيح العملية وتساعدنـا في تحديد سبب التغير. وهي وإن كانت مبنية على علم الإحصاء فإن استخدامها اليومي لا يحتاج لمتخصصين في الإحصاء بل هي وسيلة ينبغي أن يستخدمها عامل التشغيل نفسه.

افترض أنك مشرف إنتاج وتقوم بمتابعة العمل كل ساعة. وفي يوم من الأيام كانت نسبة العيوب في كل 100 منتج كالتالي: 3، 1، 5، 6، 7، 4. ما هو رد فعلك؟ ما هو الرقم الذي سيجعلك تتدخل للبحث عن السبب؟ هل مجرد زيادة النسبة من 1 إلى 5 يستدعي توقف الإنتاج حتى يتم تحديد سبب هذا الانهيار؟ ما هي مرجعية قرارك؟ هل 5 يعتبر رقم طبيعي أم لا؟ هل 7 يعتبر رقم مقبول؟ هل مستوى العملية قد تغير تغيراً ملحوظاً أم لا؟ ثم هل يعتبر رقم 1 إنجازاً أم لا؟ كيف ستحدد ذلك؟ في الحقيقة يصعب الإجابة عن هذه الأسئلة ولكن خرائط الضبط (المراقبة) تجيبنا عن ذلك.



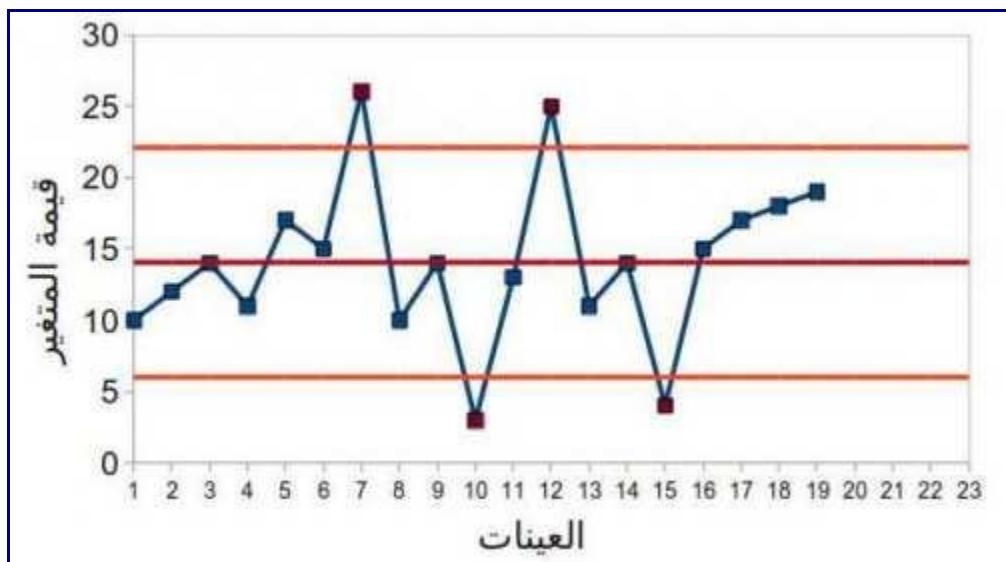
خريطة التحكم هي خريطة تبين لنا القيمة المتوسطة للمتغير الذي نتابعه وكذلك القيمة الدنيا والقصوى. فعندما نبدأ في استخدام خرائط التحكم فإننا نجمع بعض العينات ونسجل القيمة المتوسطة لكل عينة للمتغير الذي نقيسه مثل طول المنتج أو قطره أو درجة الحرارة. بعد ذلك نحسب القيمة المتوسطة وبذلك نرسم أول خط في خريطة التحكم والذي يمثل المتوسط. أما القيمة القصوى فيتم حسابهما بجمع المتوسط مع ثلاثة أمثل الانحراف المعياري. والقيمة الدنيا أو الحد الأدنى فيتم حسابه بطرح ثلاثة أمثل الانحراف المعياري من المتوسط. وسوف أبين كيفية رسم الخريطة بالتفصيل والتبسيط بمشيئة الله في مقالة تالية.

ما معنى القيمة القصوى والدنيا؟ افترض أن العملية التي نتابعها هي عملية مستقرة تنتج جودة مقبولة طبقاً للمواصفات المطلوبة. وافتراض أن التغير في قطر المنتج يتراوح بين 9.5 و 10.5. وافتراض أن هذا المدى ثابت يومياً ففي كل يوم يتراوح القطر بين هاتين القيميتين فمثلاً تكون النتائج: 10.1 - 10.2 - 10.0 - 9.9 - 10.5 - 9.6 - 10.3 - 9.5 - 9.8. فلو وجدنا في يوم ما أن القطر يساوى 10.6 أو 9.4 فإننا نستنتج فوراً أن تغيراً خارج الحدود

المعادة قد حدث في هذه العملية وبالتالي فإننا نوقف العمل ونبحث عن السبب ونحاول علاجه. هذا هو المقصود بالقيمة الدنيا والقصوى ببعض التبسيط.

في الواقع فإن التغير يختلف من يوم لآخر فلا نستطيع تحديد مدى التغير على وجه الدقة ولكن باستخدام الانحراف المعياري فإننا نرسم الحد الأدنى والأقصى اللذين يمثلان 95% من قيم المتغير. معنى أنه لو وقعت نقطة خارج هذه الحدود فإن ذلك يعني أننا متأكدين أن تغييرا غير عادي قد حدث ونسبة التأكيد هي 95%.

وقد تكون القيم داخل المدى المحدد ولكننا نفهم من الخريطة أن هناك تغيرا غير طبيعى وذلك بسبب اتخاذ القيم لأشكال محددة سنناقشها لاحقا إن شاء الله. وفي نفس الوقت فإن القيم قد تتغير حول المتوسط صعودا وهبوطا في ما بين الحدا الأدنى والأقصى ونكون مطمئنين إلا أنه لا يوجد تغير غير طبيعى أي أن التغير هو نفسه التغير الطبيعي للعملية. وهذا ما يظهر في الشكل السابق حيث أن هناك تغيرات كثيرة ولكنها في الحدود الطبيعية للعملية.



وبالتالي فإننا نوقع النتائج على خريطة التحكم بشكل دوري فنستطيع أن نحكم ما إذا كان التغير يعتبر تغيرا عاديا أو بسبب مؤثر خاص. لاحظ أننا هناك نحكم على أن العملية ما زالت مستمرة بنفس التغير ولا نحكم على أنها مناسبة. ففي بداية التنفيذ ينبغي أن تكون العملية تحقق الجودة المطلوبة وإلا فإننا سنحافظ عليها في الوضع الخاطئ. فلو نظرنا للشكل أعلاه للاحظنا تغيرات كثيرة في قيمة المتغير الذي نقيسه والذي قد يكون زمن الخدمة أو بعد من أبعاد المنتج وباستخدام هذه الطريقة ندرك بمجرد النظر أن العملية غير منضبطة إحصائيا فهناك تغيرات غير طبيعية مبينة بالنقطاط الملونة باللون الأحمر والتي خرجت عن الحد الأدنى أو الأقصى. لذلك فإن علينا أن نبحث عن أسباب التغير غير الطبيعي في هذه العينات. هل استخدمنا مادة خام مختلفة أم أن المشغل كان قليل الخبرة أم تم تغيير طريقة العمل أم حدث انهيار لجزء ما بالماكينة أم ...؟ أما التغير حول خط المنتصف وداخل الحدين الأقصى والأدنى فإنه أمر طبيعي ولا يستدعي أي تدخل.

فكما ترى فهي وسيلة مفيدة حيث أنها تعطينا تحذيرات مبكرة وتبيّن لنا ما إذا كان التغير مؤثرا أم لا. لذلك فإن خرائط المراقبة أو الضبط قد شاع استخدامها.

خلفية تاريخية:

تعزى نشأة خرائط المراقبة إلى ولتر شوهارت Walter Shewhart في العشرينيات من القرن الماضي حيث كان يعمل في شركة بل Bell Labs للاتصالات. وقد كان هناك حاجة لقليل العيوب في أجهزة الاتصالات التي تنتجها الشركة. وقد صاغ د. شوهارت التغير في صورة تغير طبيعي وتغير غير طبيعي (خاص) ثم اقترح خرائط التحكم

كوسيلة للتفريق بينهما ولمتابعة التغير والتدخل لإعادة العملية إلى طبيعتها. وقد ساهم [إدوارد دمنج Edward Deming](#) في نشر هذا الأسلوب في عدة شركات بالولايات المتحدة ثم بعد الحرب العالمية الثانية ولعدة عقود في اليابان التي تبنت أفكاره وطبقتها بكل جدية. ولذلك فإن الضبط الإحصائي للعمليات باستخدام خرائط التحكم هو أحد الأدوات التي تستخدم في نظام تويوتا الإنتاجي.

الاستخدامات:

لعلك استنتجت مما قرأت حتى الآن أن خرائط الضبط (المراقبة) هي وسيلة تستخدم في العمليات التصنيعية فقط. في الواقع إن هذه الوسيلة تستخدم في شتى المجالات فهي تستخدم في متابعة الأداء سواء في المصانع أو المؤسسات الخدمية. وهي تستخدم لمتابعة أداء العمليات اليومية وكذلك تحليل أداء المؤسسة. وهناك مؤسسات عربية تستخدم خرائط الضبط كجزء من العمل اليومي. فيمكن أن ترسم خرائط ضبط لأبعد المنتج أو لعدد الأخطاء أو لجودة المادة الخام أو لزمن العملية أو لدرجة رضا العميل عن الخدمة أو لحجم المبيعات أو وقت الانتظار أو عدد شكاوى العملاء أو سرعة الاستجابة أو غير ذلك.

بعد أن تعرفنا على هذه الخرائط المفيدة فإننا نستكمل الرحلة في المقالات التالية إن شاء الله ونتعرف على أنواع هذه الخرائط وكيفية بنائها وكيفية قرائتها.

من مراجع المقالة:

[مفهوم خريطة المراقبة – د. محمد عيشوني](#)

Operations Management, Russel & Taylor, 3rd Edition, Prentice Hall, 2000

Competitive Manufacturing Management, Nicholas, 1998, Irwin/McGraw-Hill

Lean Six Sigma Pocket ToolBook, George et al., McGraw Hill, 2005

موقع ذات صلة:

[استعمال برنامج إكسيل في مجال ضبط الجودة](#)

[Statistical Process Control in Health Service](#)

[Control Charts- Wikipedia](#)

المدرج التكراري Histogram

يناير 25, 2010

المدرج التكراري Histogram هو من الأدوات الشهيرة في تحليل البيانات لبساطته وتوضيحه لتوزيع البيانات. والكثير من التحاليل الإحصائية تبدأ برسم المدرج التكراري لمعرفة توافق توزيع البيانات الحقيقي مع بعض التوزيعات المعروفة مثل التوزيع الطبيعي Normal Distribution. ولهذا الأمر ارتباط بخراطط المراقبة لذلك فضلت أن أخصص هذه المقالة لمناقشة المدرج التكراري ثم المقالة التالية لمناقشة منحنى التوزيع الطبيعي ثم نستكمل الرحلة بمشيئة الله مع خرائط المراقبة (الضبط).

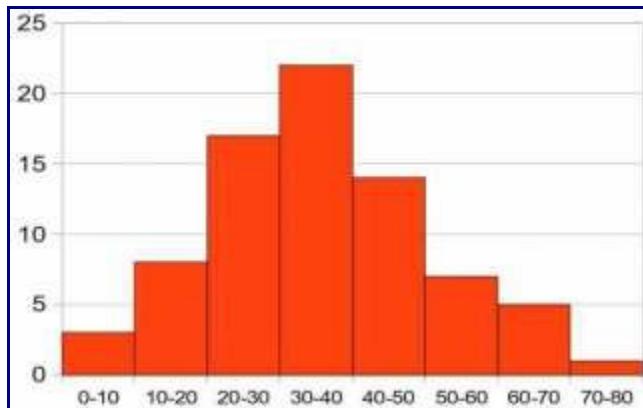
المدرج التكراري

المدرج التكراري هو أحد الرسومات البيانية التي تعطي معلومات غزيرة في شكل بسيط. فهو يمكنك من فهم البيانات وتوزيعها وبالتالي يمكنك من تحليل البيانات والوصول إلى قرارات إدارية مهمة. دعنا نستعرض مثلاً يوضح الأمر.

افرض أننا سجلنا أعمار مجموعة من الناس خرجوا في رحلة جماعية وكان عددهم 40 شخصاً. وبعد جمع البيانات أحبينا أن نعرف عدد الناس الذين سنهم أقل من 10 سنوات وهم لا يزيدون عن 10 سنين وبين 10 و 20 عاماً ثم بين 20 و 30 وهكذا. نقوم بوضع البيانات في جدول كالتالي حيث يمثل العمود الأيسر الشريحة العمرية والعمود الأيمن يمثل عدد الناس في كل شريحة:

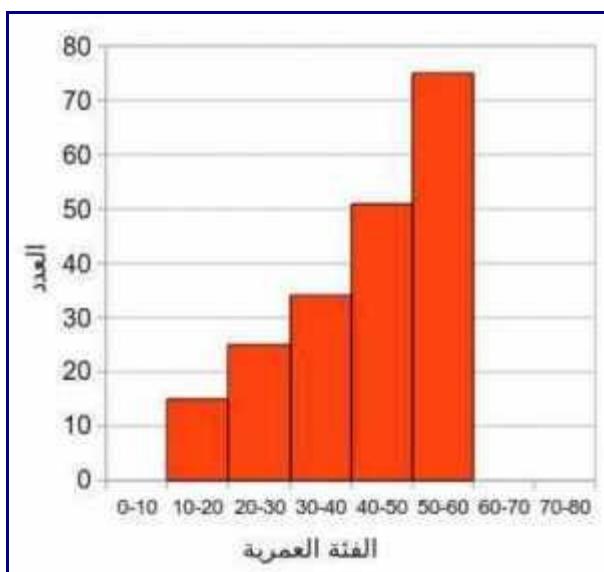
Age	Number of People
0-10	3
10-20	8
20-30	17
30-40	22
40-50	14
50-60	7
60-70	5
70-80	1

وبعد ذلك يمكننا رسم هذا الجدول في رسم هو ما يسمى بالمدرج التكراري. كل عمود من هذه الأعمدة يبين عدد الناس الذين يقعون في هذه الشريحة العمرية. بنظرية سريعة يمكنك أن تدرك أن معظم هذه المجموعة من الفئة العمرية المتوسطة أي بين العشرين والخمسين. ومن الملاحظ أن هناك قلة متساوية تقريباً من الفئات العمرية الصغيرة والكبيرة. ومن الواضح أن أكبر فئة عمرية هي بين الثلاثين والأربعين. ولاشك أن هذه معلومات مهمة نحصل عليها من الشكل بسرعة وسهولة.



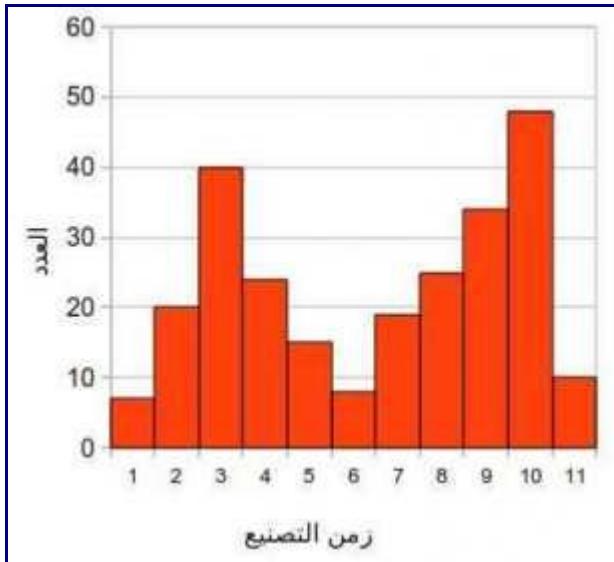
فالشكل أعلاه يشبه المدرج أو السلم ومن هنا سمي بالمدرج. وهو يبين تكرار قيمة ما بين مجموعة البيانات فهو في هذا المثال يبين تكرار كل فئة عمرية بين مشتركي الرحلة ولذلك سمي بالتكاري.

