الإحصاء الحيوي

لطلاب السنة الثانية والثالثة - كلية العلوم الصحية

جامعة دمشق

”الجلسة الرابعة“

مدرس المقرر: أ. سلام الجرّاح

إعداد: د. ياسر الزعيم

المدرس في قسم الإحصاء الرياضي

كلية العلوم - جامعة دمشق

نيسان– 2023

نتكلم اليوم عن:

- تذكرة بما تحدثنا عنه
- **البحث الثالث: مقاييس النزعة المركزية**
- مقاييس النزعة المركزية في SPSS

تذكرة بما تحدثنا عنه

label	شدة الإصابة
1	ضعيفة جداً
2	ضعيفة
3	متوسطة
4	قوية
5	قوية جداً

مثال: شدة الإصابة العضلية لدى لاعبي فريق ليفربول.

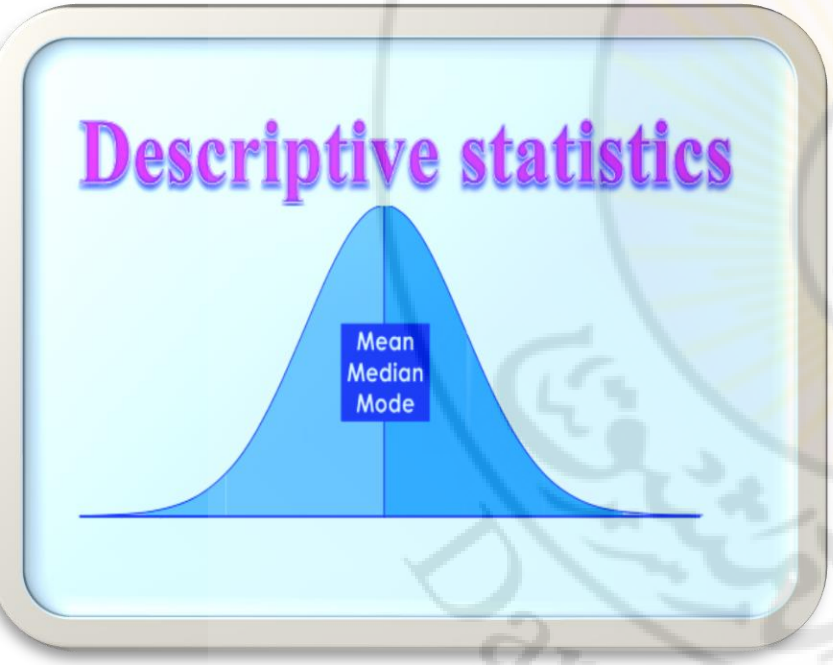
مثال آخر: سرعة استجابة المريض لمحفز ضوئي (مقاسة بالثانية):

$$3, 3, 4, 3, 8, 9, 8, 7, 3, 10, 20, 4$$

وهو الجدول التكراري

X_i	f_i
3	4
4	2
7	1
8	2
9	1
10	1
20	1
Σ	12

مقاييس النزعة المركزية Central Tendency Measures
أو Central Tendency Measurements (اختصاراً CTM)



وبعض الكتب تسميها:

Measures of Central Location

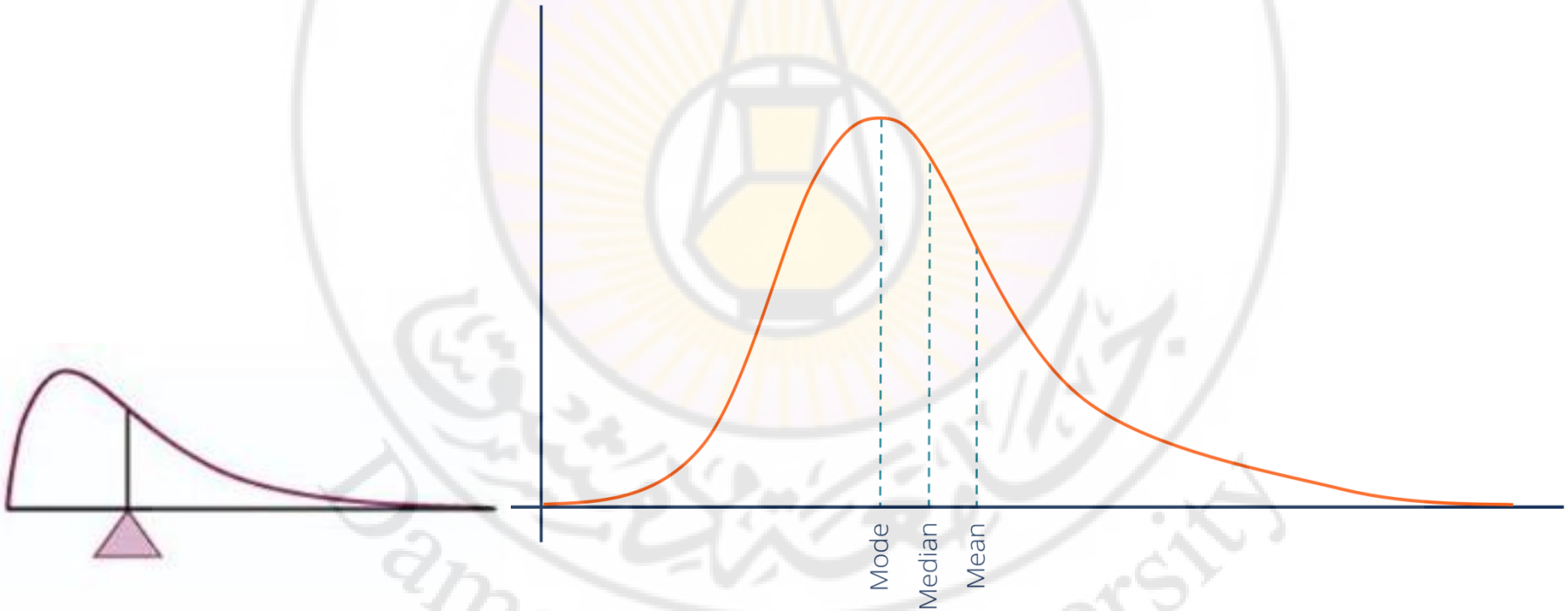
وهي أعداد (أو مؤشرات) تدل على المكان الذي يتجمع عنده معظم مشاهدات عينة.

بالتالي هي محسوبة من المشاهدات نفسها وتتوضع قريباً من مركز المشاهدات (أو هي المركز ذاته).

البحث 3: مقاييس النزعة المركزية

أولاً: المتوسط الحسابي Average or Arithmetic Mean or Mean

وهو القيمة القريبة من مركز المشاهدات و هو يسمى أيضاً بـ مركز الثقل (مركز التوازن) balancing point.



البحث 3: مقاييس النزعة المركزية

أ) المتوسط الحسابي للبيان الخام raw data

إن المتوسط هو $\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$ ونقرؤه $X\ bar$ $X_1, X_2, \ldots, X_N$

بالعودة **لمثال** زمن الاستجابة لمحفز ضوئي إنّ:

$$\bar{X} = \frac{3+3+4+3+8+9+8+7+3+10+20+4}{12} = 6.83$$

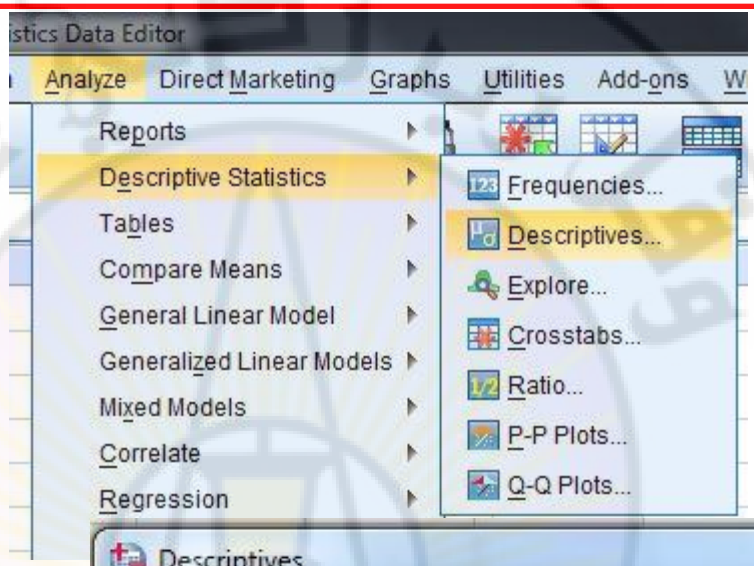
في SPSS

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
VAR00001	12	3.00	20.00	6.8333	4.91442
Valid N (listwise)	12				

وذلك كما يلي:

البحث 3: مقاييس النزعة المركزية



ب) المتوسط الحسابي للبيان المرتب في جدول تكراري:

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i}$$

$$\bar{X} = \frac{\sum f_i X_i}{N}$$

X_i	f_i
X_1	f_1
X_2	f_2
$\vdots$	$\vdots$
X_k	f_k

البحث 3: مقاييس النزعة المركزية

مثال: سرعة استجابة المريض لمحفز ضوئي (مقاسة بالثانية):

X_i	f_i
3	4
4	2
7	1
8	2
9	1
10	1
20	1
Σ	12

يصبح الجدول:

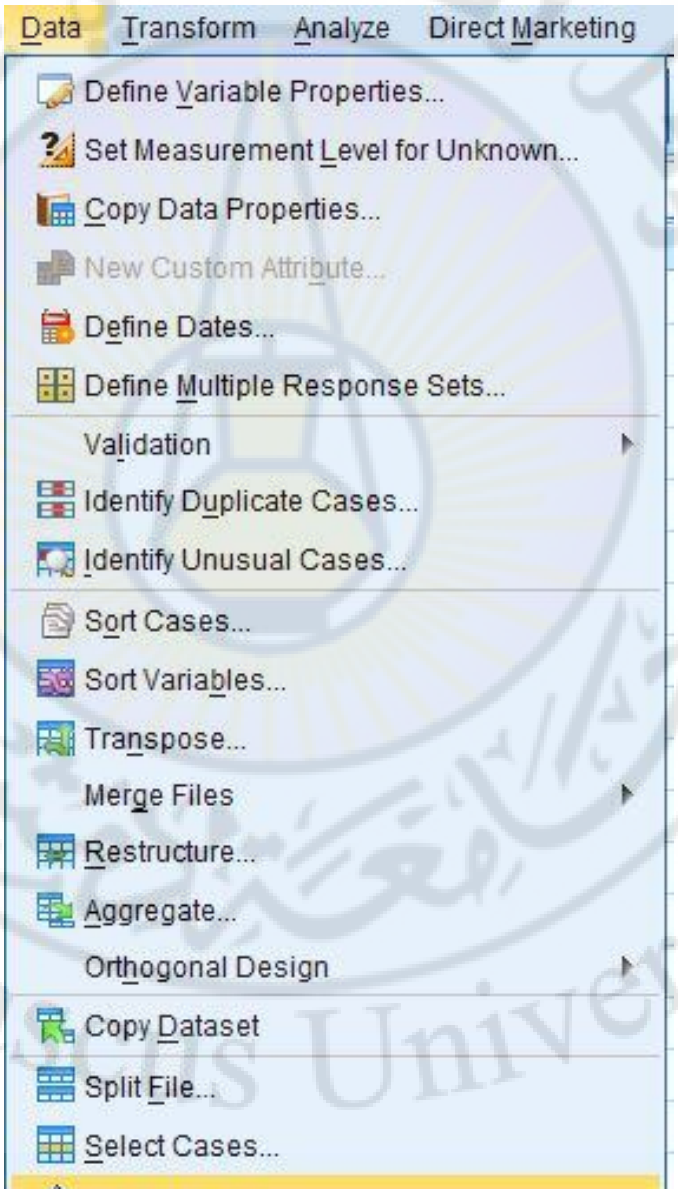
X_i	f_i	$X_i f_i$
3	4	12
4	2	8
7	1	7
8	2	16
9	1	9
10	1	10
20	1	20
Σ	12	82