افرض أننا قمنا بنفس التحليل ولكن لعينة من 200 فرد من العاملين في مؤسسة ما. انظر إلى الشكل التالي.



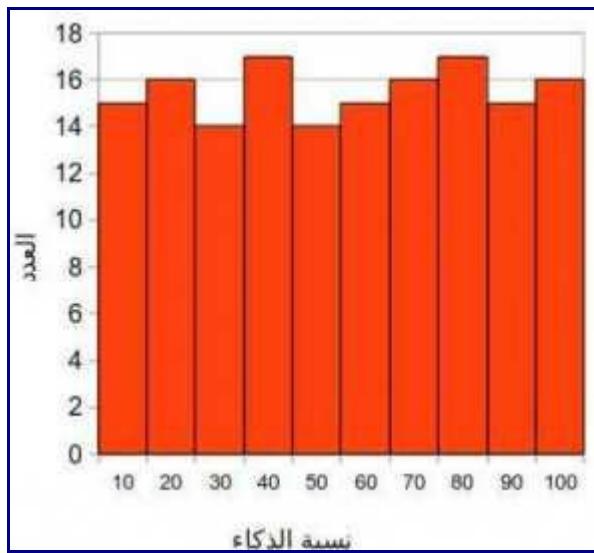
هل يمكنك التعليق على ذلك؟ بالطبع لا يوجد أطفال ولا يوجد من تعدوا الستين. هل ترى مشكلة لدى هذه المؤسسة؟ هل المستقبل يبدو مشرقاً لهذه المؤسسة؟ لابد أنك لاحظت مشكلة في التوزيع العمري للعاملين فأكبر شريحة هي الشريحة بين الخمسين والستين أي الشريحة التي من المفترض أن تتراوح في السنوات المقبلة. من الواضح ضعف التعويض من الشباب وبالتالي فهناك خطورة في ضعف أداء العاملين فجأة نتيجة لقاء أكثر من 30% من العاملين خلال السنوات العشر المقبلة. معلومات إدارية مفيدة استنتجناها بمجرد النظر لهذا الرسم.

انظر للمدرج التكاري التالي والذي يبين نتيجة قياس زمن تصنيع منتج ما داخل نفس الشركة:



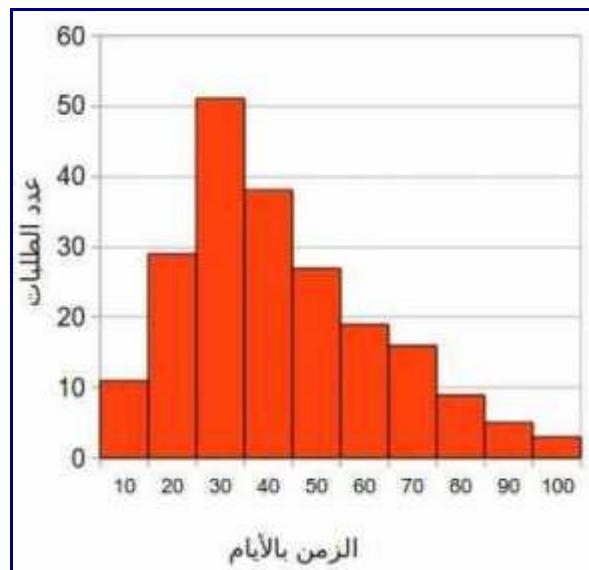
هل تلاحظ شيئاً؟ إن هناك منطقتان تتركز فيهما معظم القراءات. فيبدو كأن زمن التصنيع في الغالب حوالي 3 دقائق أو حوالي 9 أو 10 دقائق. أليس هذا غريباً؟ بل إنه لغريب فالطبيعي أن تكون هناك قيمة متوسطة في المنتصف تقريباً فمثلاً يكون الزمن الغالب هو 5 دقائق وأحياناً يقل أو يزيد عن 5.5. فما معنى ظهور الشكل هكذا؟ إن هذا يبين أن هناك حالتين لهذا المنتج وفي إحداهما نستغرق ما يقترب من 3 دقائق وفي الأخرى نحتاج حوالي 10 دقائق. فما هما الحالتان؟ هذا سؤال علينا كمدربين أو محللين أن نبحث عنه. ربما يكون هناك ماكيينات للإنتاج إحداهما جيدة والأخرى سيئة أو أن بعض العمال يستخدم طريقة محددة في العمل والبعض الآخر يستخدم طريقة أخرى أو أننا نستقبل نوعين من المواد الخام فإحداهما سهلة التصنيع والأخرى صعبة التصنيع. أمور مهمة جداً بدأنا نفكر فيها ونبحث عنها بمجرد النظر للدرج التكراري.

انظر للمثال التالي الذي يبين نسبة ذكاء مجموعة من الأطفال.



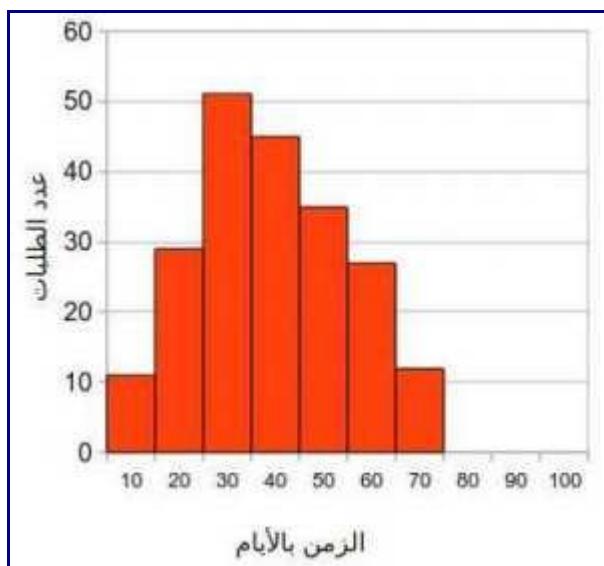
ما هو أوضح شيء في هذا المدرج التكراري؟ إن توزيع نسب الذكاء متقاربة من صفر إلى مائة بالمائة. ربما لا يبدو مقنعاً ولكن هذا ما نفهمه من هذا المثال الافتراضي. فلو صح ذلك فمعنى هذا أنه لا توجد نسبة ذكاء غالبة بل إن ذكاء هؤلاء الأطفال موزع بالتساوي طرقباً بين كل نسب الذكاء. بل يمكن أن تقول إن احتمالية أن يكون ذكاء أي طفل في هذا المجتمع 90% هي كاحتمالية أن يكون ذكاؤه 10% وهي كاحتمالية أن يكون ذكاؤه 40 أو 70%. فالدرج

التكاري يجعلنا نتصور احتمالية حدوث قيمة محددة.
انظر إلى المدرج التكاري الآتي الذي يبين زمن تنفيذ طلبات شراء المواد الخام أي الفترة الزمنية بالأيام من وقت الطلب لوقت وصول المواد. ماذا تلاحظ؟



هل رسم هذا الشكل يفيدنا كمديرين أو كمهندسين صناعيين حاول تطوير العمليات؟ إن أول ما تلاحظه أن الشكل ليس متماثلا وأن هناك طلبات شراء قليلة تستغرق وقتاً طويلاً جداً. في بينما تجد أن 10 طلبات يتم تنفيذها في أقل من 10 أيام فإن 3 طلبات يتم تنفيذها في حوالي 100 يوم. ومن الملاحظ أن أكثر الطلبات يتم تنفيذها في حوالي 20 إلى 50 يوماً. من هنا يمكننا فهم زمن التقدم Lead Time لطلبات الشراء.

قارن بين الشكل السابق والشكل التالي:



أيهما أفضل من الناحية الإدارية لمؤسسة ما؟ إن الفارق بينهما ضئيل ولكنه مهم. فعلى الرغم من أن معظم الطلبات يتم تنفيذها في حوالي 20 إلى 50 أو 60 يوماً في كلا الحالتين فإن الشكل الثاني أفضل لأن مدى التغير أقل. ففي الحالة الثانية لا يوجد طلبات شراء تستغرق أكثر من 70 يوماً بينما في الحالة الأولى هناك طلبات شراء يتم تنفيذها في مائة

يوم. فالدرج التكراري يساعدنا على مقارنة حالتين أو أكثر لنرى الصورة بوضوح. هناك استخدامات متعددة للدرج التكراري فيمكن استخدامه لرسم عدد الأعطال أو عدد عيوب الجودة في كل أسبوع أو كل شهر ويمكن استخدامه لرسم حجم المبيعات شهريا.

كيف نرسم الدرج التكراري؟

افترض أننا جمعنا مجموعة بيانات مثل أعمار الناس أو أطوالهم أو عدد العيوب في الإنتاج كل ساعة أو درجات الطلبة أو تقييم الموظفين أو غير ذلك. ونريد أن نرسم الدرج التكراري لكي نحل هذه البيانات. هناك طريقتان:

الطريقة الأولى:

1- حدد المدى أي الفرق بين أكبر قيمة وأقل قيمة. فمثلاً أقل درجة وأعلى درجة أو الفرق بين أقصر شخص في المجموعة وأطول شخص وهكذا حسب نوع البيانات التي نقيسها.

2- قسم هذا المدى إلى عدة أقسام. في المثال الأول قسمنا المدى إلى عدة أقسام تمثل كل منها شريحة قدرها عشر سنوات. عملية التقسيم تتطلب توازناً بحيث لا نقسم لأقسام صغيرة جداً فيصبح الدرج التكراري غير واضح خاصة في حالة صغر عدد الملاحظات أو صغر العينة، ولا نجعلها كثيرة جداً بحيث يكون الدرج التكراري غير واضح. فلو قسمنا الأعمار في المثال الأول إلى أقسام بحيث يمثل كل قسم عامين فقط فإننا سنجد أن هناك فراغات في الدرج التكراري نتيجة لأنه ليس هناك أحد في هذه المجموعة يقع سنه في هذه الشريحة. ولو قسمنا المدى إلى ثلاثة أقسام مثلاً فستنتهي بأحدة لا تساعدنا حقاً على فهم البيانات. وأتصور أن عدد الأعمدة يفضل أن يكون بين خمسة وعشرة للبيانات القليلة مثل درجات نجاح 50 طالباً، وربما يزيد إن كان لدينا بيانات كثيرة مثل درجات نجاح ألف طالب.

3- قم بحصر عدد القراءات التي تقع في كل شريحة. فمثلاً قم بحصر عدد الأفراد أقل من 10 سنين ثم عدد الأفراد بين 10 و 20 سنة وهكذا.

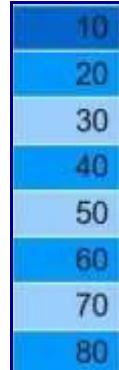
4- ضع نتيجة الحصر في جدول كما في أعلى المقالة.

5- ارسم رسم بياني يعبر عن هذا الجدول. يمكن رسم الجدول يدوياً أو باستخدام الحاسوب.

الطريقة الثانية:

الخطوات الأوليان لا يتغيران. فنبدأ بتحديد مدى القراءات ثم تقسيمها لأقسام. بعد ذلك نستخدم برنامج إكسيل أو كالك (في المكتب المفتوح). افترض أن البيانات هي لأعمار مجموعة من الناس.

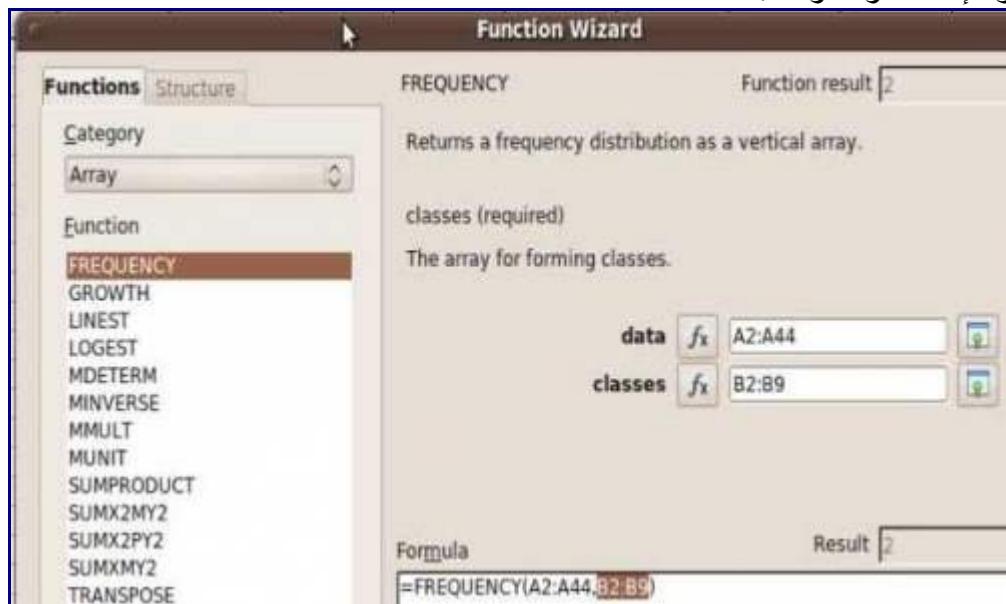
3- اكتب حد كل قسم في خلية كما بالشكل.



4- استخدم الدالة Frequency لكي يقوم الحاسوب بحصر بيانات كل شريحة بدلا من أن تفعل ذلك بنفسك. ظلل الخلايا المجاورة للخلايا التي سجلت فيها مدى كل شريحة كما بالشكل

	A	B	C
1	Age		
2	8	10	
3	9	20	
4	11	30	
5	12	40	
6	18	50	
7	19	60	
8	22	70	
9	23	80	
10	24		
11	26		
12	26		
13	28		
14	28		
15	32		
16	32		
17	32		

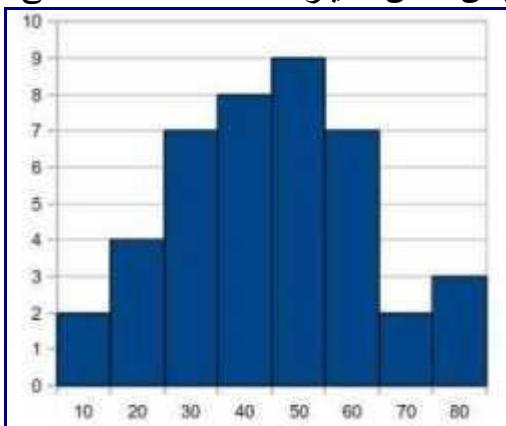
اختر دالة التكرار Frequency واملأ الخانات المطلوبة. الخانة الأولى هي الخلية التي تحتوي على البيانات وهي في مثالي هذا A2:A44. والخانة الثانية هي الخلية التي سجلت فيها الأقسام وهي نفس المثال الحالي B2:B9. ثم اضغط على زر الإدخال أو الموافقة.