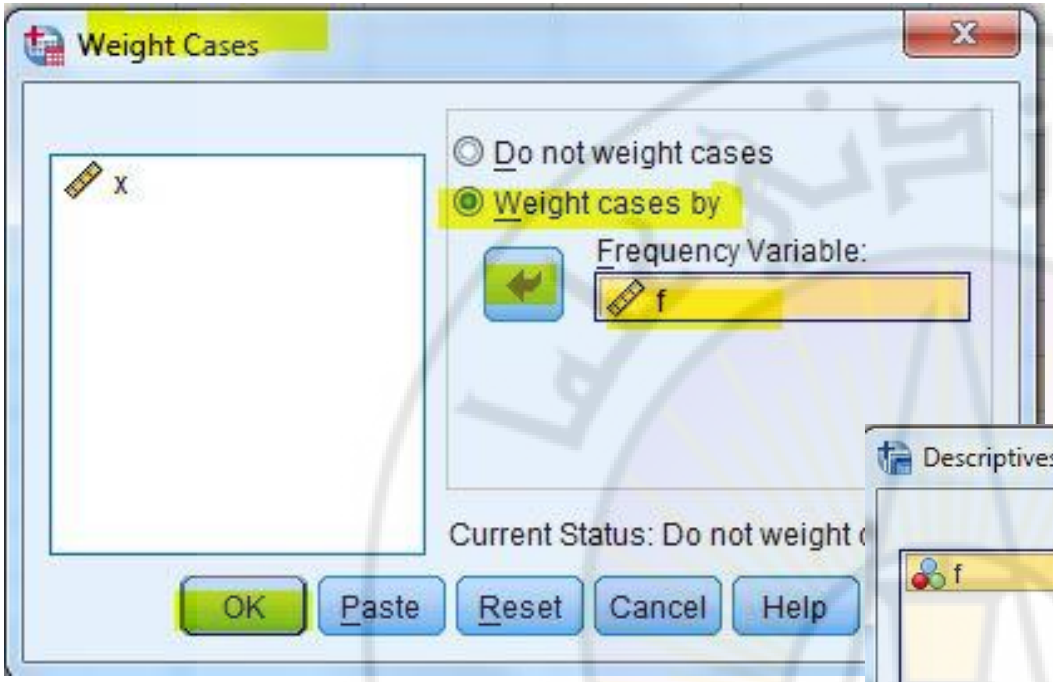
$$\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i} = \frac{82}{12} = 6.833$$

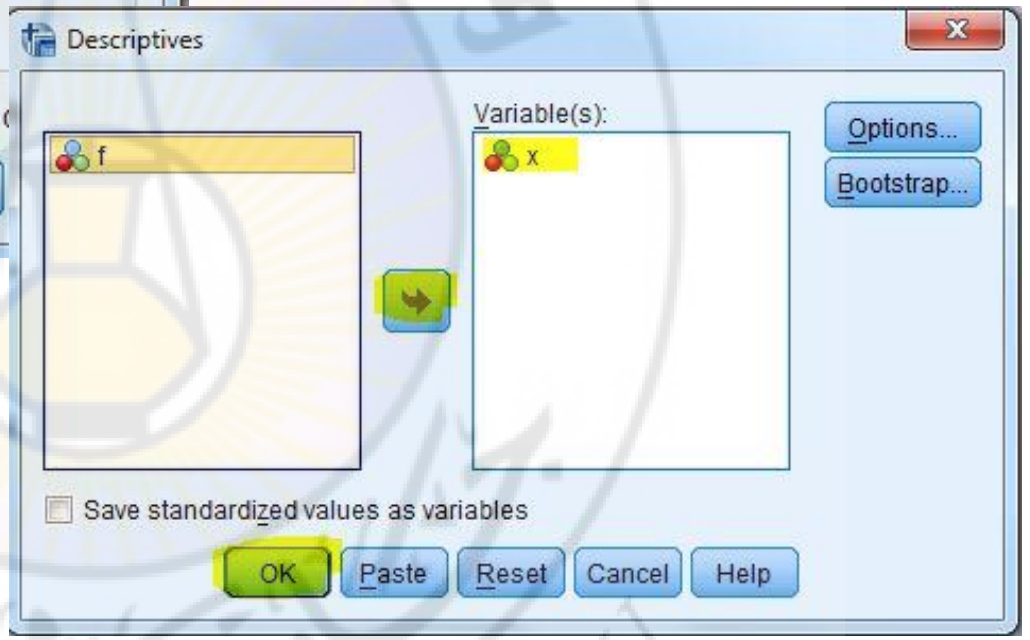
البحث 3: مقاييس النزعة المركزية

	x	f
1	3.00	4.00
2	4.00	2.00
3	8.00	2.00
4	9.00	1.00
5	7.00	1.00
6	10.00	1.00
7	20.00	1.00

في SPSS كما يلي:







Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
x	12	3.00	20.00	6.8333	4.91442
Valid N (listwise)	12				

البحث 3: مقاييس النزعة المركزية

Kids weight_extended.sav

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
weight	209	24.00	53.00	39.6459	5.75535
Valid N (listwise)	209				

ثانياً الوسط Median

وهي القيمة التي تتوضع في مركز المشاهدات أي هي منتصف المشاهدات.

عينة عشوائية N من المتغير X: $X_1, X_2, \ldots, X_N$. لحساب الوسط $med(X)$

نرتب تصاعدياً (من الأصغر إلى الأكبر) فيصبح $X_{(1)} \le X_{(2)} \le \ldots \le X_{(N)}$:

أ) إذا N فردي odd فالوسط هو المشاهدة التي ترتيبها $\frac{N+1}{2}$، أي:

$$med(X) = X_{\left(\frac{N+1}{2}\right)}$$

ب) إذا N زوجي even فالوسط هو متوسط المشاهدتين اللتين ترتيبهما $\frac{N}{2}, \frac{N}{2}+1$،

أي:

$$med(X) = \frac{X_{\left(\frac{N}{2}\right)} + X_{\left(\frac{N}{2}+1\right)}}{2}$$

البحث 3: مقاييس النزعة المركزية

بالعودة **لمثال** سرعة الاستجابة: $3,3,4,3,8,9,8,7,3,10,20,4$

البيان الخام: $X_1 = 3, X_2 = 3, X_3 = 4, \ldots, X_1 = 3, X_{12} = 4$

البيان المرتب: ordered data

$$X_{(1)} = 3 \le X_{(2)} = 3 \le X_{(3)} = 3 \le X_{(4)} = 3 \le X_{(5)} = 4 \le X_{(6)} = 4 \le$$
$$X_{(7)} = 7 \le X_{(8)} = 8 \le X_{(9)} = 8 \le X_{(10)} = 9 \le X_{(11)} = 10 \le X_{(12)} = 20$$

وكون $N = 12$ فإن الوسط هو:

$$med(X) = \frac{X_{\left(\frac{12}{2}\right)} + X_{\left(\frac{12}{2}+1\right)}}{2} = \frac{X_{(6)} + X_{(7)}}{2} = \frac{4+7}{2} = 5.5 \text{ (sec.)}$$

ثالثاً المنوال Mode

وهي القيمة الأكثر تكراراً من بين المشاهدات ونرمز له $mod(X)$.

مثال: سرعة الاستجابة لمحفز ضوئي

إنّ المشاهدة (القيمة) $X = 3$ مكررة أربع مرات وهي أعلى تكراراً من باقي المشاهدات و لذا $mod(X) = 3$

X_i	f_i
3	4
4	2
7	1
8	2
9	1
10	1
20	1
Σ	12

ملاحظة أ:

لا يوجد منوال عندما يكون لكل المشاهدات نفس التكرار.

مثال: $1,1,2,2,4,4,7,7,35,35,20,20$

لاحظ كل مشاهدة مكررة مرتين لذا إن هذا البيان الإحصائي لا يحوي منوال.

البحث 3: مقاييس النزعة المركزية

ملاحظة ب: نفس التكرار لمشاهدتين (أو أكثر، وليس الكل)

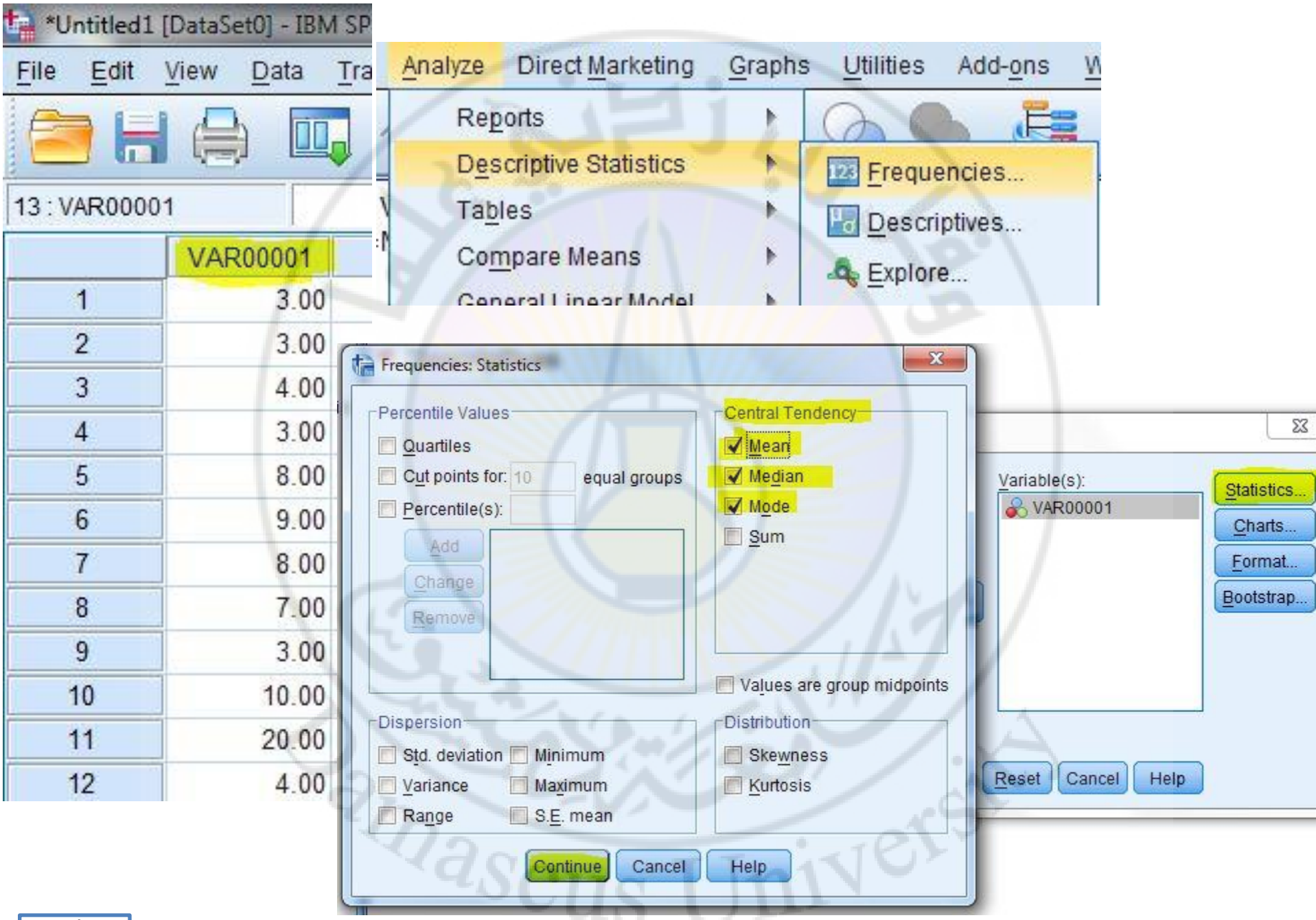
- **المشاهدتان متجاورتان: المنوال هو متوسط هاتين المشاهدتين.**
- **المشاهدتان متباعدتان: يوجد منوالان.**

مثال: $1,2,2,3,3,3,6,6,6,12,12,13$ إن المشاهدة 3 والمشاهدة 6 لهما نفس التكرار و لاحظ أنهما (بعد ترتيب المشاهدات تصاعدياً) متجاورتان

ولذا يكون المنوال $mod(X) = \frac{3+6}{2} = 4.5$

مثال: $1,2,2,3,3,3,6,6,12,12,13,13,13$ إن القيمة 3 والقيمة 13 لهما نفس التكرار و هما متباعدتان و بالتالي يوجد منوالان لهذا البيان الإحصائي

$$mod_1(X) = 3\ ,\ mod_2(X) = 13$$


*Untitled1 [DataSet0] - IBM SP
File Edit View Data Tra
13 : VAR00001
VAR00001
1 3.00
2 3.00
3 4.00
4 3.00
5 8.00
6 9.00
7 8.00
8 7.00
9 3.00
10 10.00
11 20.00
12 4.00
Analyze Direct Marketing Graphs Utilities Add-ons W
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Frequencies...
Descriptives...
Explore...
Frequencies: Statistics
Percentile Values
Quartiles
Cut points for: 10 equal groups
Percentile(s):
Add
Change
Remove
Central Tendency
Mean
Median
Mode
Sum
Values are group midpoints
Dispersion
Std. deviation
Minimum
Variance
Maximum
Range
S.E. mean
Distribution
Skewness
Kurtosis
Continue
Cancel
Help
Variable(s):
VAR00001
Statistics...
Charts...
Format...
Bootstrap...
Reset
Cancel
Help

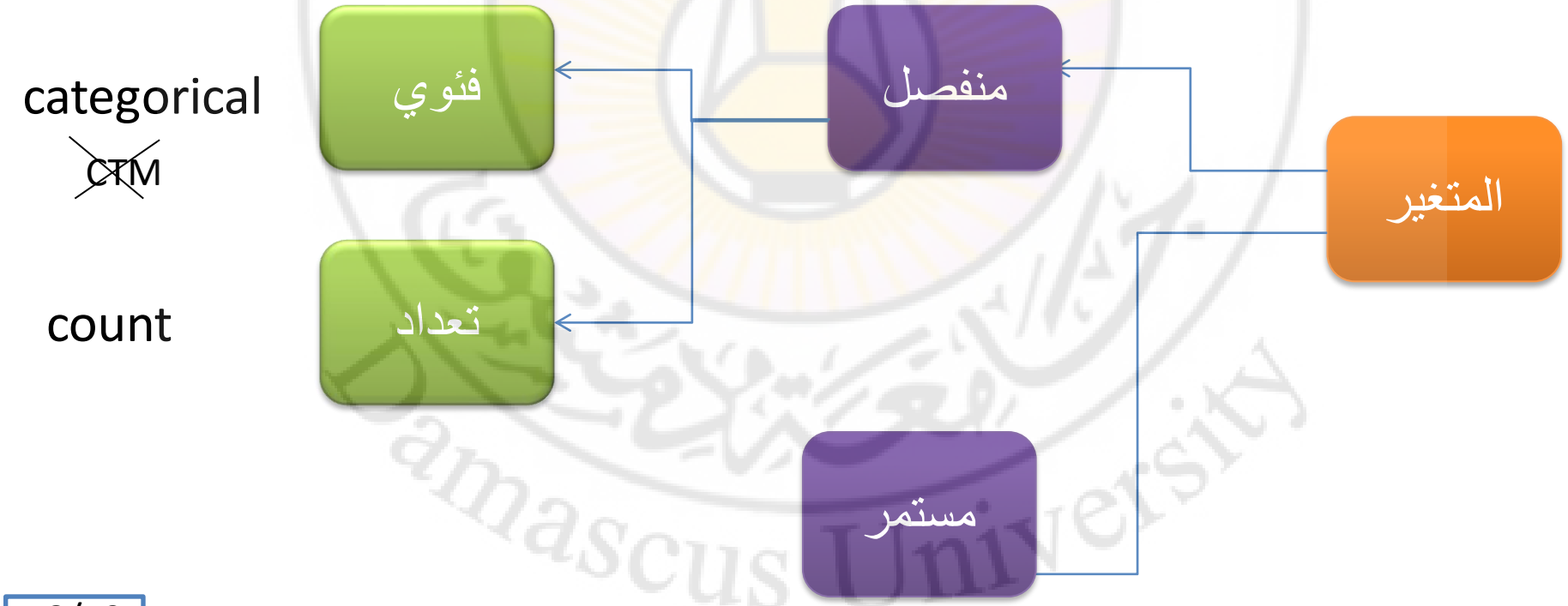
البحث 3: مقاييس النزعة المركزية

Statistics

VAR00001

N	Valid	12
	Missing	0
Mean		6.8333
Median		5.5000
Mode		3.00

المتغيرات بلغة SPSS



عدد الفئات الأمثل لتمثيل المتحول المستمر فئوياً

Herbert Sturges 1882-1958

توجد محاولات كثيرة لاختيار عدد الفئات المناسب $\mathbf{I}$ لتقسيم مشاهدات عينة حجمها N من متحول مستمر.
ومن أشهر تلك المحاولات:

1) قانون سترجس

$$\mathbf{I} = 1 + 3.322 \log(N)$$

مثال: Bacteria.sav إنّ $\mathbf{I} = 14.288$
لذا نحتاج 15 فئة لتمثيل المتحول data فئوياً.

2) قانون رايس $\mathbf{I} = 2\left\lceil \sqrt[3]{N} \right\rceil$

مثال: Bacteria.sav إنّ: $\mathbf{I} = 2\left\lceil \sqrt[3]{10000} \right\rceil = 2 * \lceil 21.54 \rceil = 2 * 22 = 44$
لذا نحتاج 44 فئة وفق قاعدة Rice.

الإحصاء الحيوي

لطلاب السنة الثانية والثالثة - كلية العلوم الصحية

جامعة دمشق

" الجلسة الأولى"

مدرس المقرر: أ. سلام الجرّاح

إعداد: د. ياسر الزعيم

المدرس في قسم الإحصاء الرياضي

كلية العلوم - جامعة دمشق

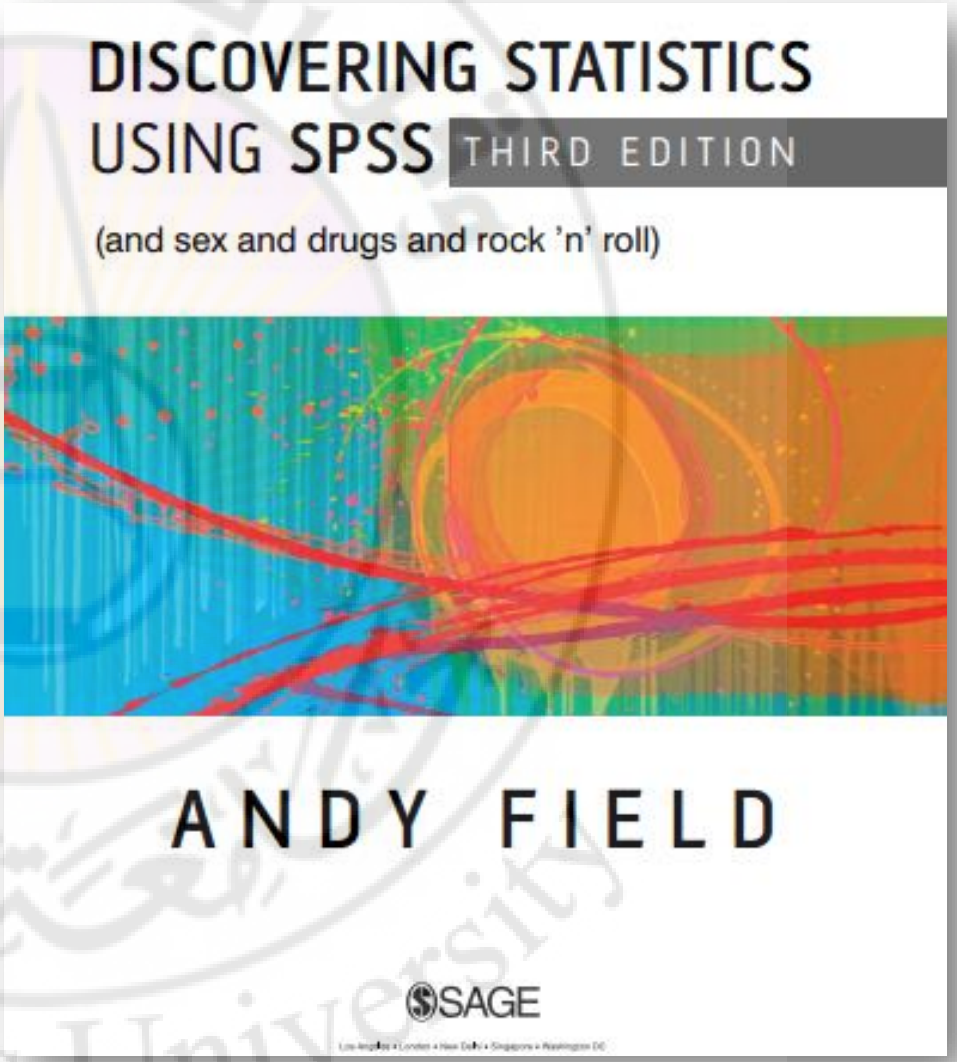
آذار– 2023

نتكلم اليوم عن:

- مقدمة
- علم الإحصاء
- أهمية علم الإحصاء في ميادين البحث العلمي
- الاستدلال الإحصائي
- بعض الموضوعات التي ندرسها

المرجع

Field, A. (2009). Discovering statistics using SPSS, *3rd Edition*, SAGE.



هو فن جمع البيانات وتلخيصها وتحليلها بالطريقة الصحيحة الهادفة إلى فهم ظاهرة معينة، ومن ثم بناء توقعات مستقبلية حول هذه الظاهرة.

مثل:

نمذجة والتنبؤ بمعدل انتشارومعدل وفيات مرض ما في العالم عن طريق بناء نماذج إحصائية.

الاستدلال الإحصائي

الاستدلال (أو الاستقراء أو الاستنباط) الإحصائي هو العلم الذي يعتني ببناء الأحكام والقرارات الكلية الشاملة وذلك بالاعتماد فقط على معلومات جزئية.

أقسام الاستدلال الإحصائي:

1. نظرية التقدير Estimation Theory
2. اختبار الفرضيات Hypotheses Testing

الموضوعات التي ندرسها

البحث 1: مفاهيم أساسية في علم الإحصاء
البحث 2: التوزيعات التكرارية و ترسيم البيانات
البحث 3: مقاييس النزعة المركزية
البحث 4: مقاييس التشتت
البحث 5: المتغير العشوائي و بعض التوزيعات
البحث 6: الانحدار و الارتباط

بالإضافة إلى بحوث أخرى

البحث 1: مفاهيم أساسية في علم الإحصاء

هدف الإحصاء الرياضي هو وصف و تفسير نتائج متعلقة بالخصائص العددية للمجتمع (المجتمع الإحصائي) من خلال تحليل مشاهدات ذاك المجتمع.

أولاً المجتمع (المجتمع الإحصائي)

مجموعة من الأشياء والمواد والنبات وأي شيء يخطر في البال من موجودات.

(1) المجتمع المحدود finite population

(2) المجتمع غير المحدود infinite population

ثانياً الخصائص العددية للمجتمع numerical properties

مثل: عمر الشخص، الوزن، لون العينين، ضغط الدم، طول القامة، الجنس، ...

ثالثاً العينة العشوائية random sample

هي مجموعة جزئية (صغيرة) من المجتمع المدروس، يتم انتقاءها بطرق إحصائية علمية وتسمى اصطلاحاً بالعينة العشوائية.

أهمية العينة: تقليل التكلفة البدنية والزمنية والاقتصادية مقارنة بدراسة المجتمع الإحصائي بأكمله.

رابعاً المعلَمة و المقدّر

المعلمة parameter هي خاصة تصف المجتمع بينما **المقدر estimate** فهو خاصة تصف العينة العشوائية المسحوبة منه.

المتوسط الحسابي $\bar{x}$ للعينة هو مقدر للمتوسط الحسابي μ للمجتمع

الانحراف المعياري s للعينة هو مقدر للانحراف المعياري σ للمجتمع.

خامساً المتغير variable

هو الخصيصة (الصفة) الدالة على أن عناصر مجموعة ما تختلف فيما بينها.

مثل متغير الجنس في عينة تحتوي على الذكور والإناث

سادساً الثابت constant

هو الخصيصة الدالة على أن عناصر مجموعة ما لا تختلف أي اختلاف فيما بينها.

مثل متغير الجنس في عينة تحتوي فقط ذكوراً

سابعاً العنونة (التسمية) label

ترقيم مفترض لتوصيف طبقات مجموعة تختلف عناصرها فيما بينها.

مثال:

Identity (id)	gender
1	Male
2	Male
3	Male
4	female

1 = Male
2 = Female

Identity (id)	gender
1	1
2	1
3	1
4	2

ثامناً أنواع المتغيرات

تنقسم المتغيرات وفق نوع قيمها إلى نوعين (مستمر و منفصل) بينما تنقسم

المتغيرات وفق طريقة قياسها إلى أربعة أقسام (اسمي، ترتيبي، مجالي ونسبي).

1) المتغير المستمر و المتغير المنفصل

المتغير مستمر continuous إذا كانت مقاديره غير قابلة للعد، مثل متغير العمر.

المتغير منفصل (متقطع) discrete إذا كانت مقاديره قابلة للعد، مثل متغير الجنس.

2) المتغيرات الاسمية والترتيبية والمجالية والنسبية

المتغير الاسمي nominal variable وهو المتغير الذي تختلف مقاديره فيما بينها اسمياً ولا يمكن ترتيبها بشكل ذي معنى.

مثال: لون العينين: بني – عسلي – أزرق – أخضر – أسود

مثال: لون العينين مع تسمية:

اللون	label
بني	1
عسلي	2
أزرق	3
أخضر	4
أسود	5

المتغير الترتيبي (الرتبي) ordinal variable وهو المتغير الذي تختلف مقاديره فيما بينها اسمياً وصفةً بحيث أن الترتيب له معنى ولكن لا معنى لتساوي المسافات (الفروقات) بين قيمه.