يظهر لك عدد الناس في كل شريحة عمرية كالتالي:

10	2
20	4
30	7
40	8
50	9
60	7
70	2
80	3

5- استخدم البرنامج لرسم هذا الجدول مع تحديد أن العمود الأول هو مجرد أسماء الأعمدة Labels ومع تصغير المسافة بين الأعمدة إلى الصفر وهي من ضمن الخيارات المتاحة فتحصل على الشكل التالي



هذه الطريقة تعمل في إكسل واكالك. وتوجد طريقة أخرى في إكسل لا تختلف كثيراً وهي اختيار Tools ثم Data Analysis ثم Histogram Analysis

[من مراجع المقالة](#):

Lean Six Sigma Pocket ToolBook, George et al., McGraw Hill, 2005

Operations Management, Russel & Taylor, 3rd Edition, Prentice Hall, 2000

التوزيع الطبيعي وأهميته... Normal Distribution

فبراير 11, 2010

منحنى التوزيع الطبيعي Normal Distribution Curve هو من الأدوات كثيرة الاستخدام في التحاليل الإحصائية التي يحتاجها المدير والمهندس الصناعي. فدائماً ما تسمع عن المنحنى الذي يشبه الناقوس وهو منحنى التوزيع الطبيعي. ومن أشهر تطبيقاته الإدارية تقييم المسؤولين طبقاً لهذا المنحنى أي بحيث يحقق التقييم نفس شكل التوزيع الطبيعي لضمان قدر من العدالة. ولمنحنى التوزيع الطبيعي استخدامه في دراسة الباقي في تحليل الانحدار وله علاقة وطيدة بخراطط الضبط Control Charts. لذلك فضلت أن نُمعن النظر في منحنى التوزيع الطبيعي قبل أن نستقيض في خراطط الضبط (المراقبة). وإنني أحاول في هذه المقالة توضيح مفهوم منحنى التوزيع الطبيعي دون الدخول في تعقيدات حسابية.

ما معنى التوزيع الاحتمالي Probability Distribution

يمكن فهم التوزيع (التوزيع الاحتمالي) كشكل مشابه للمدرج التكراري Histogram ولكن المدرج التكراري يصف توزيع البيانات الحقيقية بينما التوزيعات الرياضية (النظرية) مثل التوزيع الطبيعي وغيرها هي توزيعات نظرية لها معدلات محددة وجداول تبين الاحتمالات المختلفة ولذلك تسمى توزيعات احتمالية. فعندما نرسم المدرج التكراري لمتغير ما فإننا نحاول أن نتعرف على التوزيع الاحتمالي الذي يُشبهه لكي نستخدم هذا التوزيع الاحتمالي في التحاليل الإحصائية.

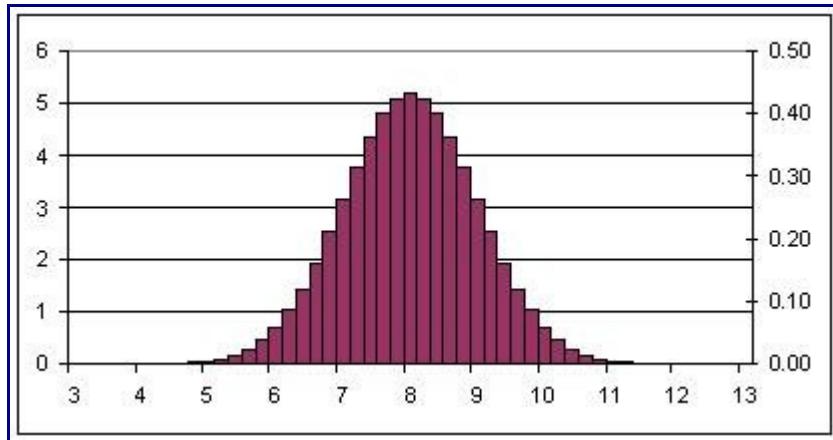
التوزيع يبين احتمالية أن يأخذ المتغير الذي ندرسه قيمة معينة أو أن يأخذ أقل أو أكثر من قيمة ما. فالتوزيع المنتظم Uniform يبين أن احتمالية أن يأخذ المتغير قيمة ما في مدى محدد متساوية بينما تجد الاحتمالات مختلفة في التوزيع الطبيعي. ففي التوزيع الطبيعي تكون الاحتمالية أعلى إذا كانت القيمة قريبة من المتوسط وتكون قليلة كلما ابتعدنا عن المتوسط. وهذه الاحتمالية يمكن تحديدها باستخدام الحاسوب أو الجداول.

افترض أنك تريد حساب محيط ومساحة منزلك. في البداية تقيس أبعاد الغرف ثم تقوم برسمها. بعد ذلك تبدأ في البحث عن أشكال هندسية تتشابه لأشكال الغرف مثل الشكل المستطيل أو المثلث أو شبه المنحرف أو المربع. وبعد تحديد الشكل الهندسي المشابه للغرفة تبدأ في حساب المحيط والمساحة باستخدام قوانين الهندسة الخاصة بكل شكل. هذا هو نفس الأمر بالنسبة للتغير متغير ما. إنك تقيس قيم هذا المتغير في فترة ما ثم تقوم برسمها كمدرج تكراري. بعد ذلك تبحث عن توزيع احتمالي يُشبه هذا المدرج التكراري. وبعد تحديد التوزيع الاحتمالي المناسب تبدأ في استخدام جداوله أو استخدام الحاسوب ل القيام ببعض التحاليل الخاصة بهذا المتغير.

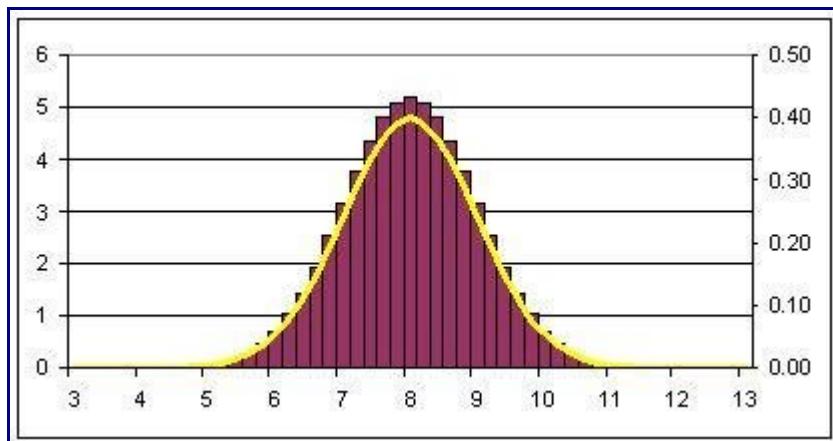
الكثير من التحاليل الإحصائية تعتمد على توزيع البيانات بنفس التوزيع الطبيعي ولذلك فإننا نرسم المدرج التكراري ونحاول مقارنته بمنحنى التوزيع الطبيعي. وهناك تطبيقات تفترض توزيع أسي Exponential Distribution مثل نظرية خطوط الانتظار (الطاوبيـر) أي أنها مبنية على افتراض أن زمن الخدمة يأخذ شكل التوزيع الأسي.

والتوزيعات الاحتمالية لها أهمية في عمليات المحاكاة Simulation حيث يقوم بتحديد أقرب توزيع احتمالي للمدرج التكراري أي للتغيرات الحقيقة. وبناء عليه فإننا نستخدم هذا التوزيع في نموذج المحاكاة حيث يتم محاكاة التغير بنفس التوزيع وبنفس القيم الحقيقة.

افتراض أننا قمنا برسم المدرج التكراري لمجموعة بيانات وحصلنا على الشكل التالي.



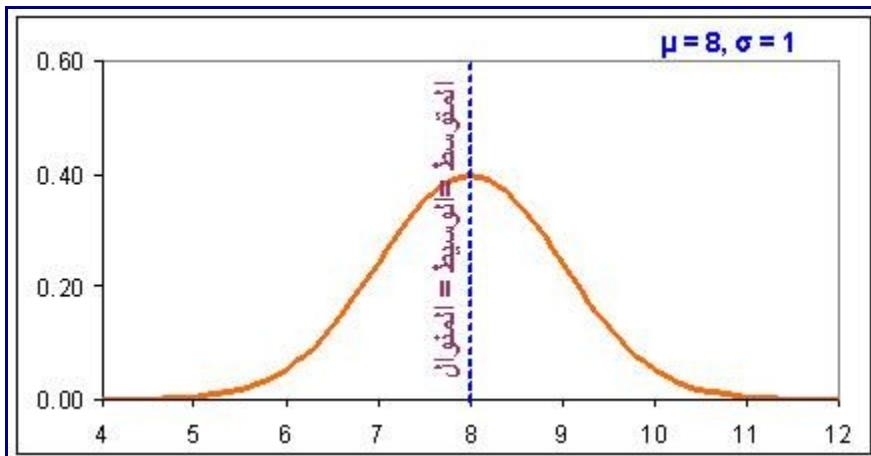
يمكنا البحث عن توزيع رياضي يشبه هذا المدرج التكراري والذي نرسمه بالخط الأصفر في الرسم التالي. في هذه الحالة فإن التوزيع المناسب هو التوزيع الطبيعي.



التوزيع الطبيعي؟

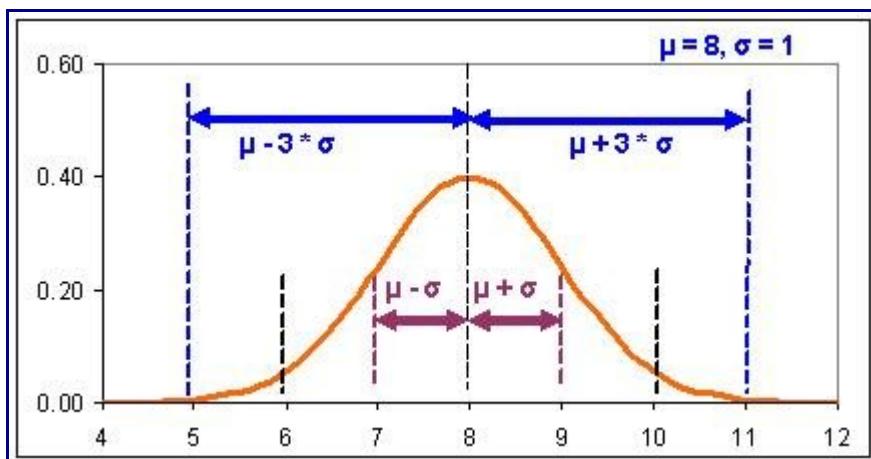
التوزيع الطبيعي Normal Distribution هو أشهر التوزيعات الاحتمالية وذلك لسبعين. السبب الأول هو أن الكثير من الظواهر تتبع منحنى التوزيع الطبيعي. السبب الآخر هو أن هناك نظرية تقول أن متوسط قيم عينات متعددة يأخذ شكل التوزيع الطبيعي ولو لم يكن توزيع المتغير نفسه يتبع التوزيع الطبيعي. لذلك فإن التوزيع الطبيعي هو شيء محوري في علم الإحصاء.

منحنى التوزيع الطبيعي يشبه الجرس (الناقوس) ويتميز بوجود تماثل بين جانبيه الأيمن والأيسر حول المتوسط. ومن سمات منحنى التوزيع الطبيعي أن المتوسط يساوي الوسيط ويساوي المنوال. يتم تعريف منحنى التوزيع الطبيعي بقيمتين: المتوسط والانحراف المعياري. ويرمز عادة للمتوسط بـ μ وللإنحراف المعياري بـ σ . الرسم التالي يبين شكل منحنى التوزيع الطبيعي وفي هذا المثال المتوسط $8 = \mu$. لاحظ أن تماثل المنحنى يعني أن 50% من القيم هي أقل من المتوسط و 50% من القيم هي أكبر من المتوسط وهذا يعني أن الوسيط يساوي المتوسط.



*** إذا لم تكن مصطلحات المتوسط والوسيط والمنوال والأنحراف المعياري مألوفة للقارئ الكريم برجاء الرجوع [للمقالات التاليتين: التعامل مع البيانات، تلخيص البيانات](#). وكتكررة سريعة فإن المتوسط هو مجموع القيم كلها مقسوما على عددها. والوسيط هو القيمة التي تكون 50% منها قيم أكبر منها. والمنوال هو القيمة الأكثر تكرارا. والأنحراف المعياري هو مقياس بعد جميع القيم عن المتوسط أي مقياس لتشتت القيم.

ولمحتى التوزيع الطبيعي سمات رئيسية منها أن 68% من الاحتمالات تقع في حدود المتوسط \pm الأنحراف المعياري. و99.7% من الاحتمالات تقع في حدود المتوسط $\pm 3\sigma$ * الأنحراف المعياري. فلو عرفنا المتوسط والأنحراف المعياري يمكننا حساب هذه الاحتمالات. لاحظ أن احتمال وقوع المتغير بين قيمتين تمثل بالمساحة تحت المنحنى بين هاتين القيمتين. ولذلك يمكننا بمجرد النظر أن نقول إن وقوع قيمة المتغير في الرسم أدناه بين 8 و 9 هي أعلى بكثير من وقوعه بين 10 و 11 لأن المساحة تحت المنحنى بين 8 و 9 أكبر بكثير منها بين 10 و 11.



وفي الشكل أعلاه يمكننا أن نقول أن قيمة هذا المتغير في 99.7% من الحالات تقع بين 5 و 11. وأن قيمة هذا المتغير تتراوح بين 7 و 9 في 68% من الحالات.

فعلى سبيل المثال لو وجدنا أن زمن التصنيع يتبع التوزيع الطبيعي بمتوسط 30 دقيقة وانحراف معياري 2 دقيقة فإنه يمكننا أن نقول أن 99.7% من الإنتاج يستغرق

$$30 \pm 2 \times 3 = 24 \text{ إلى } 36 \text{ دقيقة}$$

ولو وجدنا أن طول القطعة التي ننتجها يتبع التوزيع الطبيعي بمتوسط 10 م وانحراف معياري 0.01 مم فإنه يمكننا مقارنة ذلك بالمواصفات المطلوبة. فمثلا يمكننا أن نقول أن 99.7% من الإنتاج سيحقق طول =

$$= 0.01 * 3 \pm 10 \text{ مم} = 9.97 \text{ إلى } 10.03$$

فلو كانت الموصفات تسمح بأن يكون هذا البعد بين 9.96 و 10.04 مم فإننا نستنتج أننا في الجانب الآمن فيما يزيد عن 99.7% من الحالات. أما لو كانت الموصفات تشرط أن يكون هذا البعد بين 9.99 و 10.01 مم فإن المخاطرة ستكون كبيرة. فنحن نعلم أنه في 68% من الحالات يكون هذا الطول مساويا

$$= 0.01 * 1 \pm 10 \text{ مم} = 9.99 \text{ إلى } 10.01$$

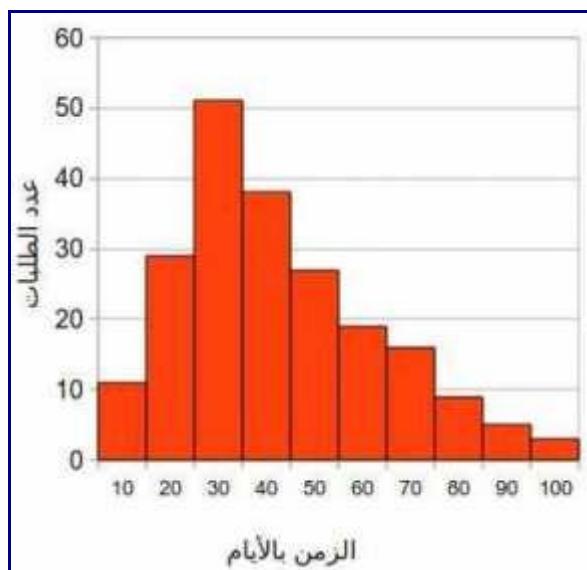
وبالتالي فإننا في هذه الحالة نتوقع أن حقق الموصفات في 68% من الكمية المنتجة أي أن 32% من المحتمل أن يتجاوز الموصفات المطلوبة. ومن هنا نفكر في عدم القيام بهذه العملية أو استخدام طريقة إنتاج أخرى.

ولا يتوقف الأمر عند هذا الحد بل يمكننا تحديد احتمالية تجاوز أي قيمة وذلك من خلال الجداول أو باستخدام الحاسوب.

والتوزيع الطبيعي هو جزء أساسي من فكرة خرائط المراقبة. فالحدود القصوى والدنيا توضع عند $\mu \pm 3\sigma$. لماذا؟ لأنه في حالة التوزيع الطبيعي فإن احتمالية وقوع القيم في هذا المدى هي 99.7% كما ذكرنا منذ قليل. أي أن القيمة لو كانت خارج هذا المدى فهي لا تتنبئ بنفس التوزيع أي أن شيئاً غير طبيعي قد حدث.

المساحة تحت المنحنى... لماذا؟

كما علمت فإن احتمالية وقوع المتغير بين قيمتين تقادس بالمساحة تحت المنحنى بين هاتين القيمتين. ولكن من أين لنا هذا المفهوم؟ دعنا نرجع إلى المدرج التكراري Histogram. انظر إلى المدرج التكراري أدناه والذي يبين زمن عملية ما بالأيام.

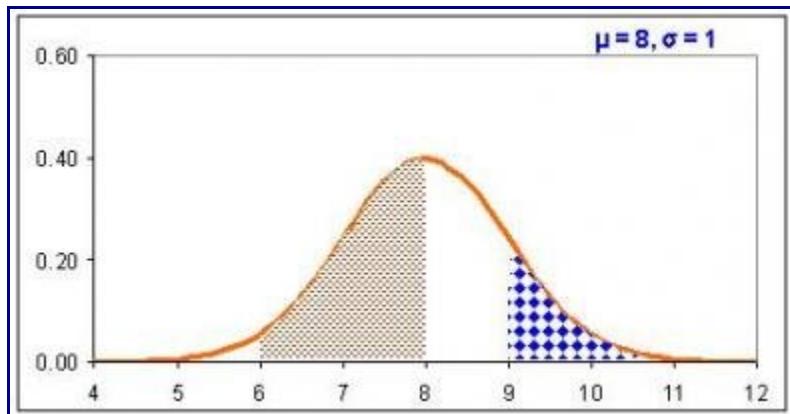


من الواضح أن الزمن متغير ولكن إن سألك ما هي احتمالية أن يكون زمن العملية بين 20 و 40 يوماً؟ كيف ستتقر في الأمر؟ إنك ستتظر إلى الأعمدة التي تبين وقوع المتغير في هذا المدى. من لواضح أنهما أطول عمودين وبالتالي لإغتن احتماليتهما كبيرة.

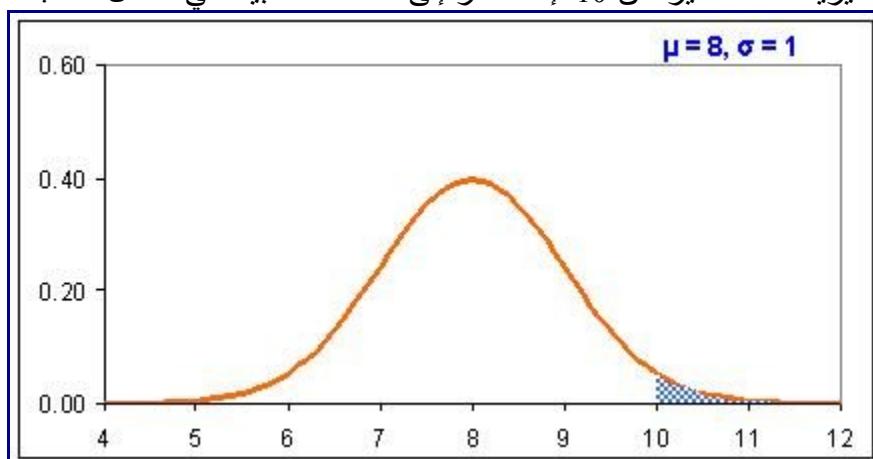
ماذا لو سألك عن المقارنة أن يكون الزمن من 90 إلى 100 يوم وبين أن يكون من 30 إلى 50 يوماً؟ إنك ستجيب بمنتهي الثقة بأن احتمالية أن يكون الزمن من 90 إلى 100 يوم أقل بكثير من احتمالية أن يكون من 30 إلى 50 يوماً. لماذا؟ لأنك وجدت أن العمود الذي يمثل وقوع المتغير من 90 إلى 100 مائة قصير جداً بالنسبة للعمودين اللذين يمثلان وقوع المتغير من 30 إلى 50 يوماً. فالواقع أنك تجمع طول الأعمدة وتقارنها لتحديد الاحتماليات. وطول

الأعمدة يتناسب تماما مع المساحة التي تمثلها هذه الأعمدة لأن المساحة هي حاصل ضرب هذه الأطوال في عرض كل عمود والذي هو ثابت يساوي عشرة في مثالنا هذا.

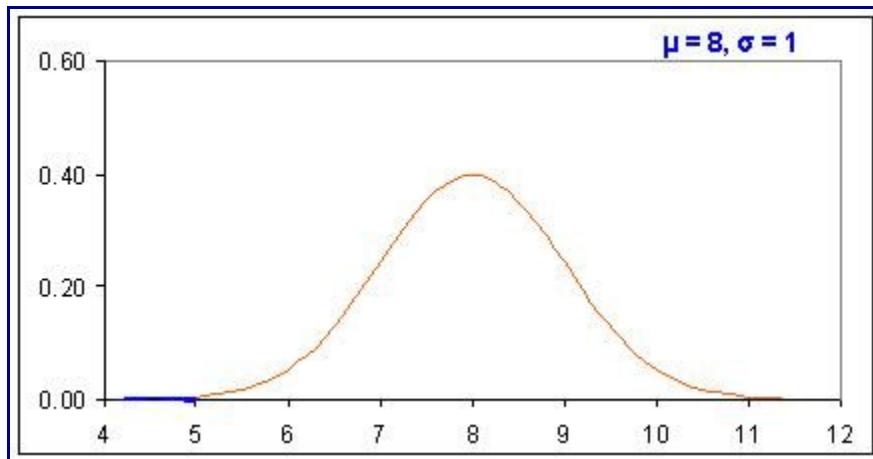
ولذلك فإننا عندما نستخدم توزيع احتمالي مثل التوزيع الطبيعي أو المنتظم أو الأسني وغيرهم فإننا نحدد الاحتماليات بالنظر للمساحة تحت المنحنى. فلو نظرنا للشكل أدناه لعلمنا أن وقوع هذا المتغير بين 6 و 8 (المساحة البنية اللون) هي أكبر بكثير من وقوعه بين 9 و 11 (المساحة الزرقاء اللون). فهي نفس فكرة النظر للأعمدة في المدرج التكراري.



ويمكننا بنفس الطريقة تقدير احتمالية أن يتجاوز المتغير قيمة ما أو يقل عنها. فمثلا لو أحبينا أن نعرف احتمالية أن يزيد هذا المتغير عن 10 فإننا ننظر إلى المساحة المبينة في الشكل أدناه.



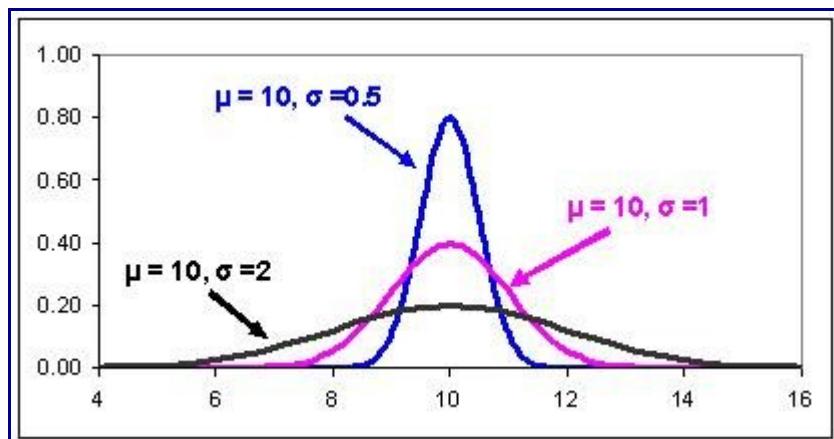
ولو أحبينا أن نعرف احتمالية أن يقل هذا المتغير عن 5 فإننا ننظر إلى المساحة تحت المنحنى من قيمة 5 فما أقل وهي مساحة صغيرة جدا تقترب من الصفر (المساحة الزرقاء في الشكل أدناه).



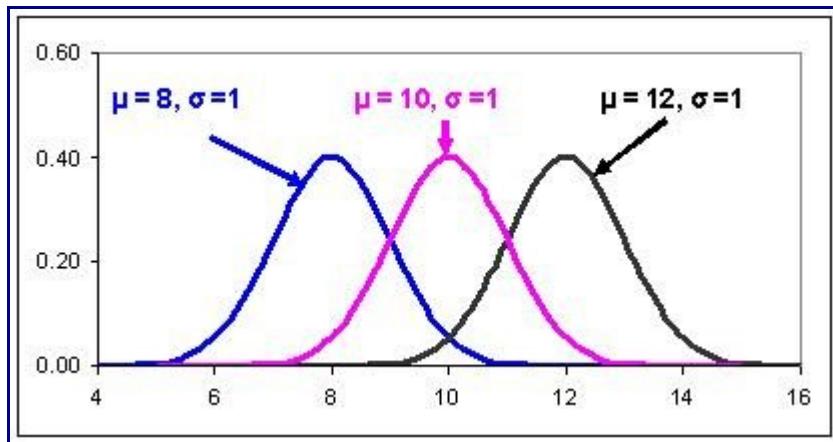
ومن هنا نعرف لماذا كانت معضظم القيم (99.7%) في حدود $\mu \pm 3\sigma$ أي في هذا المثال من 5 إلى 11 لأن المساحة تحت المنحنى من 5 إلى 11 تكاد تكون هي المساحة كلها وتنقى مساحة ضئيلة جاعلى الجانبين. وعملية حساب احتماليات وقوع المتغير بين قيمتين أو أكبر من قيمة ما أو أقل من قيمة ما يتم تقديره على وجه الدقة باستخدام الجداول التي تعطي المساحة تحت المنحنى في كل جزء منه أو باستخدام الحاسوب.

تأثير تغير قيمة المتوسط أو الانحراف المعياري

الشكل التالي يبين تأثير تغير الانحراف المعياري مع ثبات المتوسط. إن ما يحدث هو أن المنحنى يقل انباجا كلما زادت قيمة الانحراف المعياري. وهذا مرتبط بأن الانحراف المعياري هو مقياس لتشتت المنحنى وبالتالي فكلما زاد الانحراف المعياري فإن هذا يعني أن المنحنى ينتشر على مدى أوسع. فعندما كان الانحراف المعياري يساوي 0.5 كان التوزيع قريب جداً من المتوسط بينما ازداد اتساعاً عندما زادت قيمة الانحراف المعياري إلى 1 ثم ازداد اتساعاً عندما وصلت قيمة الانحراف المعياري إلى 2.



أما تغير المتوسط فيظهر في الرسم التالي. فالانحراف المعياري لكل منحنى من هذه المنحنيات متساوٍ بينما المتوسط مختلف. لاحظ أن المنحنيات الثلاثة متشابهة تماماً ولكن كل منها يتوزع حول متوسط مختلف.



بهذا نكون قد تعرفنا على منحنى التوزيع الطبيعي وفي المقالة التالية إن شاء الله نتعرف أكثر على هذا المنحنى وبعض التوزيعات الأخرى.

[من مراجع المقالة](#):

Applied Statistical Methods, W. Carlson and B. Thorne, Prentice Hall, 1997

Statistics for Managers, Levine et al., Prentice Hall, 1999

Lean Six Sigma Pocket ToolBook, George et al., McGraw Hill, 2005

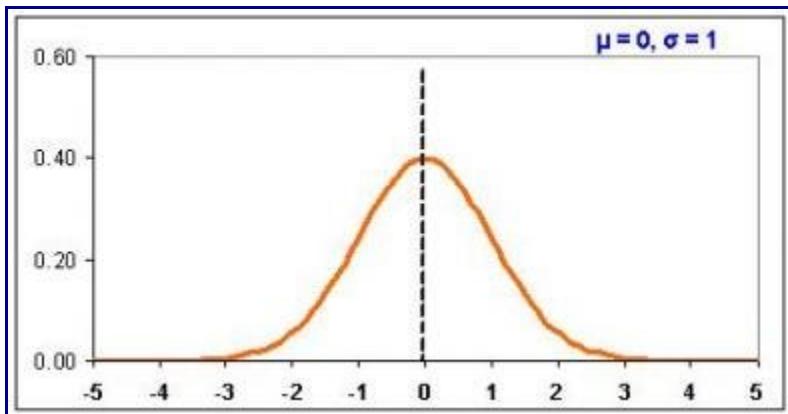
منحنى التوزيع الطبيعي القياسي ... Standard Normal Distribution

فبراير 16, 2010

تعرفنا في المقالة السابقة على منحنى التوزيع الطبيعي وخصائصه. في هذه المقالة نلقي المزيد من الضوء على التوزيع الطبيعي وذلك باستعراض التوزيع الطبيعي القياسي.

التوزيع الطبيعي القياسي (المعياري)؟

كما تعلم فإن منحنى التوزيع الطبيعي يُعرف بالمتوسط μ والانحراف المعياري σ . وقد يأخذ المتوسط أي قيمة ويأخذ الانحراف المعياري أي قيمة موجبة. أما منحنى التوزيع الطبيعي القياسي Standard Normal Distribution فهو توزيع طبيعي له متوسط يساوي الصفر وانحراف معياري يساوي واحد.



ويستخدم منحنى التوزيع الطبيعي القياسي لتحديد احتمالية أن يأخذ متغيراً يتبع التوزيع الطبيعي قيمة في مدى محدد. افترض أننا ندرس متغيراً ما مثل أخطاء الإنتاج اليومية أو أطوال مجموعة من الناس أو زمن عملية ما ووجدنا أنه يتبع توزيعاً طبيعياً بمتوسط يساوي 35 وانحراف معياري يساوي 2 ونريد أن نقدر احتمالية أن تكون قيمة هذا المتغير أكبر من 40. إننا بحاجة لجدول تبين المساحة تحت هذا المنحنى لأن هذه المساحة كما بينا في المقالة السابقة، تعبّر عن الاحتمالات. وبالتالي فإننا سنحتاج جدول لكل منحنى توزيع طبيعي وهذا أمر معقد جداً. لذلك فإننا نستخدم معادلة بسيطة لتحويل قيمة المتغير لمنحنى التوزيع القياسي وبالتالي يمكننا استخدام جدول واحد فقط وهو منحنى التوزيع الطبيعي القياسي.