مثال: شدة الإصابة بمرض نفسي ما:

label	شدة الإصابة
1	ضعيفة جداً
2	ضعيفة
3	متوسطة
4	قوية
5	قوية جداً

إن القيمة 1 أصغر من القيمة 2، كذلك إن شدة المرض 5 أكبر من شدة المرض 3 وهكذا...
ولكن: لا نستطيع القول أن الفرق (المسافة) بين 1 و2 هي نفسها الفرق (المسافة) بين 3 و4.
كذلك لا نستطيع القول أن 4 هي ضعف القيمة 2، فليس لهذا الضعف أي معنى في هذا المثال.

البحث 1: مفاهيم أساسية في علم الإحصاء

المتغير المجالي interval variable هو المتغير الذي تختلف مقاديره اختلافاً اسمياً وصفةً ترتيبيةً ويوجد معنى لتساوي الفروقات بين قيمه ولكنه لا يحتوي الصفر الحقيقي.

مثال: بفرض درجات الحرارة (درجة مئوية) : 12 و24 و36
إن درجة الحرارة 24 هي أعلى من الدرجة 12، وإن الدرجة 24 هي أقل من 36 لذا الترتيب موجود وله معنى. كما أن الفرق بين الدرجة 24 و12 يساوي 12 درجة مئوية وهو يساوي الفرق بين الدرجة 36 والدرجة 24.

ولكن: لا يمكن القول بأن درجة الحرارة 36 هي ثلاثة أضعاف درجة الحرارة 12، لأنه في هذا المثال لا يوجد الصفر الحقيقي (أو المسمى بالصفر المطلق) true (or pure) zero.

فما هو الصفر المطلق؟

الصفر الحقيقي (أو المسمى بالصفر المطلق أو الصفر البحت): هو الصفر الذي يدل على انتفاء وانعدام الصفة تماماً، فمثلاً إذا قلنا أن وزن الشخص هو صفر فهذا يعني أن الوزن غير موجود تماماً.

بالعودة إلى **مثال** درجات الحرارة، إذا قلنا أن درجة الحرارة هي صفر، فهذا لا يعني أبداً انعدام الحرارة، بل الحرارة موجودة وهي صفرية.

مثال آخر: في التقويم الميلادي نحن اليوم نعيش في تاريخ 17/11/2020، لاحظ أن الفرق بين 17/11 و 20/11 هو ثلاثة أيام وهو يساوي الفرق بين اليوم 1/11 و4/11 لذا التقويم الميلادي هو متغير فوق ترتيبي، أي هو متغير مجالي. الآن هل هذا المتغير يحتوي الصفر المطلق؟

الجواب لا، لأن تاريخ 0/11/2020 يعني أننا في اليوم 31/10/2020 (أو غير ذلك) وبالتالي إن الصفر لم ينفي صفة التاريخ ولم يختفي التاريخ لذا هو صفر غير حقيقي.

البحث 1: مفاهيم أساسية في علم الإحصاء

المتغير النسبي ratio variable هو المتغير الذي تختلف قيمه اسماً وترتيباً وهناك معنى للفروقات بين قيمه ويشتمل على الصفر الحقيقي.

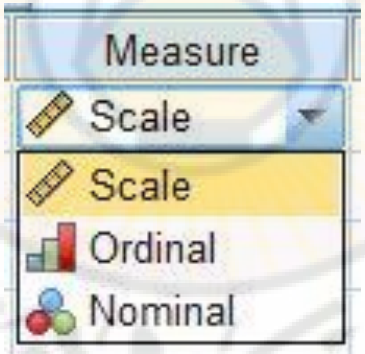
مثال: علامات الطلاب: 20, 60, 80, 90, 95, 100 إن العلامة 60 أكبر من العلامة 20. إن الفرق بين العلامتين 90 و80 هو 10ويساوي الفرق بين العلامتين 80 و90.

إن العلامة 60 هي ثلاثة أضعاف العلامة 20 وكذلك كمثال إن العلامة 20 هي ربع العلامة 80. وهذا صحيح فقط وفقط لأن الصفر الحقيقي موجود، فلو فرضنا أن أحد الطالب علامته صفر أي صفر كامل محض فهذا يعني أنه ليست هناك أية فرصة وأية حظوظ ليكون لهذا الطالب علامة أخرى.

ملاحظة ختامية إن المتغير المجالي والمتغير النسبي يتم تسميتهما بالمتغير المقياسي (القياسي) scale، وذلك اختصاراً و تسهيلاً للدراسة.

مثلاً في برنامج SPSS تم إدراج الأنواع:

الاسمي- الترتيبي- المقياسي



SPSS (v27), ..., SPSS (V21), ...

إصدار ليس أقدم من SPSS (V18).

الإحصاء الحيوي

لطلاب السنة الثانية والثالثة - كلية العلوم الصحية

جامعة دمشق

”الجلسة الخامسة“

مدرس المقرر: أ. سلام الجرّاح

إعداد: د. ياسر الزعيم

المدرس في قسم الإحصاء الرياضي

كلية العلوم - جامعة دمشق

آيار– 2023

نتكلم اليوم عن:

- **البحث الرابع: مقاييس التشتت**
- مقاييس التشتت في SPSS

البحث 4: مقاييس التشتت

مقاييس التشتت (التبعثر)

Dispersion (Scatter, Variability or Spread) Measures

هي أعداد غير سالبة، تساوي الصفر إذا كانت جميع المشاهدات متساوية بالقيمة، وتكبر هذه الأعداد كلما اختلفت وتباعدت المشاهدات عن بعضها.

أهمها:

المدى- التباين- الانحراف المعياري

البحث 4: مقاييس التشتت

أولاً: المدى Range

وهو الفرق بين أكبر قيمة من بين المشاهدات وأصغرها أي:

$$range(X) = \max_i(X_i) - \min_i(X_i)$$

مثال سرعة استجابة المريض لمحفز ضوئي (مقاسة بالثانية):

$$3, 3, 4, 3, 8, 9, 8, 7, 3, 10, 20, 4$$

$$range(X) = \max_i(X_i) - \min_i(X_i) = 20 - 3 = 17 \ \text{sec}.$$

ثانياً: التباين Variance

مقياس يعبر عن كمية التشتت الموجودة في العينة، فكلما كبر بالمقدار كان مؤشراً على ابتعاد المشاهدات عن متوسط العينة، وكلما قل بالمقدار كان دليلاً على اقتراب المشاهدات من المتوسط.

البحث 4: مقاييس التشتت

إنّ **تباين المجتمع (population)**

$$\sigma^2 = \frac{\sum_{i=1}^{N_p} (X_i - \mu)^2}{N_p}$$

مجهول (في معظم الدراسات) لذا يتم تقديره عن طريق

تباين عينة (sample) مسحوبة من ذاك المجتمع

$$s^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N - 1}$$

مثال أوزان الأطفال:

إنّ $s^2 = 33.124 \ kg^2$

❖ في بعض المراجع إنّ

$$s^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}$$

ثالثاً: الانحراف المعياري

Standard Deviation (std., sd., std. deviation)

وهو الجذر التربيعي للتباين أي:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}{N-1}}$$

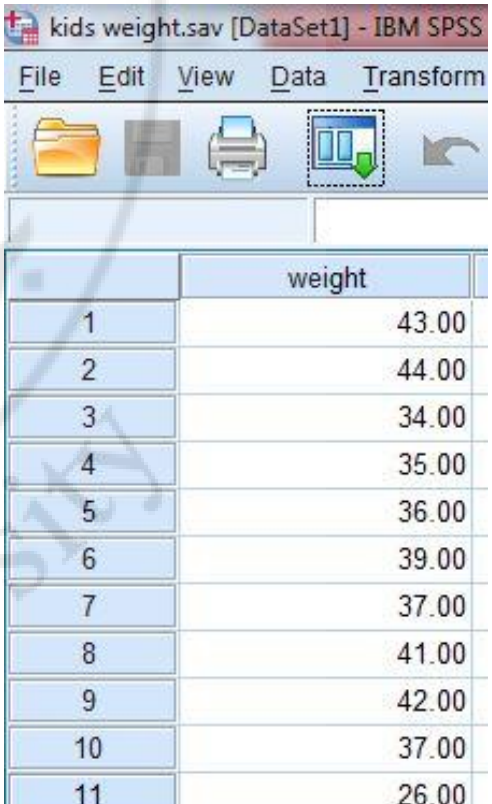
مثال أوزان الأطفال:

إنّ:

$$s^2 = 33.124 \ kg^2$$

وبالتالي يكون:

$$s = \sqrt{s^2} = 5.75 \ kg$$



	weight
1	43.00
2	44.00
3	34.00
4	35.00
5	36.00
6	39.00
7	37.00
8	41.00
9	42.00
10	37.00
11	26.00

البحث 4: مقاييس التشتت

تمرين:

لتكن مجموعة القياسات 50,35,70,90,40 التي تمثل درجات خمسة طلاب في مقررما في أحد الكليات، احسب التباين والانحراف المعياري.

الحل:

المتوسط:

$$\bar{x} = \frac{50+35+70+90+40}{5} = 57$$

وبالتالي التباين:

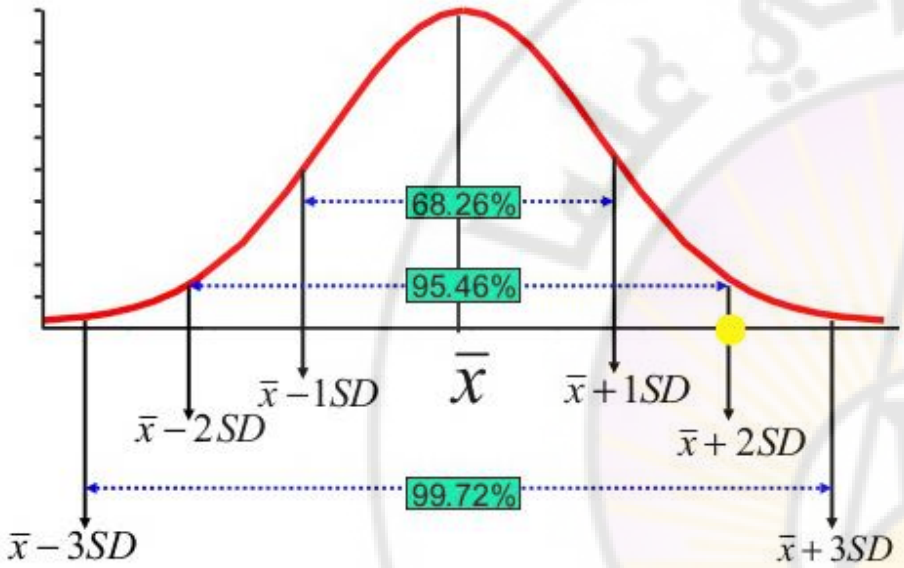
$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

$$= \frac{(49 + 484 + 169 + 1089 + 289)}{4} = 520$$

الانحراف المعياري:

$$s = \sqrt{s^2} = \sqrt{520} = 22.8$$

البحث 4: مقاييس التشتت



لندرس المثال التالي:

لنفرض بأن المنحني التكراري لمستوى ذكاء عينة من الأطفال كان بالشكل الأحمر. ولنفرض بأن متوسط مستوى الذكاء intelligence quotient (IQ) لدى مجموعة من الأطفال هو $\bar{X} = 100$ وبفرض أن الانحراف المعياري $s = 15$