وعملية التحويل من أي توزيع طبيعي للتوزيع الطبيعي القياسي تتم باستخدام معادلة بسيطة حيث نرمز للمتغير الأصلي بـ X ولما يقابله في المنحنى القياسي (المعياري) بـ Z . ويتم التحويل باستخدام المعادلة التالية:

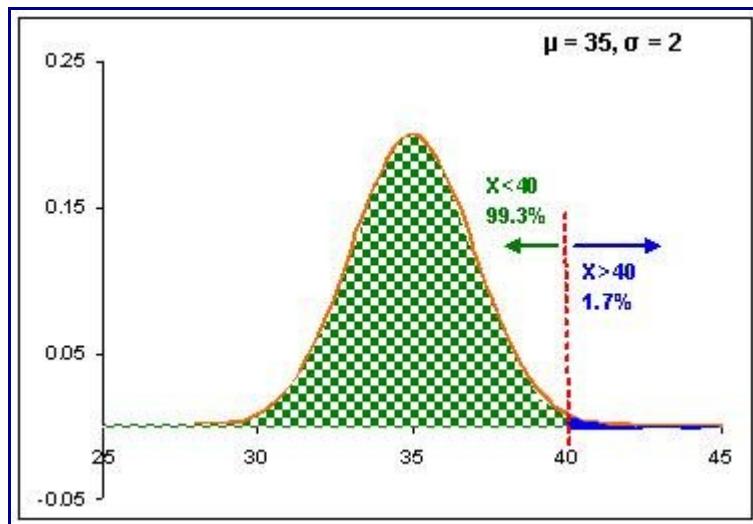
$$Z = \frac{X - \mu}{\sigma}$$

حيث μ هو المتوسط و σ هو الانحراف المعياري. ففي المثال السابق تكون قيمة Z المطلوبة لـ $X=40$ هي $2.5 = 2\frac{1}{2} (35 - 40)$

وبالتالي فإننا نبحث في جدول التوزيع الطبيعي القياسي عن قيمة 2.5 والتي نجدها تناهز 0.993 أي أن المساحة على اليسار تساوي هذه القيمة والتي تناهز أن تكون X أقل من 40. ولكننا نبحث عن احتمالية X أكبر من 40. وبالتالي فإننا نبحث عن المساحة على يمين المنحنى وهي $1 - 0.993 = 0.017$. أي أن احتمالية أن تتجاوز X الأربعين هي

.%1.7

لاحظ أن المساحة الكلية تحت منحنى التوزيع الطبيعي تساوي 1 في كل الأحوال ولذلك فإننا طرحتا القيمة التي حصلنا عليها من 1 لكي نحصل على المساحة على يمين المنحنى.



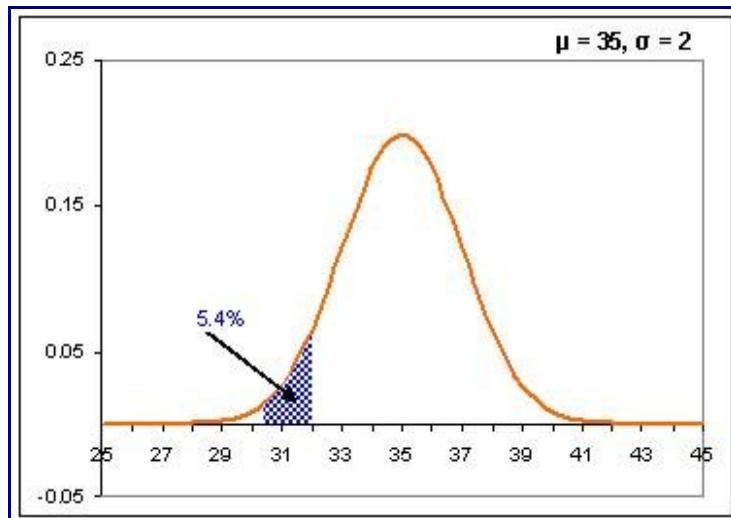
ويمكن الوصول لنفس النتيجة باستخدام برنامج إكسل Excel أو برنامج كالم Calc باستخدام الدالة NORMSDIST فنكتب في أي خلية $NORMSDIST(2.5) = 0.993$

ولكن علينا الانتباه إلى أن هذه هي المساحة على يسار الـ 2.5 فهي تعني احتمالية أن تكون X أقل من 40. هل يمكن تحديد احتمالية أن تكون X بين 30.5 و 32؟ نعم، علينا أن نحسب المساحة تحت المنحنى على يسار كل قيمة ثم نطرحهما لنحصل على المساحة بين هاتين القيمتين وهي كما تعلم تساوي احتمالية وقوع X بين هاتين القيمتين.

$$Z1 = (30.5 - 35) / 2 = -2.25$$

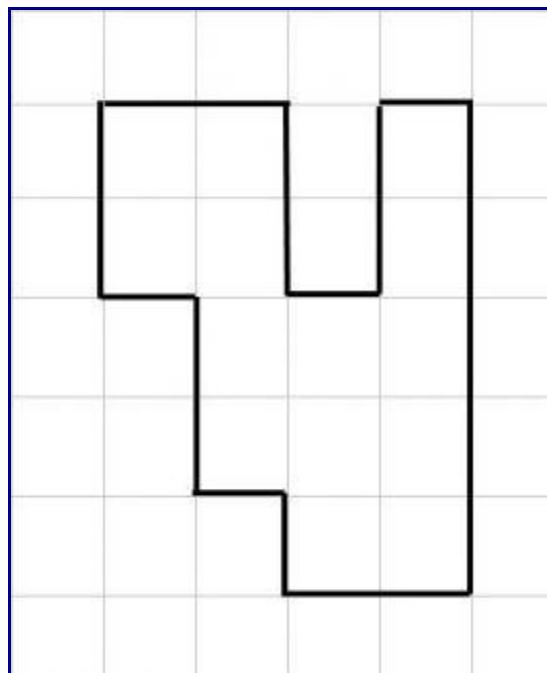
$$Z2 = (32 - 35) / 2 = -1.5$$

وباستخدام الجداول أو الحاسوب نجدا أن المساحتين هما 0.122 و 0.066 والفارق بينهما يساوي 0.054 أي أن احتمالية وقوع X بين 30.5 و 32 هي .% 5.4



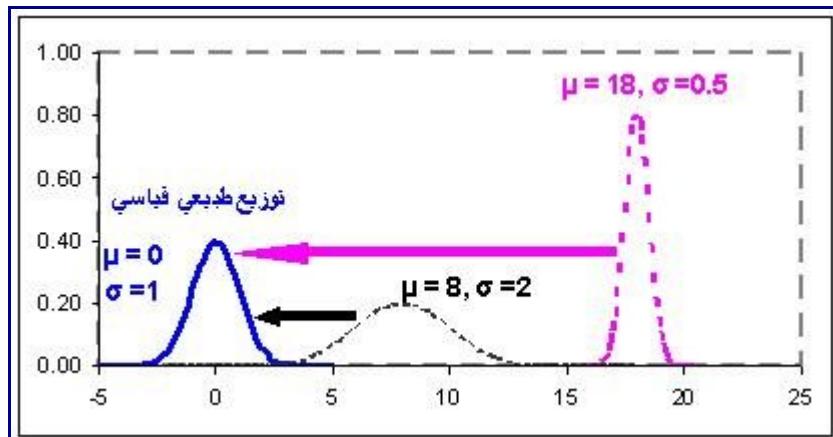
مفهوم التحويل لمنحنى التوزيع الطبيعي القياسي:

قد تبدو فكرة استخدام منحنى التوزيع الطبيعي القياسي لحساب الاحتمالات لمنحنيات طبيعية غير قياسية فكرة غريبة وغير واضحة ولكنها في الحقيقة شبيهة بأمور كثيرة مرت عليك من قبل. عملية التحويل لمنحنى التوزيع الطبيعي القياسي شبيهة بقياس مساحة ما بالبوصة المربعة ثم استخدام معامل التحويل لتحويلها إلى المتر المربع. وهي شبيهة كذلك برسم البلاد الكبيرة جدا على خريطة صغيرة باستخدام مقاييس الرسم. ويمكن تشبّه الأمر كذلك بقياس مساحة الشكل أدناه باستخدام مساحة المستطيلات الصغيرة التي تبلغ مساحتها 1 سنتيمتر مربع فنجد أن المساحة تساوي 14 سنتيمتر مربع.

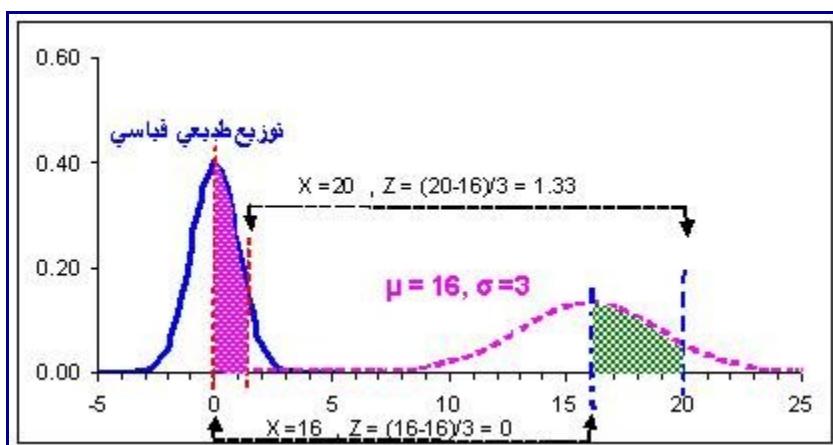


فمنحنى التوزيع القياسي هو وسيلة لحساب الاحتمالات (المساحة تحت المنحنى) لأي منحنى توزيع طبيعي. فيمكّنا تحويل القيمة (X) لأي متغير يتبع توزيعاً طبيعياً غير قياسي إلى نظيرتها (Z) في منحنى التوزيع الطبيعي القياسي وبالتالي نتمكن من تقدير المساحة تحت المنحنى. فالتحول من X إلى Z والعكس شبيه باستخدام مقاييس الرسم في

الخراط. وحساب المساحة تحت المنحنى الأول باستخدام المساحة تحت المنحنى القياسي تشبه قياس مساحة الشكل باستخدام المربعات الصغيرة معلومة المساحة.



والشكل أدناه يبين مثلاً لعملية التحويل. فلدينا توزيع طبيعي بمتوسط = 15 وانحراف معياري يساوي 3. ونريد أن تُقدر احتمالية أن يقع هذا المتغير بين 16 و 20. نستخدم التحويل التحويل القيمتين 16 و 20 لنظيرتيهما في التوزيع القياسي وهما 0 و 1.33. ما معنى هذا التحويل؟ معنى هذا التحويل أن المساحة التي نريد حسابها أصلًا والملونة باللون الأخضر والواقعة أسفل المنحنى الأصلي بين القيمتين 16 و 20 تساوي المساحة تحت المنحنى القياسي بين القيمتين 0 و 1.33 والملونة باللون الأحمر على الرغم من اختلاف الشكل. وبالتالي فالتحويل يمكننا من تقدير المساحة الملونة باللون الأحمر باستخدام جداول التوزيع الطبيعي القياسي أو باستخدام الحاسوب. وبذلك تكون قد وصلنا للمساحة الأصلية (الخضراء) والتي هي مُعبرة عن احتمالية أن تكون قيمة المتغير تحت الدراسة بين 16 و 20. وفي هذا المثال نجد هذه المساحة تساوي 0.40. أي أن المساحة بين 0 و 1.33 في المنحنى القياسي تساوي 0.40 وهي متساوية لمساحة تحت المنحنى الأصلي بين 16 و 20 وهذا يعني أن احتمالية وقوع المتغير بين 16 و 20 هي 40%.

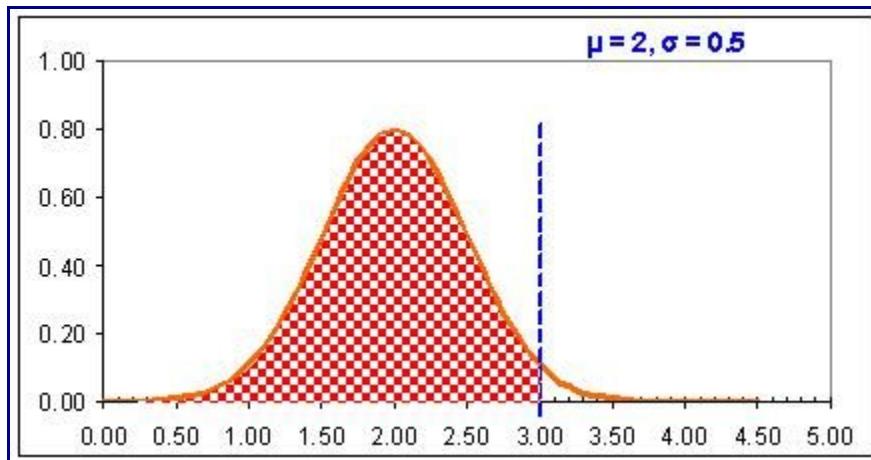


أمثلة:

المثال الأول: افترض أن زمن إعداد مشروب ما في مطعم يتغير من مرة لأخرى بمتوسط يساوي دقيقة وانحراف معياري يساوي 0.5 دقيقة. ما هي احتمالية أن يكون زمن إعداد المشروب أقل من 3 دقائق؟
أولاً نحسب قيمة Z المكافئة لـ X

$$Z = (3-2) / 0.5 = 2$$

باستخدام الجداول أو الحاسوب نجد أن المساحة تحت المنحنى على يسار القيمة 3 (الحمراء) تساوي 97.7% أي أن احتمالية أن يكون زمن إعداد المشروب أقل من 3 دقائق هو 97.7%.

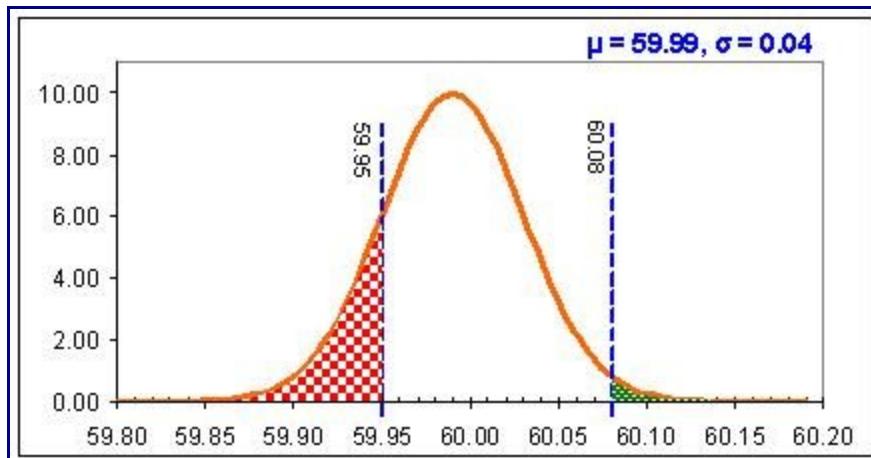


ويمكننا أن نستنتج أن احتمالية أن يكون زمن إعداد المشروب أكبر من 3 دقائق هي $1 - 97.7\% = 2.3\%$.

المثال الثاني: افترض أن طول قطعة يتم إنتاجها هو 60 سم ويطلب العميل أن يكون الطول في حدود 59.95 سم و 60.08 سم. وبمتابعة العملية الإنتاجية وجدنا أننا ننتج القطعة بمتوسط 59.99 سم وبانحراف معياري 0.04 سم. ما هي احتمالية تجاوز التفاوت الذي يسمح به العميل؟



الشكل أدناه يبين منحنى التوزيع الطبيعي الذي يمثل تغير طول هذه القطعة في الإنتاج. والمطلوب هو حساب المساحة على يمين 60.08 (الخضراء) والمساحة على يسار 59.95 (الحمراء).



نحسب قيمة Z المكافئة لـ 59.95 فجدها

$$Z = (59.95 - 59.99) / 0.04 = -1$$

باستخدام الجداول أو الحاسوب نجد أن المساحة على يسار هذه القيمة تساوي 15.87 %. هل هذه هي القيمة التي نبحث عنها أم ينبغي أن نظر لها من 1 كما فعلنا في المثال السابق؟ نحن نبحث عن احتمالية أن يقل الطول عن هذه القيمة فنحن فعلاً نريد المساحة على يسار هذه القيمة.

ثم نحسب قيمة Z المكافئة لـ 60.08 فنجد أنها

$$Z = (60.08 - 59.99) / 0.04 = 2.25$$

باستخدام الجداول أو الحاسوب نجد أن المساحة على يسار هذه القيمة تساوي 98.78 %. هذه القيمة تبين احتمالية أن يقل الطول عن 60.08 سم ولكننا نسأل ما هي احتمالية أن يزيد الطول عن ذلك. فعلينا أن نطرح هذه القيمة من 1 (المساحة الكلية تحت المنحنى) فتحصل على 1.2%.

وبالتالي فإن احتمالية تجاوز الحد الأدنى للطول هي 15.87% واحتمالية تجاوز الحد الأقصى هي 1.2%. ويمكن أن نجمعهما ونقول أن احتمالية تجاوز التفاوت المحدد للطول هي 17.07%.