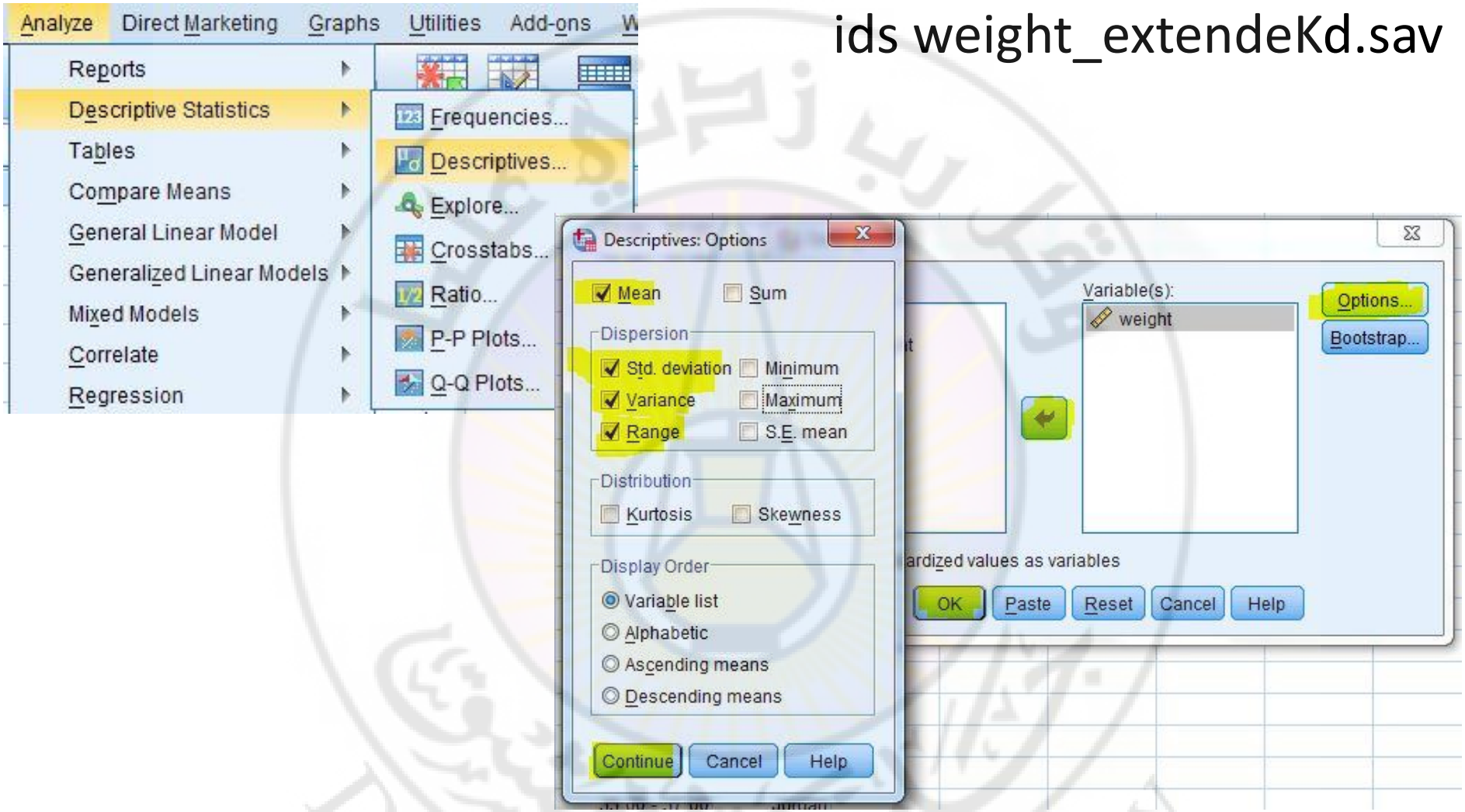
الآن لنفرض بأن طفلاً ما كان لديه مستوى الذكاء $X_i = 130$

عندئذ نستنتج بأن الطفل هذا له مستوى ذكاء يفوق المتوسط بما يعادل ضعفي الانحراف المعياري أي $X_i = \bar{X} + 2s$

الفائدة من هذا المثال:

إنّ الانحراف المعياري مهم لتشخيص تموضع المشاهدات بالنسبة للمتوسط.

ids weight_extendeKd.sav



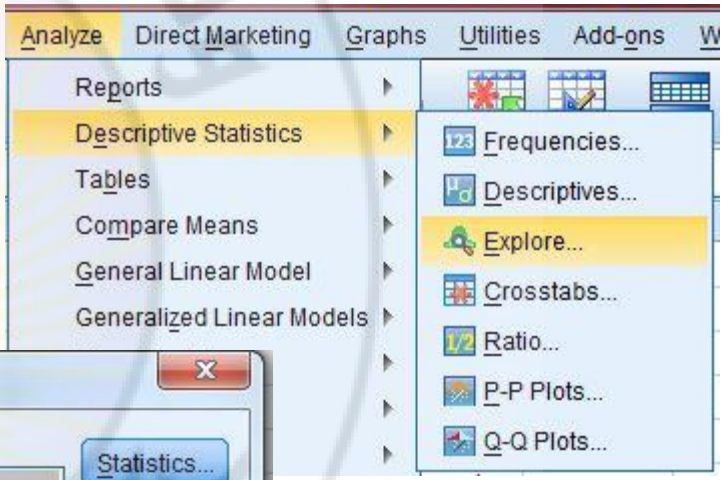
Descriptive Statistics

	N	Range	Mean	Std. Deviation	Variance
weight	209	29.00	39.6459	5.75535	33.124
Valid N (listwise)	209				

البحث 4: مقاييس التشتت

	weight	gender
1	43.00	boy
2	44.00	boy
3	34.00	girl
4	35.00	girl
5	36.00	girl
6	39.00	girl
7	37.00	girl
8	41.00	girl
9	42.00	girl

مقاييس النزعة المركزية والتشتت لأوزان الأطفال الذكور و الإناث على حدة



Descriptives

	gender			Statistic	Std. Error
weight	boy	Mean		40.4344	.53707
		95% Confidence Interval for Mean	Lower Bound	39.3712	
			Upper Bound	41.4977	
		5% Trimmed Mean		40.5000	
		Median		40.5000	
		Variance		35.190	
		Std. Deviation		5.93211	
		Minimum		24.00	
		Maximum		53.00	
		Range		29.00	
		Interquartile Range		8.00	
		Skewness		-.107	.219
		Kurtosis		-.346	.435
	girl	Mean		38.5402	.57219
		95% Confidence Interval for Mean	Lower Bound	37.4028	
			Upper Bound	39.6777	
		5% Trimmed Mean		38.5619	
		Median		39.0000	
		Variance		28.484	
		Std. Deviation		5.33702	
		Minimum		26.00	
		Maximum		49.00	
		Range		23.00	
		Interquartile Range		7.00	
		Skewness		-.101	.258
		Kurtosis		-.366	.511

البحث 4: مقاييس التشتت

ملاحظات حول الانحراف المعياري:

- إن ضرب جميع المشاهدات للمتحول بعدد موجب (أو سالب) سيؤدي إلى تضاعف قيمة الانحراف المعياري لهذا المتحول بمقدار القيمة المطلقة لذلك العدد.

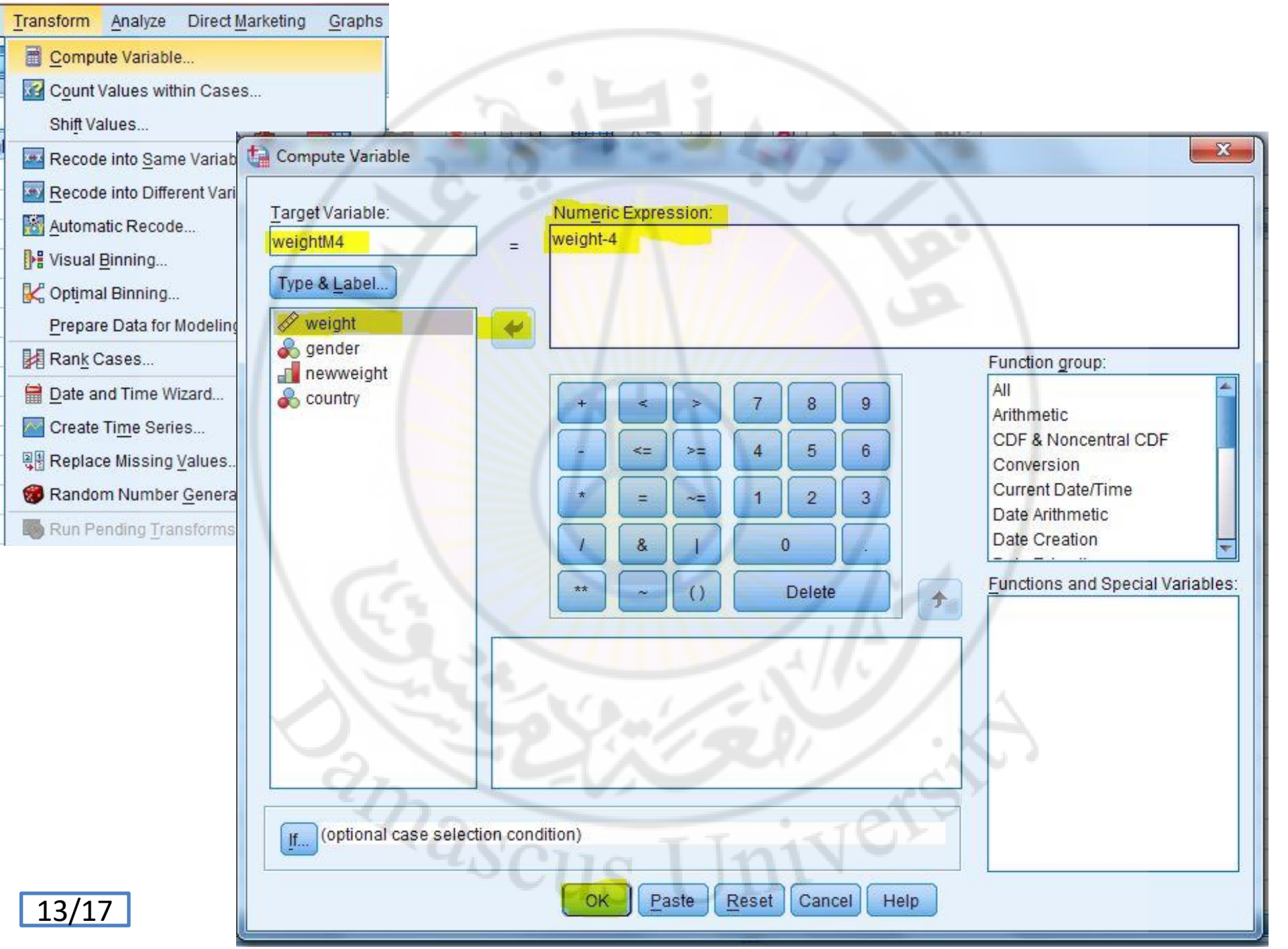
$$\forall i : X_i \longrightarrow cX_i \quad \Rightarrow \quad s \longrightarrow |c|s$$

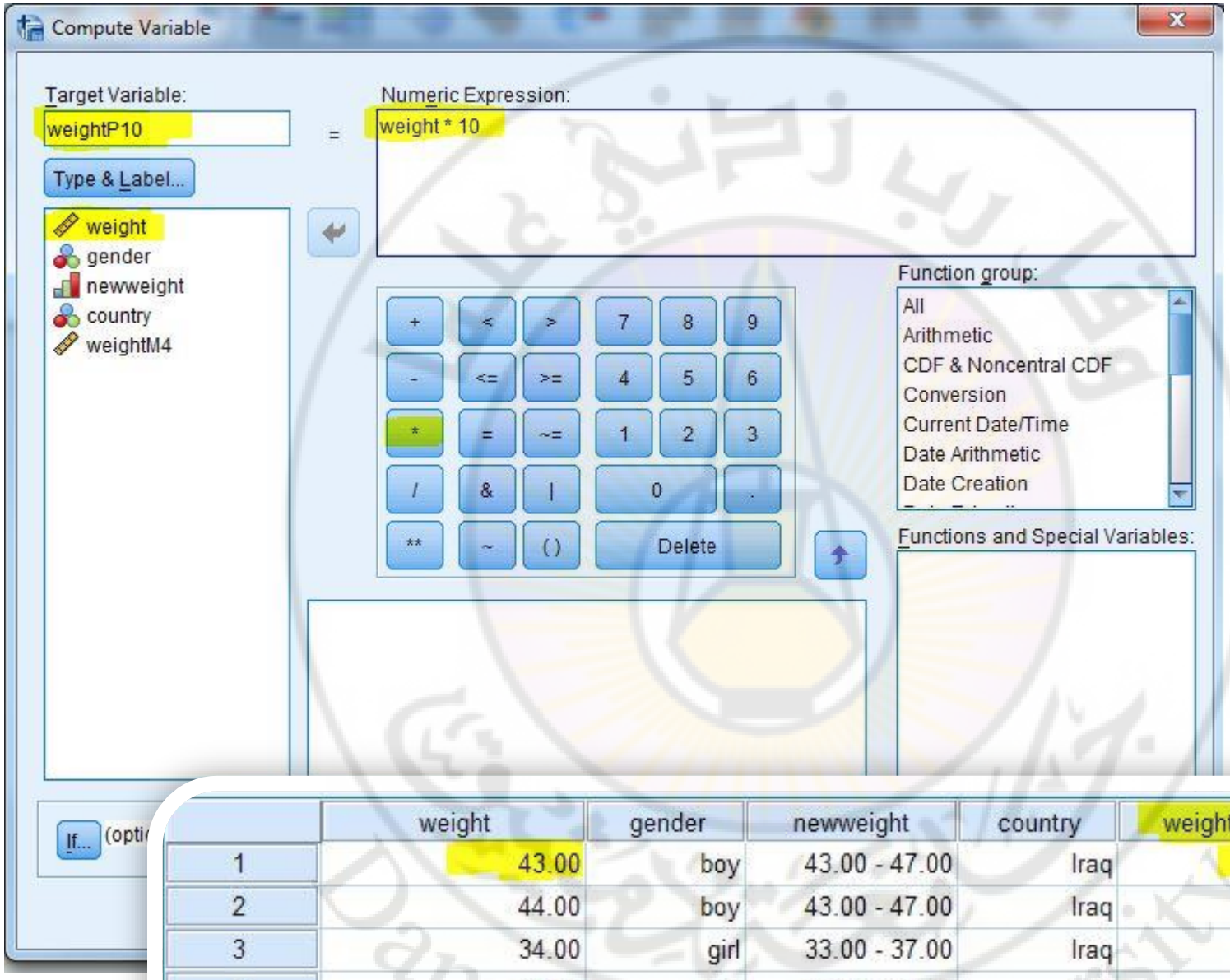
- إن جمع (أو طرح) عدد لجميع (من جميع) مشاهدات المتحول لن يغير من قيمة الانحراف المعياري لهذا المتحول.

$$\forall i : X_i \longrightarrow c + X_i \quad \Rightarrow \quad s \longrightarrow s$$

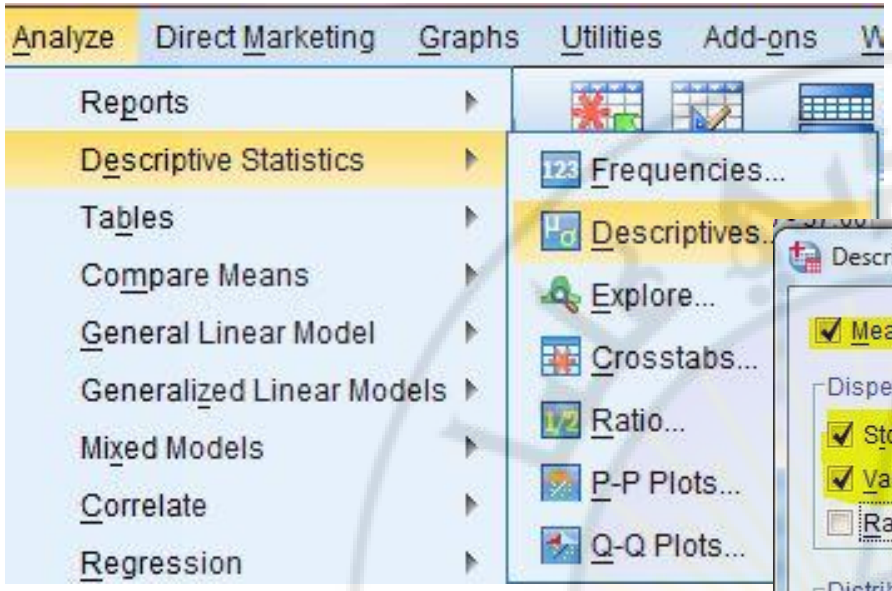
مثال: Kids weight_extended.sav

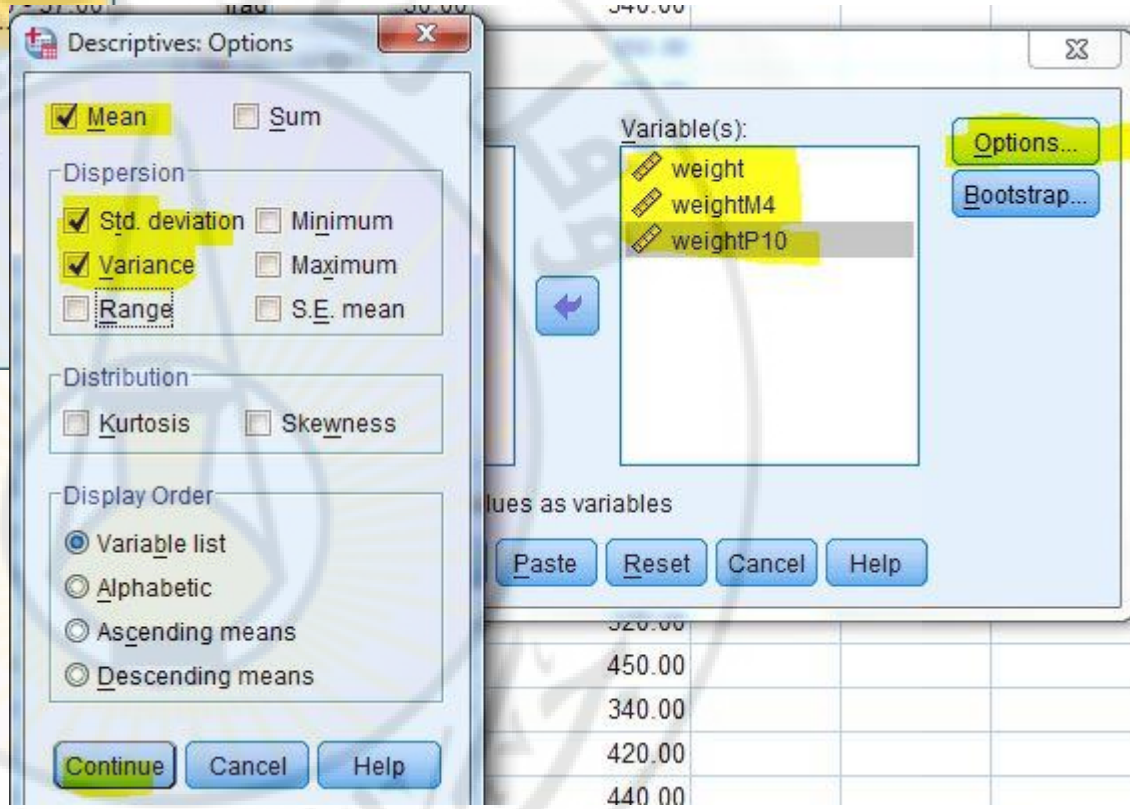
weightP10=weight*10 , weightM4=weight-4


Transform
Analyze
Direct Marketing
Graphs
Compute Variable...
Count Values within Cases...
Shift Values...
Recode into Same Variab
Recode into Different Vari
Automatic Recode...
Visual Binning...
Optimal Binning...
Prepare Data for Modeling
Rank Cases...
Date and Time Wizard...
Create Time Series...
Replace Missing Values...
Random Number Genera
Run Pending Transforms
Compute Variable
Target Variable:
weightM4
=
Numeric Expression:
weight-4
Type & Label...
weight
gender
newweight
country
+ < > 7 8 9
- <= >= 4 5 6
* = ~= 1 2 3
/ & | 0 .
** ~ ( ) Delete
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Date Creation
Functions and Special Variables:
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help



	weight	gender	newweight	country	weightM4	weightP10
1	43.00	boy	43.00 - 47.00	Iraq	39.00	430.00
2	44.00	boy	43.00 - 47.00	Iraq	40.00	440.00
3	34.00	girl	33.00 - 37.00	Iraq	30.00	340.00
4	35.00	girl	33.00 - 37.00	Iraq	31.00	350.00
5	36.00	girl	33.00 - 37.00	Iraq	32.00	360.00
6	39.00	girl	38.00 - 42.00	Iraq	35.00	390.00
7	37.00	girl	33.00 - 37.00	Iraq	33.00	370





Descriptive Statistics

	N	Mean	Std. Deviation	Variance
weight	209	39.6459	5.75535	33.124
weightM4	209	35.6459	5.75535	33.124
weightP10	209	396.4593	57.55348	3312.403
Valid N (listwise)	209			

البحث 4: مقاييس التشتت

تمرين(1): ids weight_extendeKd.sav

أنشئ المتحولين weightm10=weight-10

weightp5=weight*(-5)

كيف تغيرت مقاييس النزعة المركزية (المتوسط والوسط والمنوال) وما التغيير الذي يطرأ على مقاييس التشتت (التباين والانحراف المعياري)؟

نلاحظ بأنّ قيمة كل من المتوسط والوسط والمنوال نقصت بمقدار(10) للمتغير weightm10، وضربت بـ (-5) للمتغير weightp5.

بينما مقاييس التشتت بقيت نفسها للمتغير weightm10، وضربت بـ (5) للمتغير weightp5.

Statistics

		weight	weightm10	weightp5
N	Valid	209	209	209
	Missing	0	0	0
Mean		39.6459	29.6459	-198.2297
Median		40.0000	30.0000	-200.0000
Mode		37.00	27.00	-185.00
Std. Deviation		5.75535	5.75535	28.77674
Variance		33.124	33.124	828.101

تمرين(2):

بالعودة لتمرين درجات الطلاب (سنرمز للمتغير بـ marks)

$$50,35,70,90,40$$

حسبنا:

المتوسط: $\bar{x} = 57$، التباين: $s^2 = 520$، والانحراف: $s = 22.8$

- أنشئ المتحولين: x=marks-20، y=marks*3.
- اعتماداً على خواص ضرب وجمع قيمة لمشاهدات متغير، احسب المتوسط والوسط والمنوال (مقاييس النزعة المركزية) والتباين والانحراف المعياري (مقاييس التشتت) لكل من x وy.

الإحصاء الحيوي

لطلاب السنة الثانية والثالثة - كلية العلوم الصحية

جامعة دمشق

”الجلسة العاشرة“

مدرس المقرر: أ. سلام الجرّاح

إعداد: د. ياسر الزعيم

المدرس في قسم الإحصاء الرياضي

كلية العلوم - جامعة دمشق

حزيران– 2023

نتكلم اليوم عن:

- الاختبارات (الطرق) الوسيطية واللاوسيطية
- بيرسون ونظرائه من طرق غير وسيطية
- اختبار حول مجتمع واحد (اختبار t لعينة واحدة)
- اختبار حول مجتمعين مستقلين (اختبار t لعينتين مستقلتين)
- اختبار حول مجتمعين مرتبطين (اختبار t لعينتين مرتبطتين)
- اختبار حول أكثر من مجتمعين مستقلين (1-way *ANOVA* *المستقل*)

الطرق الوسيطية و الطرق اللاوسيطية

في علم الاستدلال الإحصائي، إنّ طرق اختبار الفرضيات تنقسم إلى قسمين وذلك بسبب طبيعة البيانات ونوع المتحولات المدروسة:

1) **الطرق الوسيطية parametric methods:** هذه الطرق تستلزم معرفة التوزع الإحصائي للبيانات المدروسة ومن أهمها أن تتوزع البيانات المدروسة توزعاً طبيعياً (بالإضافة لشروط أخرى) (وبالتالي من البديهي أن المتحولات المدروسة مستمرة).

2) **الطرق اللاوسيطية nonparametric methods:** وهي طرق لاختبار الفرضيات الإحصائية لا تشترط ولا تعتمد أي توزع احتمالي للبيانات المدروسة، ولذا يتم تسمية هذه الطرق بالطرق الخالية من الفروض (خالية من الشروط).

الطرق الوسيطية و الطرق اللاوسيطية

1) إنّ الطرق الوسيطية (parametric statistics) أقوى (more powerful) من نظرائها غير الوسيطية (nonparametric statistics) أي:
إذا كانت الشروط اللازمة لإنجاز الاختبار الوسيطي محققة، فإنّ فرصة واحتمال رفض الفرضية الصفرية وهي خاطئة أكبر من احتمال رفضها فيما لو استخدمنا الطرق اللاوسيطية لإنجاز ذلك الاختبار.
أي إن الـ sig المحسوبة من الطرق الوسيطية أقل وأصغر من الـ sig المحسوبة من نظرائها غيرالوسيطية.

2) إنّ الطرق غيرالوسيطية أكثر متانة (more robust) من نظرائها الوسيطية أي إذا اختل شرط من الشروط اللازمة لاستخدام الطرق الوسيطية (وفشلت التدابير العلاجية) فالطرق اللاوسيطية هي الأفضل.

تحليل الارتباط

رأينا أن شروط حساب معامل ارتباط بيرسون ρ_{XY} بين متحولين هي:

1) المتحولان مستمران (أي مقاسان بالمقياس القياسي scale)
2) عدم وجود نقط قاصية في أي من المتحولين قيد الدراسة

و درسنا الاختبار الإحصائي $H_0 : \rho_{XY} = 0 \ \ v.s. \ \ H_1 : \rho_{XY} \neq 0$

حيث أنّ الفرض الصفري ينص على أن المتحولان مستقلان خطياً أما الفرض البديل ينص بأن هناك علاقة ما (خط مستقيم).