هل هذا ترف أكاديمي؟ بالطبع لا، فالأمثلة التي استعرضناها تعطي أرقاماً مهمة تساعد المدير على اتخاذ القرارات. وفي المثال الأخير يبدو أن احتمال الخطأ يعتبر كبيراً وبالتالي فهو المؤسسة إما أن ترفض الالتزام بهذا العمل أو أن تطور أسلوب الإنتاج تطويراً كبيراً يقلل من نسبة الخطأ. وفي المثال الأول قد تجد إدارة المطعم أن الحفاظ على زمن إعداد المشروب أقل من 3 دقائق في 97.7% من الحالات هو أمر مقبول وقد تستهدف ما هو أفضل من ذلك للوصول إلى نسبة 99%.

في المقالة التالية إن شاء الله نستعرض المزيد من الأمثلة ونناقش كيفية قراءة جداول منحنى التوزيع الطبيعي القياسي.

[من مراجع المقالة:](#)

Applied Statistical Methods, W. Carlson and B. Thorne, Prentice Hall, 1997

Statistics for Managers, Levine et al., Prentice Hall, 1999

Lean Six Sigma Pocket ToolBook, George et al., McGraw ill, 2005

[مقالات ذات صلة:](#)

[منحنى التوزيع الطبيعي](#)

[تلویص البيانات](#)

[تلویص البيانات باستخدام برنامج إكسيل](#)

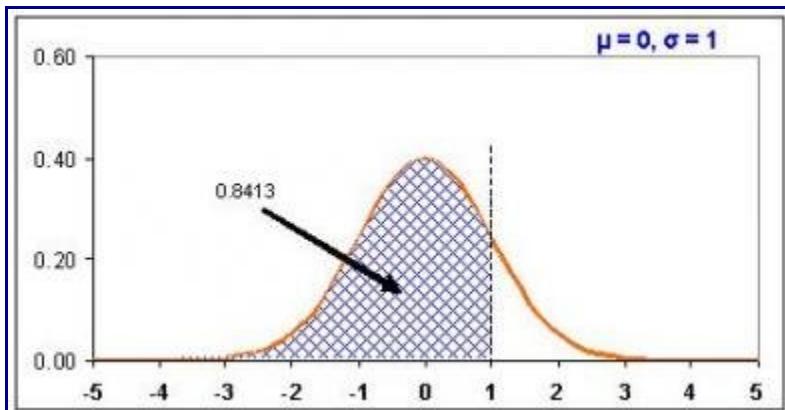
منحنى التوزيع الطبيعي القياسي - 2

فبراير 25, 2010

ناقشنا في المقالة السابقة [منحنى التوزيع الطبيعي القياسي](#) وهو منحنى له متوسط يساوي صفر وانحراف معياري يساوي 1. ويستخدم هذا المنحنى كوسيلة لتحديد المساحة تحت أي منحنى توزيع طبيعي والتي تمثل احتمالية أن يأخذ المتغير قيمًا في مدى محدد. نستكمل في هذه المقالة استعراض منحنى التوزيع الطبيعي القياسي فنناقش أنواع جداول منحنى التوزيع الطبيعي القياسي والتي تُعطي المساحة تحت المنحنى لقيم مختلفة لـ Z . ثم نستعرض بعض الأمثلة الإضافية لتطبيقات منحنى التوزيع الطبيعي.

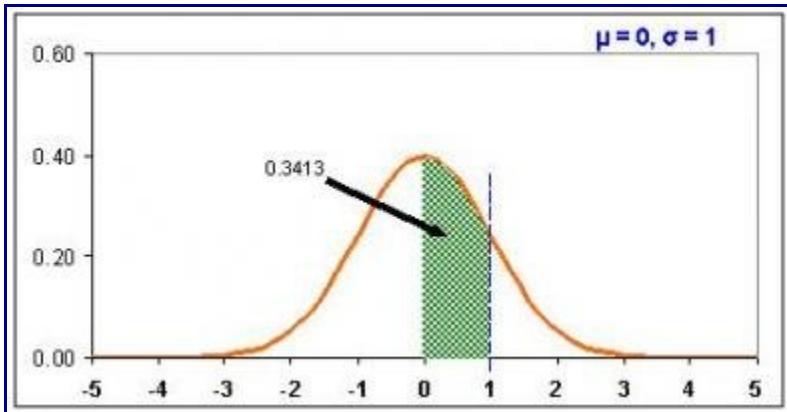
أنواع الجداول:

في هذا العصر أصبح من اليسير أن تحسب الاحتمالات الخاصة بمنحنى التوزيع الطبيعي باستخدام الحاسوب ولكن في نفس الوقت قد تحتاج أن تستخدم الجداول. وهناك أكثر من أسلوب عرض لهذه الجداول. بعض هذه الجداول يعطيك المساحة على يسار القيمة فمثلاً عند قيمة $Z=0$ يعطيك 0.5 لأن المتوسط يقسم المساحة إلى نصفين متساويين وبالتالي فالمساحة على يمين المتوسط تساوي 0.5 لأن المساحة الكلية تحت المنحنى القياسي تساوي 1. الشكل التالي يبين مفهوم هذا الجدول. فالرقم المناظر لقيمة $Z=1$ هو 0.8413 وهو المساحة الكلية على يسار $Z=1$.



Z	0.00	0.01	0.02	0.03	0.04	0.05
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596
0.2	0.5973	0.5832	0.5871	0.5910	0.5948	0.5987
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368

بعض الجداول يبدأ من المتوسط فيعطيك عند المتوسط ($Z = 0$) صفرًا. ولذلك يجب النظر إلى القيمة المناظرة لـ $Z=0$ لكي نفهم أسلوب التعامل مع الجداول. الشكل التالي يبين معنى القيم التي تحصل عليها من هذا الجدول فالرقم المناظر لقيمة $Z=1$ هو 0.3413 وهو يقل عن الرقم الذي حصلنا عليه من الجدول الأول بـ 0.5 وهي قيمة المساحة على يسار المتوسط. فهذا الجدول يعطيك لمساحة المحصورة بين الرقم والمتوسط وهي المساحة المظللة باللون الأخضر. (لاحظ أن الجدول المعروض أدناه هو جزء من الجدول وليس الجدول كله)



Z	0.00	0.01	0.02	0.03	0.04	0.05
0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596
0.2	0.0973	0.0832	0.0871	0.0910	0.0948	0.0987
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368

ما هي الأرقام الموجودة في أول صف؟

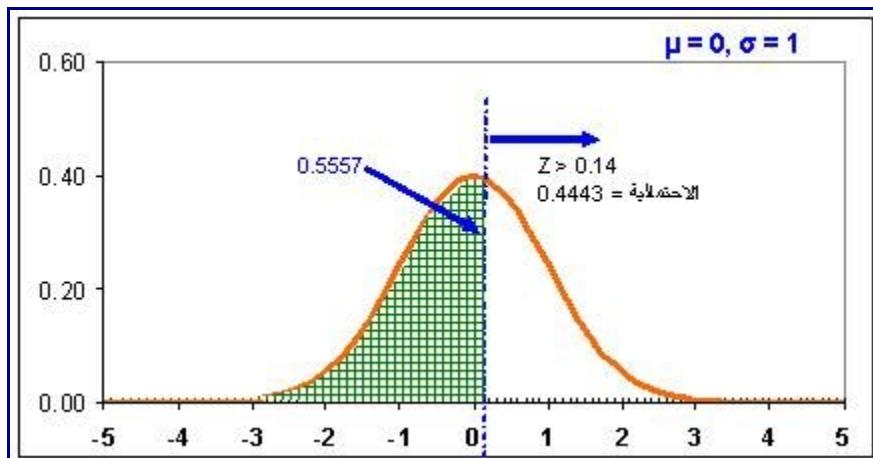
إن هذا الجدول يستخدم أسلوباً خاصاً لتحديد الكسر المئوي L Z . افترض أننا نريد القيمة المناظرة $L=0.23$. إنها في الصف المناظر $L=0.2$ والعمود المناظر $L=0.03$ أي $0.03 + 0.2 = 0.23 = 0.03 + 0.2$. والقيمة في الجدول الأخير هي 0.0910 وهي في الجدول الأول 0.5910 (الفارق بينهما 0.5 كما ذكرنا). ومثلاً القيمة المناظرة $L=0.35$ هي في الصف المناظر $L=0.3$ والعمود المناظر $L=0.05$. وهي في الجدول الأخير 0.1368 وفي الجدول الأول 0.6368.

Z	0.00	0.01	0.02	0.03	0.04	0.05
0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596
0.2	0.0973	0.0832	0.0871	0.0910	0.0948	0.0987
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368

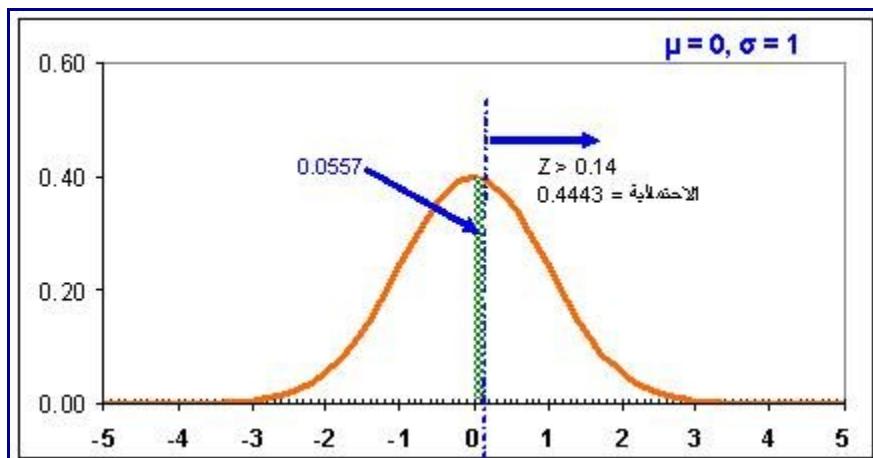
كيف نستخدم الجدول؟

على سبيل المثال لو كنا نريد أن نعرف احتمالية أن يقل المتغير عن قيمة كذا والتي هي أكبر من المتوسط فإننا نلجم الجدول وافتراض أننا وجدنا القيمة هي 0.32 وكانت القيمة عند المتوسط هي صفر فإننا نضيف 0.5 فتصبح الاحتمالية هي 0.82. أضفنا 0.5 لنضيف المساحة على يسار المتوسط. أما لو كانت القيمة في الجدول المناظرة للمتوسط هي 0.5 فإننا سنجد القيمة المناظرة لقيمة المتغير 0.82 ولنحتاج لإضافة 0.5. الأمر يبدو في البداية صعباً ولكن بفهم الجدول وفهم المساحة التي نريد حسابها نستطيع الوصول للنتيجة الصحيحة.

افرض أنك تريد احتمالية تجاوز Z لـ 0.14. سندخل للجدول الأول ونجد القيمة المقابلة هي 0.5557. ما هذه القيمة؟ إنها المساحة الكلية على يسار $Z = 0.14$. ولكننا نريد المساحة على يمين $Z = 0.14$ لأنها هي التي تمثل احتمالية زيادة Z عن 0.14. لذلك نطرح 0.5557 - 0.4443 فنحصل على الناتج وهو 0.4443.



ولو استخدمنا الجدول الأخير لوجدنا القيمة المقابلة هي 0.0557 وهي القيمة من المتوسط وحتى 0.14. ولكننا نريد المساحة بعد 0.14 لذا نطرح 0.5 - 0.0557 = 0.4443. كان من الممكن أن نضيف 0.5 لكي نحصل على المساحة الكلية على يسار 0.14 ثم نطرحها من 1 كما فعلنا في الجدول الأول.



أمثلة أخرى على منحنى التوزيع الطبيعي القياسي:

المثال الأول: افترض أننا رسمنا المدرج التكراري لحجم المبيعات اليومي ووجدنا أنه يتبع منحنى التوزيع الطبيعي بمتوسط 2600 وانحراف معياري 80. ونريد أن نعرف حجم المبيعات المتوقع في 95% من الأيام.

علينا أن نسأل أنفسنا ما هي المساحة التي تبين حجم المبيعات في 95% من الأيام؟ هل المطلوبة المساحة التي تساوي 0.95 بدءاً من اليسار أم من اليمين؟ في الواقع إننا نريد أن نستبعد الأرقام النادرة الحدوث. فمثلاً على الرغم من أن المتوسط يساوي 2600 فإننا في بعض الأيام النادرة قد نبيع 2700 أو 2200. ولكننا لكي نت Exped بعض القرارات الإدارية نريد أن نحدد مدى لحجم المبيعات فنقول إننا في 95% من الأيام نبيع ما قيمته كذا إلى كذا. لذلك فنحن نريد أن نستبعد حجم المبيعات النادر الحدوث سواء كان كبيراً أو صغيراً.

معنى هذا أننا سنستبعد المساحة أقصى اليمين والمساحة أقصى اليسار وتبقى مساحة في الوسط تساوي 0.95. فما هي المساحة على اليمين واليسار؟ لكي نصل إلى مساحة في المنتصف تساوي 0.95 فإننا سنستبعد من الجانبين ما قيمته

$1 - 0.95 = 0.05$. وهذه المساحة مقسمة بالتساوي على الجانبين أي إننا سنستبعد مساحة قدرها 0.025 من اليمين و 0.025 من اليسار.

فما هي القيم التي سنبحث عنها؟ إننا نريد القيمة المناظرة لمساحة 0.025 وذلك لنستبعد المساحة على اليسار. مادا عن المساحة على اليمين؟ نظرا لأن لمساحة الكلية تساوي واحد فإننا نبحث عن المساحة المناظرة $1 - 0.025 = 0.975$. لقد قمنا بالطرح من 1 لأن الجداول لا تعطي المساحة لعى اليمين بل تعطينا دائما المساحة على اليسار.

نستخدم الجداول فنبحث بطريقة عكسية لما لتبناه سابقا. إننا نبحث عن قيمة المساحة (الاحتمال) ثم نحدد قيمة المناظرة لها. في الأمثلة السابقة كنا نعرف Z ونريد المساحة المناظرة لها ولكننا هنا نفعل العكس حسب طبيعة السؤال. ويمكن استخدام الحاسوب ولكننا هنا سنستخدم الدالة

NORMSINV

فنكتب في أي خلية

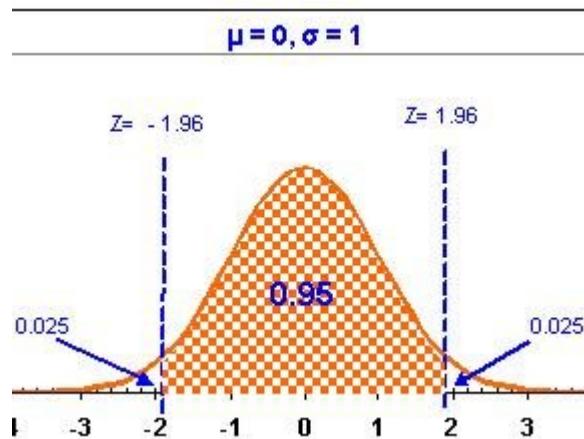
(NORMINV (0.025

فحصل على -1.96

ثم نكتب

(NORMINV (0.975

فحصل على 1.96



فحن نبيع ما يوازي $Z = -1.96$ وحتى $Z = 1.96$ في 95% من الأيام. فما هي القيمة الحقيقة المناظرة $1 - 0.95 = 0.05$ ؟ نستخدم نفس المعادلة المعتادة ولكننا هنا نريد تحديد X بمعرفة قيمة Z فتصبح المعادلة كالتالي

$$X = \mu + Z * \sigma$$

وبالتعويض

$$X = 2600 + (-1.96) * 80 = 2443$$

$$X = 2600 + (1.96) * 80 = 2757$$

أي إننا في 95% من الأيام نبيع ما يتراوح بين 2443 و 2757. ويمكن أن نقول ذلك بصيغة أخرى وهي إننا متاكدين بنسبة 95% أننا نبيع ما يتراوح بين 2443 و 2757.