فإذا اختل شرط من الشرطين السابقين عندئذ نستخدم ارتباط سبيرمن Spearman (يرمز له r_s) أو نستخدم ارتباط كندال Kendall (المسمى كندال تاو-ب) (يرمز له τ_b).

ملاحظة:

إنّ ارتباط سبيرمن يتأثر بالربطات ties لذا من الأفضل اعتماد كندال تاو-ب كبديل لاوسيطي لمعامل ارتباط بيرسون.

اختبار حول مجتمع واحد

اختبار t للعينة الواحدة *one sample t test*:

وهي طريقة وسيطية لاختبار أن متوسط مجتمع ما يساوي قيمة ثابتة معلومة أي أننا نختبر الفرضية الإحصائية التالية:

$$H_0 : \mu = \mu_0 \quad v.s. \quad H_1 : \mu \neq \mu_0$$

William Gosset 1876-1937

يتم إنجاز الاختبار باعتماد توزيع احتمالي يسمى توزيع t-ستيودنت (t-student) الذي قام بتعريفه العالم الإحصائي والكيميائي Gosset

اختبار حول مجتمع واحد

شروط إنجاز اختبار t للعينة الواحدة:

لأجل عينة من حجم كبير أي $N \geqslant 30$ يجب توفر الشروط التالية:

1) المتحول المدروس مستمر
2) خلو المتحول المدروس من النقط القاصية extreme values (فإذا وجدت القيم القاصية وتمت معالجتها) وبقي حجم العينة المدروسة كبيراً، نستطيع استخدام اختبار t)

(فإذا اختل أحد الشرطين السابقين نذهب للبديل اللاوسيطي)

أمّا إذا كانت العينة من حجم صغير أي $N < 30$ فالشروط هي:

1) المتحول المدروس مستمر
2) المتحول المدروس يتوزع طبيعياً

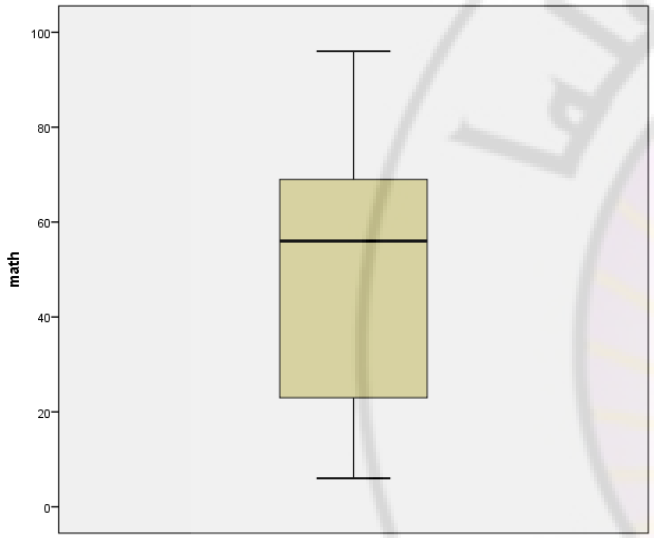
(فإذا اختل أحد الشرطين السابقين نذهب للبديل اللاوسيطي اختبار ويلكوكسن للرتب المؤشرة *Wilcoxon signed-rank test*)

مثال: comprehensive.sav

اختبر أن متوسط علامات الطلاب في مادة الرياضيات math هو 60

math

N	Valid	30
	Missing	0

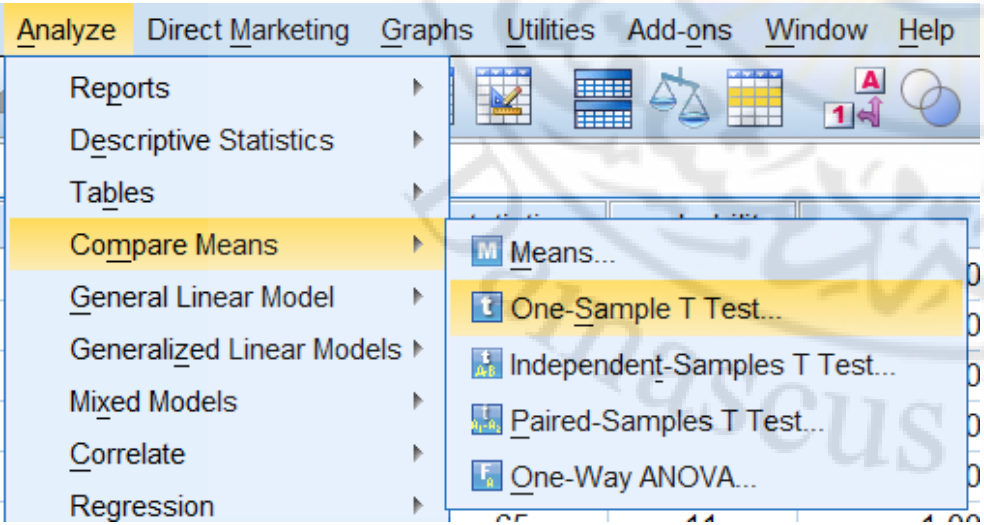


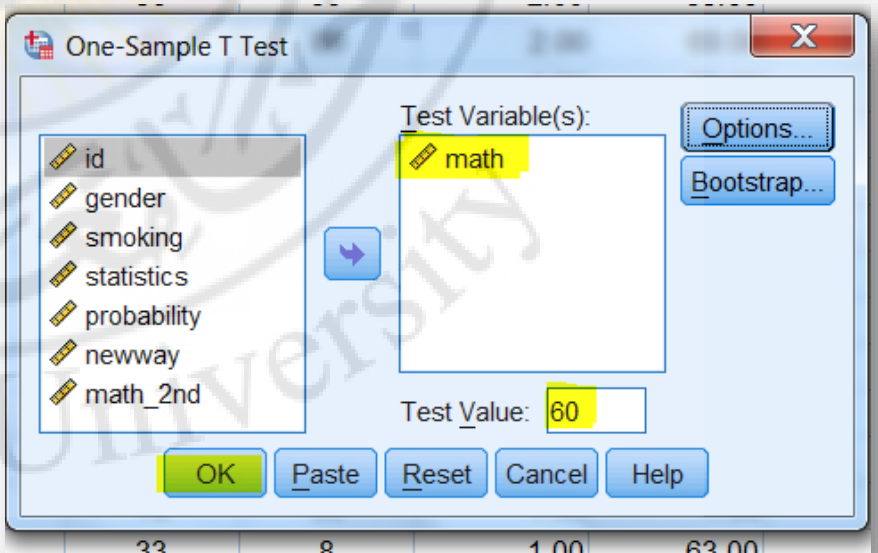
الحل: المطلوب اختبار

$$H_0 : \mu = 60 \quad v.s. \quad H_1 : \mu \neq 60$$

المتحول مستمر و إنّ حجم العينة كبير $N = 30$ و إنّ المتحول خالي من النقط القاصية.

إذاً نستطيع إنجاز الاختبار وفق اختبار t-ستيودنت.





البحث 7: اختبار حول مجتمع واحد

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
math	30	51.27	25.054	4.574

$\overline{x} = 51.27 \ , \ s = 25.05$

One-Sample Test

	Test Value = 60					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
math	-1.909	29	.066	-8.733	-18.09	.62

$H_0 : \mu = 60 \ \ v.s. \ \ H_1 : \mu \neq 60$

$sig = .066 > \alpha = .05$

قبول الفرض الصفري أي قبول الادعاء

إنّ 95% مجال ثقة للفرق بين متوسط المجتمع والقيمة المختبرة هو

$\mu - 60 \in [-18.09, +0.62]$ أي: بثقة 95% إنّ الصفر واقع في هذا المجال

لذا بثقة 95% إن $\mu - 60 \in [-, +]$ أي $\mu - 60 = 0$ أي $\mu = 60$

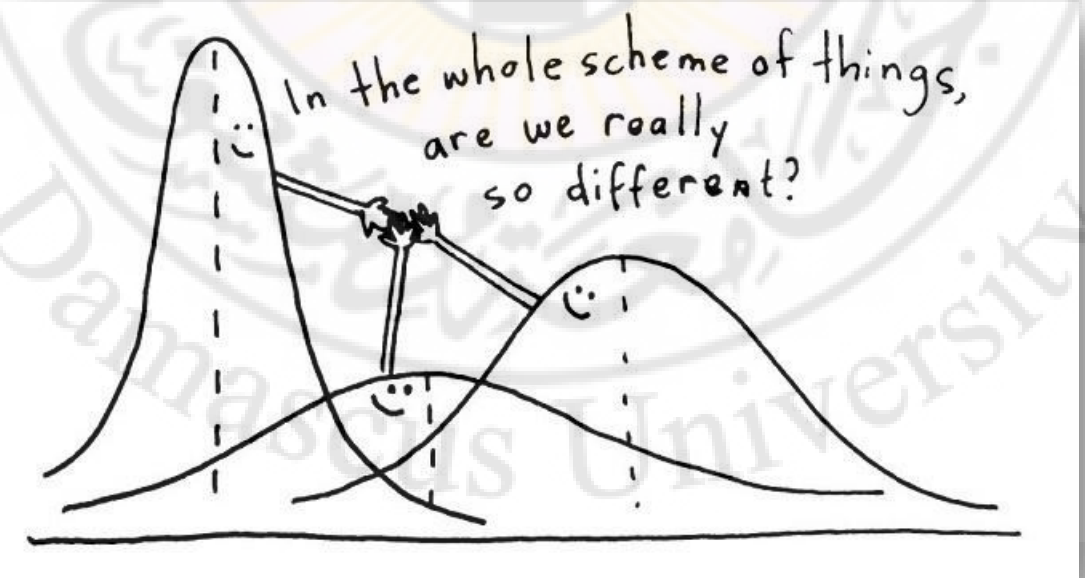
اختبار حول مجتمعين مستقلين

اختبار t للعينتين المستقلتين *independent samples t test*:

وهي طريقة وسيطية لاختبار أنّ متوسطي مجتمعين (مستقلين، أي أنّ مشاهدات أحدهما لا تعتمد ولا تتأثر ولا ترتبط بمشاهدات الآخر) متساويان بالقيمة أي أنّنا نختبر الفرضية الإحصائية التالية

$$H_0 : \mu_1 = \mu_2 \quad v.s. \quad H_1 : \mu_1 \neq \mu_2$$

ويتم إنجاز الاختبار باعتماد التوزيع الاحتمالي t-ستيودنت (t-student) إذا تحققت شروط إنجازه و إلا نستخدم البديل اللاوسيطي المناسب.



اختبار حول مجتمعين مستقلين

شروط إنجاز اختبار t للعينتين المستقلتين:

لأجل عينات من حجم كبير أي $N_1 \geqslant 30 \& N_2 \geqslant 30$ يجب تحقق ما يلي:

1) المتحولين المدروسين مستمرين و مستقلين.
2) خلو المتحولين من النقط القاصية extreme values (فإذا وجدت القيم القاصية وتمت معالجتها وبقي حجم العينة كبيراً، نستطيع استخدام اختبار t)

أمّا إذا كانت العينتين (أو إحداها) من حجم صغير $N_1 < 30 \ \& \ N_2 < 30$ فالشروط هي:

1) المتحولين المدروسين مستمرين و مستقلين
2) التوزع طبيعياً للمتحولين (أو للمتحول ذو العينة الصغير)
3) التجانس أي أن يتساوى تباين مجتمعَي المتحولين homogeneity of variance (or homoscedasticity)

اختبار حول مجتمعين مستقلين

فإذا اختل شرط ممّا سبق، نستخدم البديل اللاوسيطي: اختبار مان-وتني *Mann-Whitney test* و اختبار كولموغوروف-سميرنوف Z (*Two-sample Kolmogorov-Smirnov test*

Howard Levene 1914-2003

التجانس يعني قبول الفرض الصفري في الاختبار الإحصائي (والذي يدعى باختبار لفين Levene) التالي:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad v.s. \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

ملاحظة:

إذا كان المتحولان يتوزعان طبيعياً

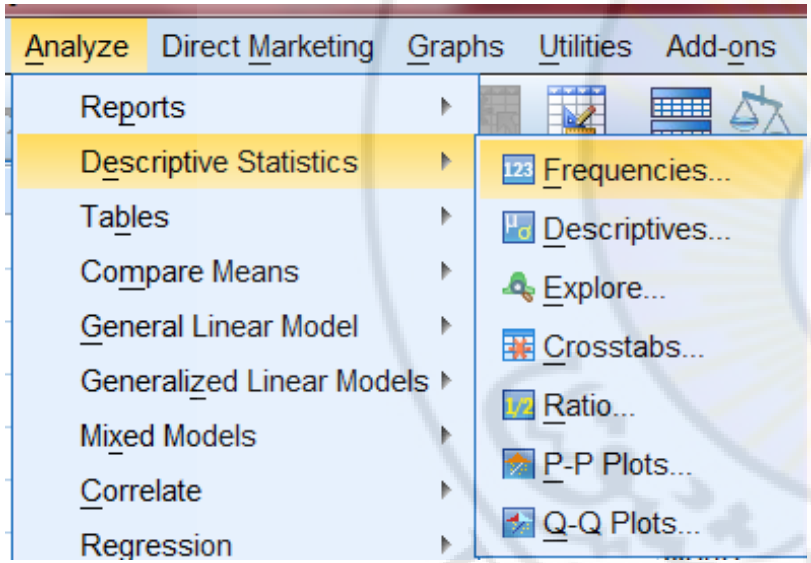
و كانت أحجام العينات $30 > N_1, N_2 \geqslant 12$ فإنّ التجانس محقق.