ما هي المساحة التي نريد حسابها؟ إننا لا نريد حساب مساحة بل نبحث عن قيمة Z المناظرة لمساحة ما. فما هي المساحة؟ هل هي 0.1 ؟ إننا لا نريد المساحة التي تنتظر Z التي ستنتجاوزها في 10% من الحالات أي لن تتجاوزها في 90% من الحالات. فنحن نريد قيمة Z المناظرة لمساحة 0.9 . وبنفس الطريقة نريد حساب Z المناظرة له 0.99 و 0.95 .

باستخدام الحاسوب والدالة NORMSINV نحصل على قيم Z كالآتي:

Z=1.28 -0.90

Z=1.64 -0.95

Z=2.32 - 0.99

ما هو الزمن المناظر لهذه القيم لـ Z أي ما هي قيمة X المناظرة لـ Z ? نستخدم نفس المعادلة

$$X = \mu + Z * \sigma$$

بالحساب نحصل على قيم X وهي على التوالي: 17.6، 18.3، 19.7. معنى ذلك أننا نقدم الوجبة في أقل من 17.6 دقيقة في 90% من الحالات ونقدمها في أقل من 18.3 دقيقة في 95% من الحالات ونقدمها في أقل من 19.7 دقيقة في 99% من الحالات

بذلك تكون قد قدمنا لإدارة المطعم معلومات عظيمة تمكّنهم من اختيار الزمن الذي سنلتازم به تجاه العميل. فعلينا أن نحسب تكلفة 10% من الوجبات و 5% من الوجبات و 1% من الوجبات ونختار ما هو مناسب من ناحية التكلفة والمنافسة

في المقالة الثالثية إن شاء الله نستعرض نظرية الحد المركبة وكذلك بعض التوزيعات الأخرى.

من مراجعة المقالة

Applied Statistical Methods, W. Carlson and B. Thorne, Prentice Hall, 1997

Statistics for Managers, Levine et al., Prentice Hall, 1999

مقالات ذات صلة:

منحنى التوزيع الطبيعي

تلخيص البيانات

تلخيص البيانات باستخدام برنامج إكسل

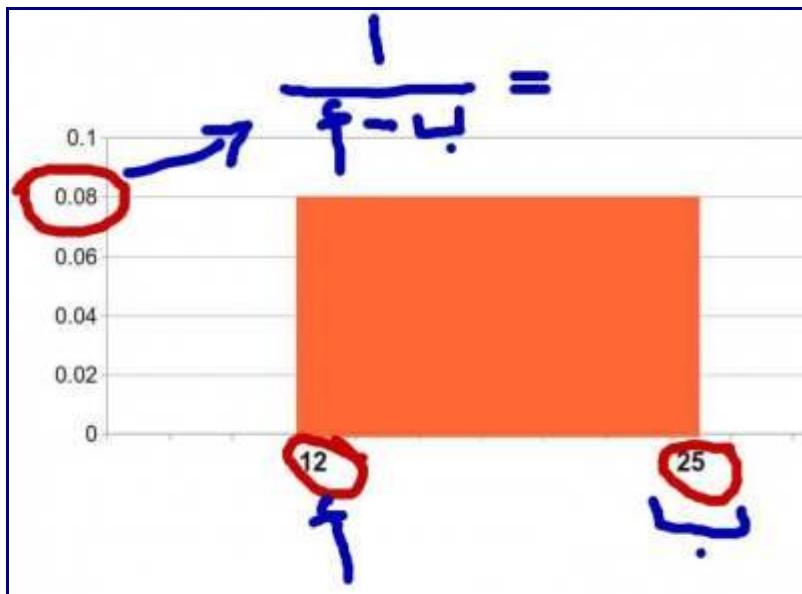
بعض التوزيعات الأخرى

مارس 5, 2010

تعرفنا على **التوزيع الطبيعي** ولكن التوزيع الطبيعي ليس وحدها توزيعات أخرى مثل التوزيع المنتظم والأسى. فعندما نجمع البيانات ونرسم المدرج التكراري قد نجد أنها تتبع التوزيع الطبيعي أو غيره. في هذه المقالة نتعرّف على توزيعات احتمالية أخرى. وهذه التوزيعات الاحتمالية ليست حسابات رياضية معقدة ولا هي أمر نظري لا علاقة له بالعمل. هذه التوزيعات الاحتمالية تستخدم في مجالات العمل فهي تستخدم لدراسة سرعة الخدمة ومعدل وصول العملاء وتستخدم لتحليل معدل المشاكل في المعدات وتستخدم لمحاكاة أي عملية وتستخدم لدراسة حجم المبيعات. ويمكننا التعامل مع هذه التوزيعات بدون الدخول في تعقيدات رياضية لأنه يمكننا استخدام الحاسوب.

التوزيع المنتظم :Uniform Distribution

التوزيع المنتظم يختلف عن التوزيع الطبيعي في أن احتمالية وقوع المتغير بين أي قيمتين لا يتغيّر. فلو كان لدينا توزيع منتظم لطول المنتج من 10.5 إلى 11.00 فإن احتمالية أن يكون طول المنتج بين 10.5 و 10.6 تساوي احتمالية أن يكون طول المنتج بين 10.7 و 10.8 وهي نفس احتمالية أن يكون طول المنتج بين 10.8 و 10.9 وهكذا. فالتوزيع منتظم ولا يزداد في المنتصف كما في حالة التوزيع الطبيعي.



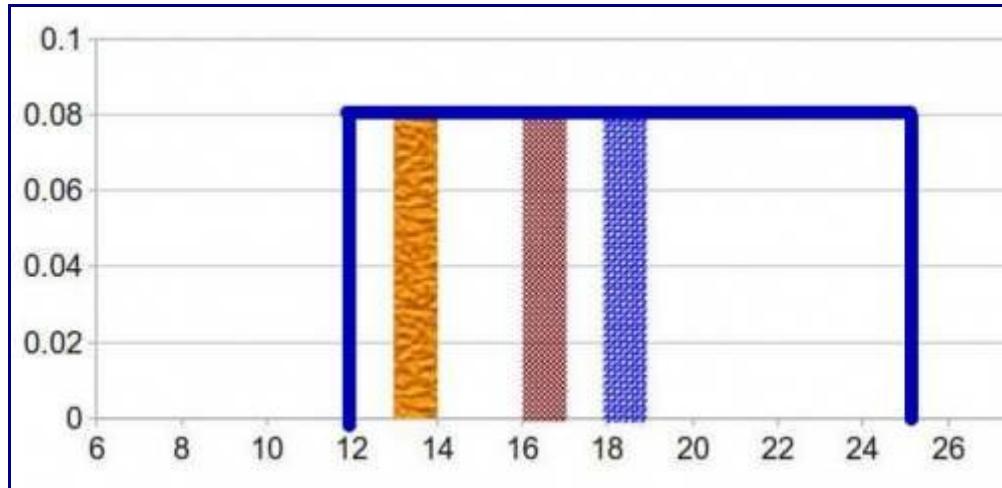
التوزيع المنتظم هو توزيع يبيّن أن المتغير يقع بين قيمتين محددتين هما a و b . فالرسم أعلاه يبيّن أن هذا المتغير يقع بين 12 و 25 فهو لا يقل عن 12 ولا يزيد عن 25. ويبين كذلك أن وقوع المتغير بين قيمتين مثل 12 و 14 تساوي احتمالية وقوع المتغير بين 17 و 19 أو بين 21 و 23 وهكذا. والمساحة تحت المنحنى كما هي الحال في منحنى التوزيع الطبيعي تساوي 1 ولذلك فإن الخط الأعلى عند

$$1 / (b - a)$$

$$\text{وهو في هذه الحالة } = 1 / (25 - 12) = 0.08$$

المتوسط في منحنى التوزيع المنتظم يساوي $(a + b) / 2$ وهو في هذه الحالة يساوي $(12 + 25) / 2 = 18.5$. أما الانحراف المعياري فيساوي الجذر التربيعي لـ $(b - a)^2 / 12$ وهو في هذه الحالة يساوي 3.75.

كيف نفهم عملية توزيع الاحتماليات توزيعاً منتظماً؟ إن الاحتمالية هنا يتم قياسها بالمساحة تحت المنحنى. فلو أردنا أن نقيس احتمالية وقوع المتغير بين 13 و 14 فإننا نحسب المساحة تحت المنحنى. وبما أن هذه المساحة هي عبارة عن مستطيل له ارتفاع يساوي 0.08 في هذا المثال فإن احتمالية وقوع المتغير بين 16 و 17 أو 18 و 19 هي نفس احتمالية وقوع المتغير بين 13 و 14 لأن المساحات متساوية.



التوزيع المنتظم يتميز بسهولة فهمه وهو يستخدم عادة لتخليق أرقام عشوائية. ويستخدم التوزيع المنتظم كافتراض مبدئي لتوزيع بيانات لا نعرف توزيعها مثل أي عملية جديدة.

التوزيع الأسّي: Exponenetal Distribution

التوزيع الأسّي هو من التوزيعات المهمة لأن له تطبيقات عديدة من أشهرها نظرية الطوابير أو خطوط الانتظار Queueing Theory. فالفترة الزمنية ما بين وصول عميل وآخر لمركز الخدمة يتبع عادة التوزيع الأسّي ولذلك فإن نظرية خطوط الانتظار تعتمد على التوزيع الأسّي. وكلمةأسّي هنا هي من الأسماء مثل أن تقول 2 أس 3 فتكلبتها 2³. والسبب في هذه التسمية أن هذا التوزيع يعتمد على معادلة رياضيةأسّية. وهذه المعادلة هي:

$$S(t) = e^{-\lambda t}$$

$$f(t) = \lambda e^{-\lambda t}$$

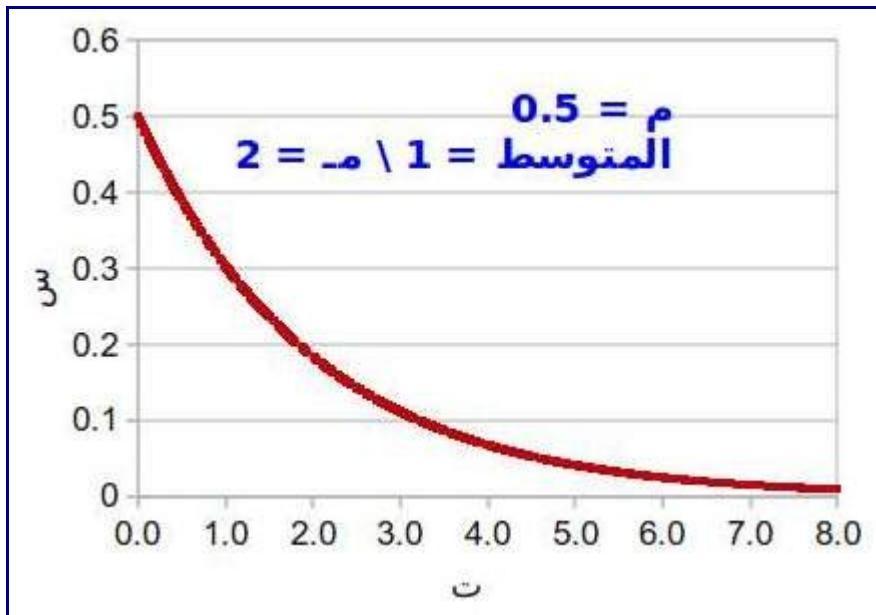
حيث $S(t)$ هي قيمة هذه المعادلة.

t أو t هي الفترة الزمنية بين وصول عميل وآخر

λ أو λ هي معدل وصول العملاء في الدقيقة (أو الثانية)

e أو e هي أساس اللوغاريتم الطبيعي أو ببساطة هو ثابت يساوي 2.718

وهذا التوزيع يأخذ الشكل التالي:



ومتوسط هذا التوزيع يتم حسابه بالمعادلة: $\text{المتوسط} = 1 / \lambda$ وهي تعبّر عن متوسط الفترة الزمنية بين وصول عميل آخر. وكما تلاحظ فالتوزيع ليس منتظماً ولا موزعاً بالتماثل حول المتوسط. والذي يحدث في الطبيعة هو أن الزمان بين وصول عميل آخر يأخذ في الأغلب قيمًا في مدى محدد فهو في هذا المثال بين 0 و 3 تقريبًا ثم قد يكون الزمان بين وصول أي عميلين طويلاً جدًا في أحيان قليلة وهذا هو ما يمثله هذا المنحنى. فيمكنك بمجرد النظر أن تلاحظ أن المساحة تحت المنحنى بين 0 و 2 وتلك أكبر من المساحة بين 2 و 3 وهكذا حتى تجد المساحة بين 7 و 8 تقترب من الصفر.

وعندما تتبع عملية الوصول لهذا المنحنى فإنه يصبح من السهل حساب احتمالية أن يكون الزمن بين كل عميلين أكبر من قيمة ما أو أقل من قيمة ما أو بين قيمتين محددين. فلمعرفة احتمالية أن يكون الزمن بين وصول عميلين أقل من قيمة ما (t_1) نستخدم المعادلة التالية:

$$1 - e^{-\lambda t}$$

ففي المثال السابق فإن معدل وصول العملاء هو نصف عميل كل دقيقة فيمكن حساب متوسط الفترة الزمنية بين كل عميلين وهي تساوي $1 / \lambda = 2$ دقيقة. مما هي احتمالية أن تكون الفترة بين عميلين أقل من 0.3 دقيقة. نستخدم المعادلة فنحصل على 0.14. أي أن احتمالية أن تكون الفترة بين عميلين أقل من 0.3 دقيقة هي 14%. ويمكن الوصول لنفس النتيجة باستخدام إكسل أو كالك (المكتب المفتوح) حيث نستخدم الدالة EXPONDIST كالتالي

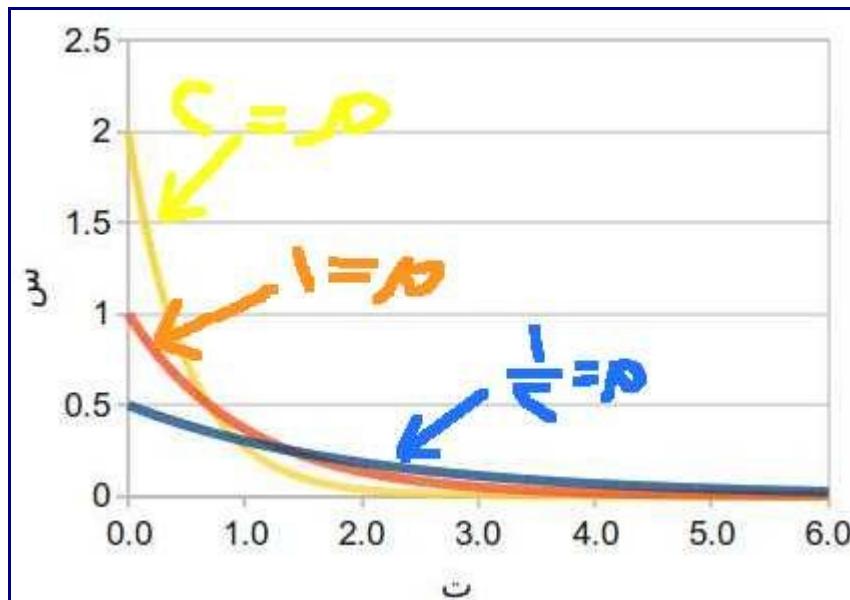
$$\text{EXPONDIST}(0.3, 0.5, 1 =$$

وبنفس الطريقة يمكننا حساب احتمالية أن يكون الزمن بين وصول عميلين أقل من 7 دقائق فنجد أنه 0.97% أي 97%. ومعنى هذا أن احتمالية أن يكون الزمن بين وصول عميلين أكبر من 7 دقائق هو 3% فقط.

التوزيع الأسوي يستخدم عادةً لتوزيع الزمن بين وصول العملاء سواء كانوا بشرًا أو معدات وكذلك لتوزيع الزمن بين حدوث عطليين في ماكينة. كما يستخدم لتوزيع زمن الخدمة. ولو فكرت في زمن خدمة العميل لوجدت أنه في الأغلب يقع في مدى محدد بين صفر وقيمة ما ولكن هذا لا يمنع أنه من آن لآخر فإن عميلاً يأخذ وقتاً أطول بكثير ليسوفي خدمته.

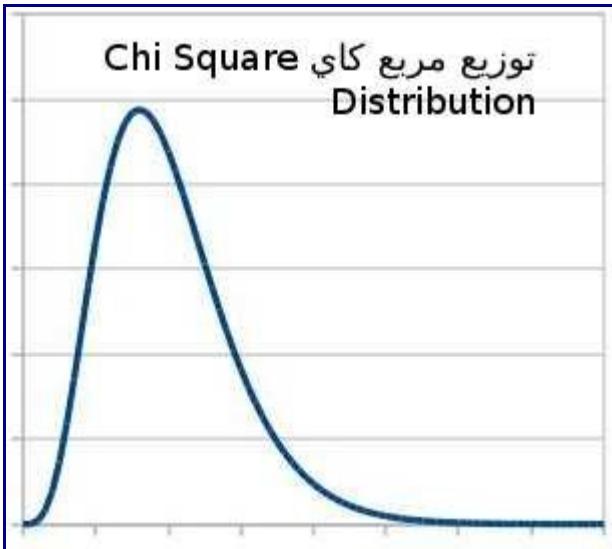
افرض أن معدل حدوث عطل في ماكينة ما هو مرة واحدة في الشهر ونريد معرفة احتمالية أن تكون الفترة الزمنية بين عطلين هي بين أسبوع وأسبوعين. يمكننا أن نستخدم الحاسوب فنحسب احتمالية أن تكون الفترة الزمنية أقل من أسبوع (0.21) ثم أقل من أسبوعين (0.37) ثم نطرح النتيجتين فنحصل على 0.16 أي 16%.

كيف يؤثر تغير معدل حدوث الأعطال أو وصول العملاء أو زمن الخدمة على شكل هذا المنحنى؟ الشكل التالي يبين ثلاثة منحنيات تمثل كل منها قيمة مختلفة لمعدل الوصول فهي على التوالي 0.5 ، 1 ، 3 . دعنا نقارن احتمالية أن يطول الفاصل بين وصول عمليين عن 2 . باستخدام الحاسوب نحسب احتمالية أن يقل الزمن عن 2 لكل منحنى فنحصل على 0.63 ، 0.86 على التوالي. ولكننا نريد احتمالية أن يطول الفاصل عن 2 وليس أن يقل عن 2 . لذلك نطرح كل قيمة من 1 فنحصل على 0.37 ، 0.14 ، 0.05 أي 37% ، 14% ، 5% .

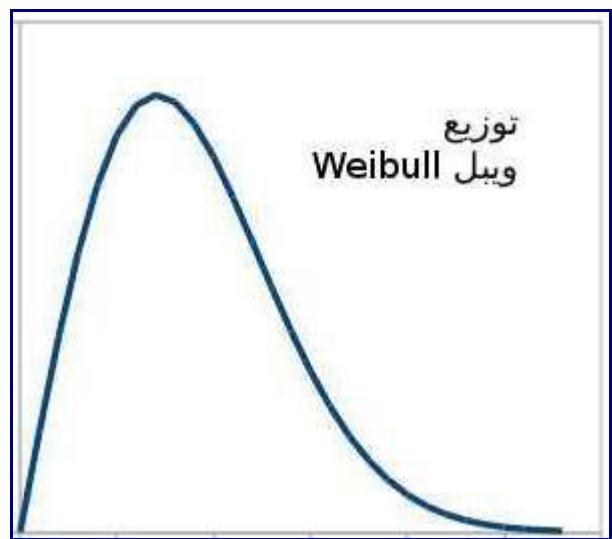


توزيعات أخرى:

وهناك توزيعات نظرية كثيرة مثل: توزيع مربع كاي Chi Sqaure Distribution والذى له بعض التطبيقات الإحصائية مثل اختبار مربع كاي. الشكل التالي يبين شكل منحنى مربع كاي. لاحظ أن شكل هذا التوزيع يختلف عن التوزيع الطبيعي في أنه غير对称 حول المتوسط بل هو منحرف تجاه اليمين.



وهناك توزيع ويبل Weibull Distribution



وهناك توزيع بيتا Beta Distribution والتوزيع المثلثي Triangular Distribution وتوزيع ف F Distribution وتوزيع ت t Distribution. وكل منها معادلة مختلفة وشكل مختلف وتطبيقات مختلفة. وهناك برامج تساعد على البحث عن التوزيع الاحتمالي المناسب للبيانات الحقيقية منها برنامج [Stat::Fit](#).

ولكن على الرغم من كثرة التوزيعات فإن منحني التوزيع الطبيعي هو أكثرها استخداماً نظراً لأن الكثير من الأمور تتبع التوزيع الطبيعي ونطراً لإمكانية استخدام التوزيع الطبيعي بدلاً من أي توزيع آخر عند تجميع البيانات باستخدام عينات وهذا ما سنناقشه في المقالة التالية إن شاء الله عند الحديث عن نظرية الحد المركزية. وهذا لا يعني الاستغناء تماماً عن التوزيعات الأخرى فإن هناك بعض الاختبارات الإحصائية التي تعتمد على هذا التوزيع أو ذاك. كما وأننا في بعض التطبيقات مثل [المحاكاة](#) نبحث عن التوزيع الذي يمثل البيانات لكي نقوم باستخدامه في نموذج المحاكاة وذلك لمحاكاة التغيرات التي تحدث في الطبيعة بنفس قيمتها وتغييرها.

[من مراجع المقالة](#):

Simulation Modeling and Analysis, Law and Lelton, Third Edition, McGrawHill, 2000

Simulation Using Promodel, Harrel et al., McGrawHill, 2000

Applied Statistical Methods, W. Carlson and B. Thorne, Prentice Hall, 1997

Statistics for Managers, Levine et al., Prentice Hall, 1999

The Six Sigma Handbook, Pyzdek, McGrawHill, 2003

نظريّة الحد المركزيّة... Central Limit Theorem...

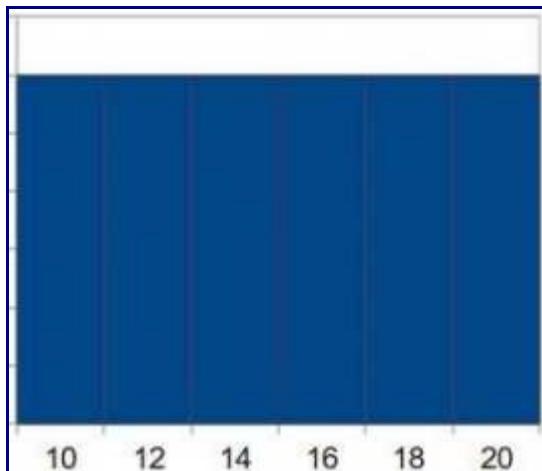
مارس 10, 2010

تعرفنا على التوزيع الطبيعي وعلى عدد من التوزيعات الأخرى. في هذه المقالة نستعرض نظرية الحد المركزيّة والتي تجعلنا نستطيع استخدام التوزيع الطبيعي في حالات لا يتبع فيها توزيع البيانات التوزيع الطبيعي.

نظريّة الحد المركزيّة: Central Limit Theory

افترض أنك جمعت بيانات عن طول المنتج ووجدها تتبع توزيعاً منتظاماً أو توزيعاً أسيّاً. في هذه الحالة لا تستطيع استخدام التوزيع الطبيعي. ولكن في الواقع فإننا نقوم بقياس 50 قطعة كل ساعتين ثم نسجل متوسط الطول وهكذا. أي أننا نأخذ عينات كل فترة زمنية ونسجل متوسط قيمة مفردات هذه العينة. معنى ذلك أننا نتعامل مع متوسط العينات. نظرية الحد المركزيّة تقول أنه يمكننا أن نستخدم التوزيع الطبيعي في هذه الحالة وفي أي حالة مماثلة. أمر عجيب ورائع. نعم رائع لأن معظم تعاملنا سيكون مع منحنى التوزيع الطبيعي الشهير والسهل ولن نشتّت جهودنا بين توزيعات كثيرة وتعقيدات حسابية. وهو أمر عجيب لأننا نقول أن المتغير الذي نقيسه لا يتبع التوزيع الطبيعي ثم نقول أنه يمكننا أن نستخدم التوزيع الطبيعي.

دعنا نفكّر في الأمر. افترض أن الطول موزع بانتظام بين 10 و 20. ما الذي يحدث عندما نأخذ 50 عينة عشوائية ثم نحسب المتوسط؟



إننا لو رسمنا توزيع الطول لهذه العينة لوجدناه يتبع التوزيع المنتظم. ولكننا لا نفعل ذلك. إننا نحسب متوسط الطول أي نحصل على متوسط 50 قطعة. ما الذي يحدث مع العينة الثانية ثم الثالثة ثم الرابعة. لا نتوقع أن يكون متوسط العينات متساوياً تماماً ولكن نتوقع أن يكون متراجحاً حول قيمة ما هي قيمة المتوسط لكل القطع المنتجة. ولكن هل شكل هذه المتوسطات أي توزيعها سيكون منتظاماً؟ لا إنه يتبع التوزيع الطبيعي. لماذا؟ لأنك لو أخذت العينات بشكل عشوائي فإن متوسطها سيكون متراجحاً حول متوسط كل العينات أو كل القطع المنتجة خلال عدة أيام.

في هذا المثال افترض أن كل عينة تتكون من ثلاثة قطع فقط. لنجرب معاً بعض العينات الافتراضية:

العينة الأولى: 11، 15، 18. المتوسط = 14.7

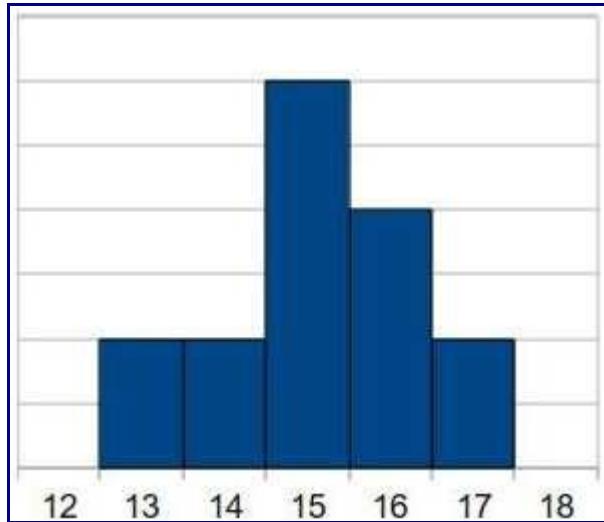
العينة الثانية: 10، 12، 18. المتوسط = 13.3

العينة الثالثة: 13، 16، 19. المتوسط = 16

العينة الرابعة: 11، 14، 20. المتوسط = 15

العينة الخامسة: 17، 13، 16. المتوسط = 15.3
 العينة السادسة: 10، 17، 18. المتوسط = 15
 العينة السابعة: 15، 13، 11. المتوسط = 13
 العينة الثامنة: 18، 17، 15. المتوسط = 16.7

افرض أن نفس النتائج تقريباً تكررت لمدة أربعة أيام. دعنا نرسم المدرج التكراري لمتوسط هذه العينات.



ألا يذكرك هذا المنظر بمنحنى التوزيع الطبيعي؟ ربما ليس دقيقاً ولكنه قريب منه لقد استخدمنا 3 قطع في كل عينة ولو استخدمنا 5 لاقربنا أكثر من التوزيع الطبيعي ولو استخدمنا 10 لاقربنا كثيراً جداً من التوزيع الطبيعي ولو استخدمنا 30 فإننا نصل إلى التوزيع الطبيعي مهما كان التوزيع الأصلي.

ما الذي يحدث؟ إنك عندما تنتقي العينات بشكل عشوائي فإنك تأخذ من أقصاها ومن أدناها فيكون المتوسط قريباً من المتوسط العام. وهذا فإن العشوائية في الاختيار تضمن أن يكون متوسط العينة قريباً من المتوسط العام. وقد يحدث أحياناً أن تكون عينة مائلة ناحية اليمين أو اليسار ولكن هذا يكون قليلاً. وهذا هو نفس وصف منحنى التوزيع الطبيعي حيث تتوزع معظم البيانات حول المتوسط وتوجد بعض القيم المتطرفة يميناً ويساراً.

هذا الأمر يحدث مهما كان التوزيع الأصلي للبيانات. ولكن قد نصل للتوزيع الطبيعي مع صغر العينة وقد نحتاج لعينة كبيرة لنصل للتوزيع الطبيعي. والمقصود بحجم العينة هنا هو عدد مفرداتها أي عدد القطع التي نقيسها في المثال السابق. ونظراً لأن القياس عن طريق العينات هو أمر شائع فإننا نستطيع استخدام المنحنى الطبيعي كثيراً.

في حالة ما إذا توزيع البيانات الأصلية يتبع التوزيع الطبيعي فمن البديهي أن توزيع متوسط أي عينات سيكون توزيعاً طبيعياً مهماً صغر حجم العينة. أما إذا لم يكن توزيع البيانات الأصلية يتبع التوزيع الطبيعي فإن توزيع متوسط العينات يتبع التوزيع الطبيعي إذا كان حجم العينات حوالي 15 إذا كان التوزيع الأصلي متماضٍ أو 30 على الأقل إذا كان التوزيع الأصلي غير متماضٍ مثل التوزيع الأسوي.

ومتوسط هذه العينات يساوي متوسط مجتمع الدراسة أما الانحراف المعياري لمتوسط العينات فيساوي الانحراف المعياري لمجتمع الدراسة مقسوماً على الجذر التربيعي لحجم العينات. ولهذه العلاقات بعض الاستخدامات الإحصائية التي قد نستعرضها في مناسبة أخرى إن شاء الله.

وهذه النظرية تساعدنا على استخدام **خراط المراقبة Control Chart** والتي تعتمد على منحنى التوزيع الطبيعي. ويمكننا الآن أن نعرف لماذا يتم رسم الحد الأدنى والأعلى على بعد (3 * الانحراف المعياري) من المتوسط ولماذا

يتم الاعتماد على خصائص منحنى التوزيع الطبيعي دون غيره في الحكم على خرائط المراقبة. وهذا هو موضوع عن الذي بدأناه في مقالة سابقة ويمكناه الآن أن نستكمله بمشيئة الله في المقالات التالية. ولكن علينا ألا ننسى أن نظرية الحد المركزية يتم تطبيقها في تحاليل إحصائية كثيرة خلاف خرائط المراقبة. كما أنه جدير بالإشارة إلى أننا لا نلغى التوزيعات الأخرى فقد تحتاجها في بعض الأحيان وبعض الاختبارات الإحصائية وبعض النظريات مثل استخدام التوزيع الأسوي في نظرية الطوابير واستخدام التوزيعات المختلفة في المحاكاة.

[من مراجع المقالة:](#)

Applied Statistical Methods, W. Carlson and B. Thorne, Prentice Hall, 1997

Statistics for Managers, Levine et al., Prentice Hall, 1999

The Six Sigma Handbook, Pyzdek, McGrawHill, 2003

The Lean Six Sigma Pocket ToolBook, George at al., McGrawHill, 2005