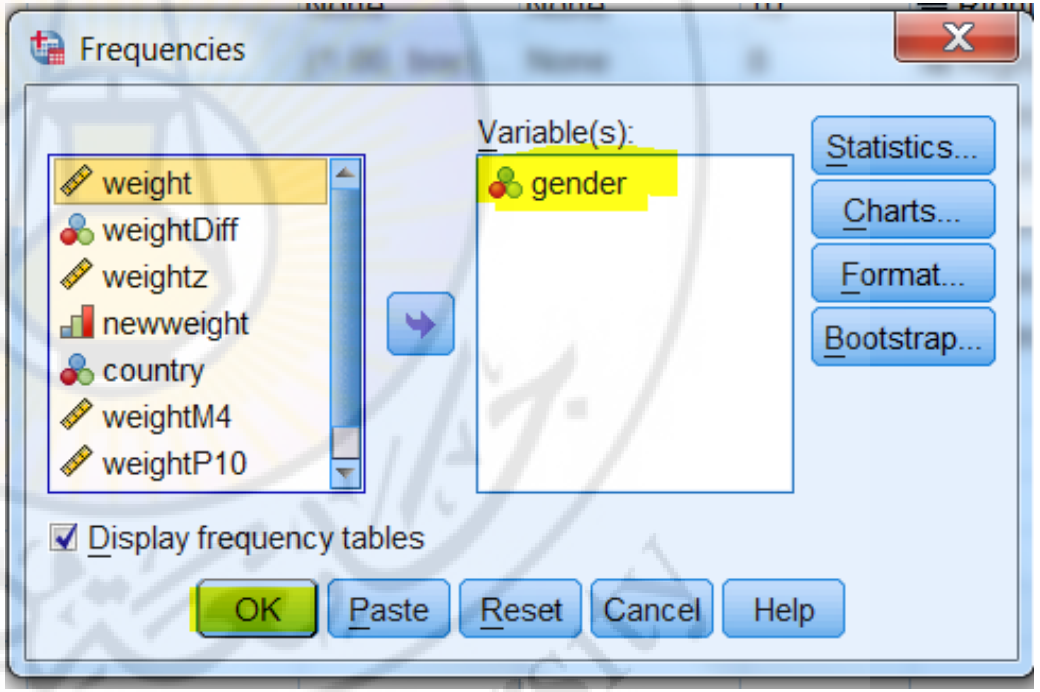
وهذا ينطبق أيضاً في حال كانت فقط إحدى العينتين من حجم صغير، أي إذا كانت عينة من حجم كبير (تتوزع طبيعياً أو لا) والعينة الأخرى من حجم صغير وموزعة طبيعياً وحجمها يزيد على 12 فإنّ التجانس موجود.

اختبار حول مجتمعين مستقلين

مثال: **kids weight_extended.sav** اختبر أن متوسطَي أوزان الأطفال الذكور والأطفال الإناث متساويان، علماً أنّ 1=boy و 2=girl و $\alpha = .05$

الحل: المطلوب اختبار $H_0 : \mu_1 = \mu_2 \quad v.s. \quad H_1 : \mu_1 \neq \mu_2$





gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	boy	122	58.4	58.4	58.4
	girl	87	41.6	41.6	100.0
	Total	209	100.0	100.0	

$N_1 = 122 \quad \& \quad N_2 = 87$

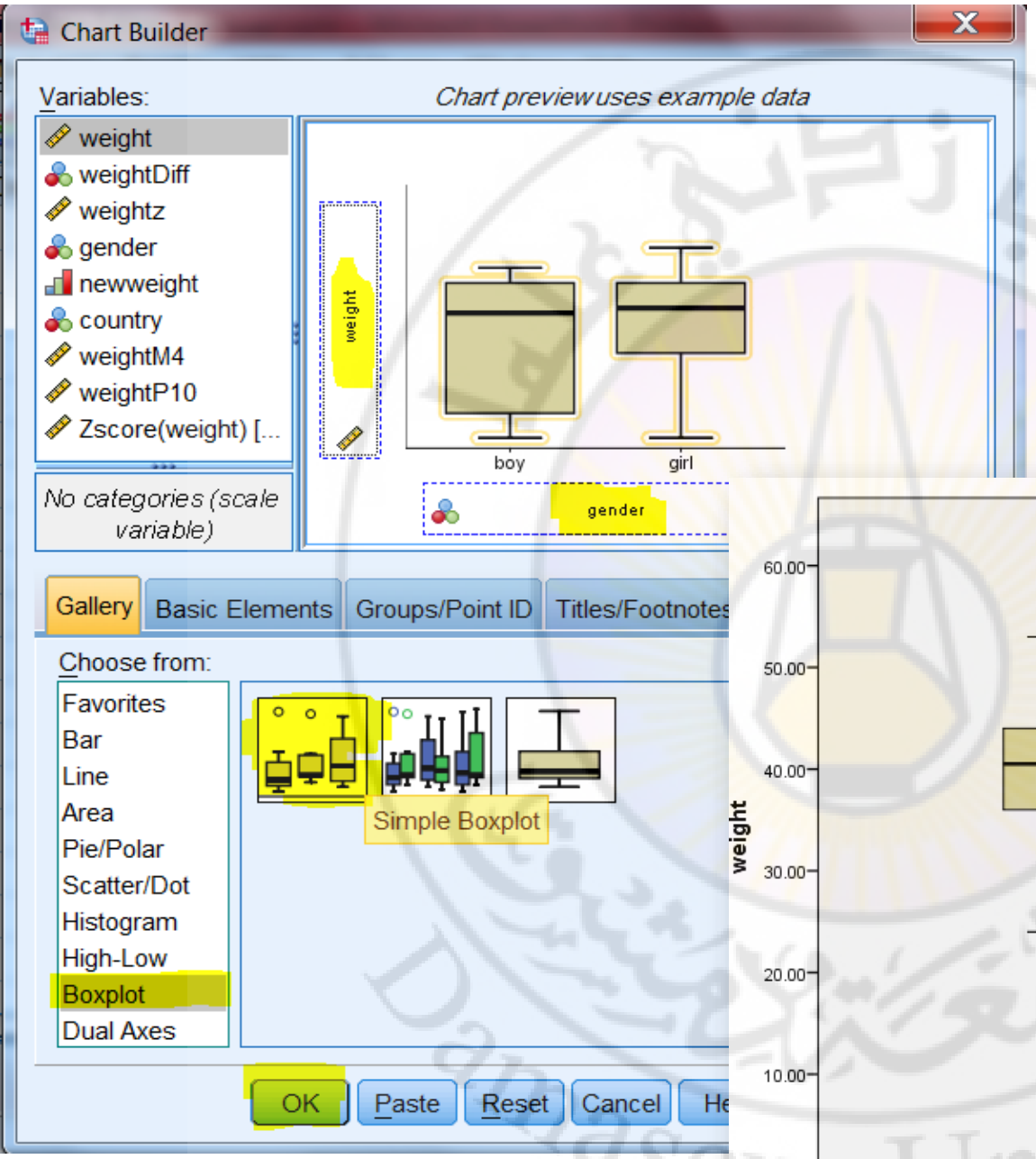
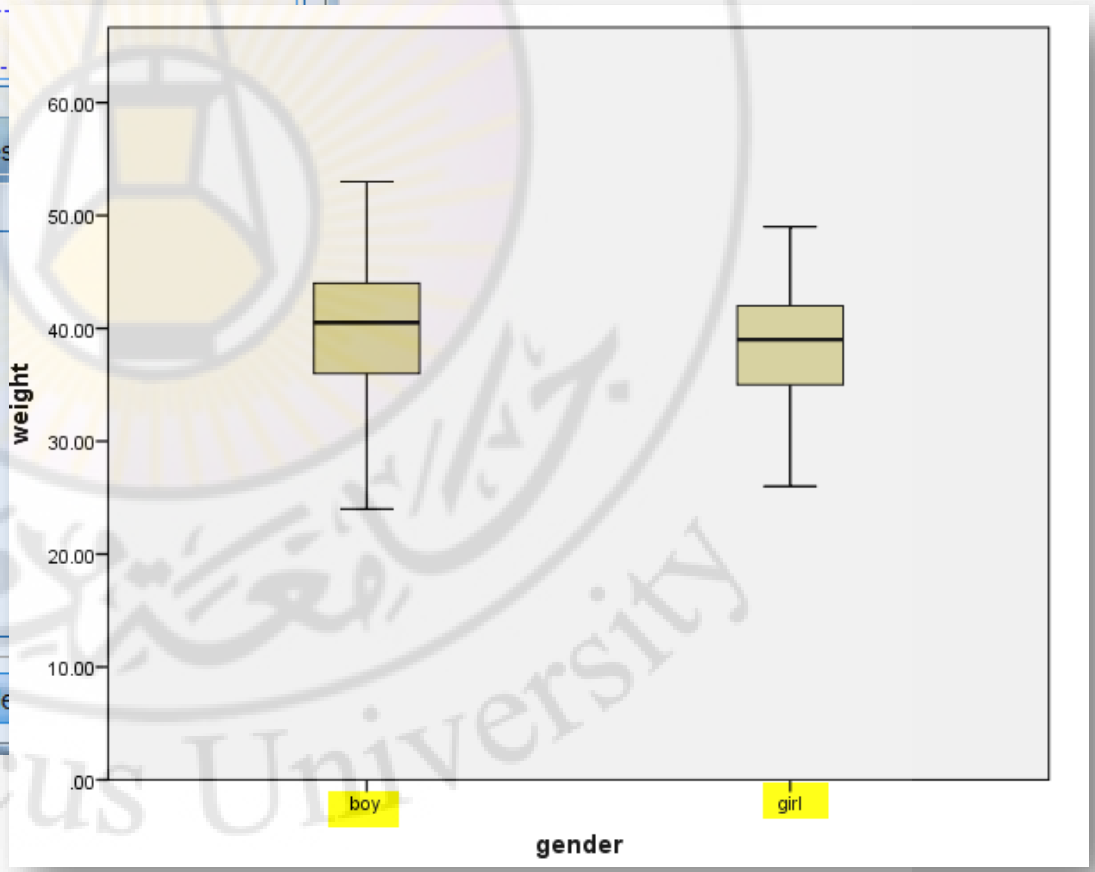
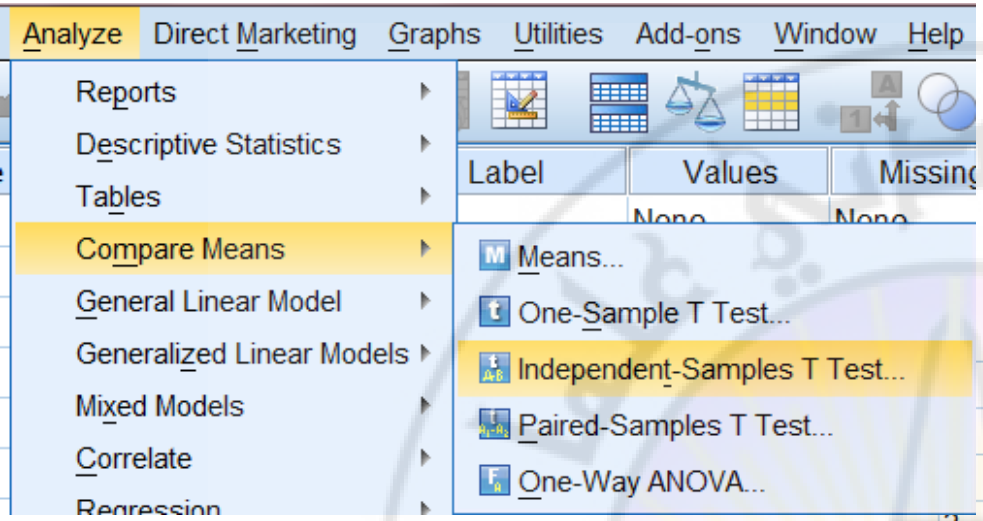
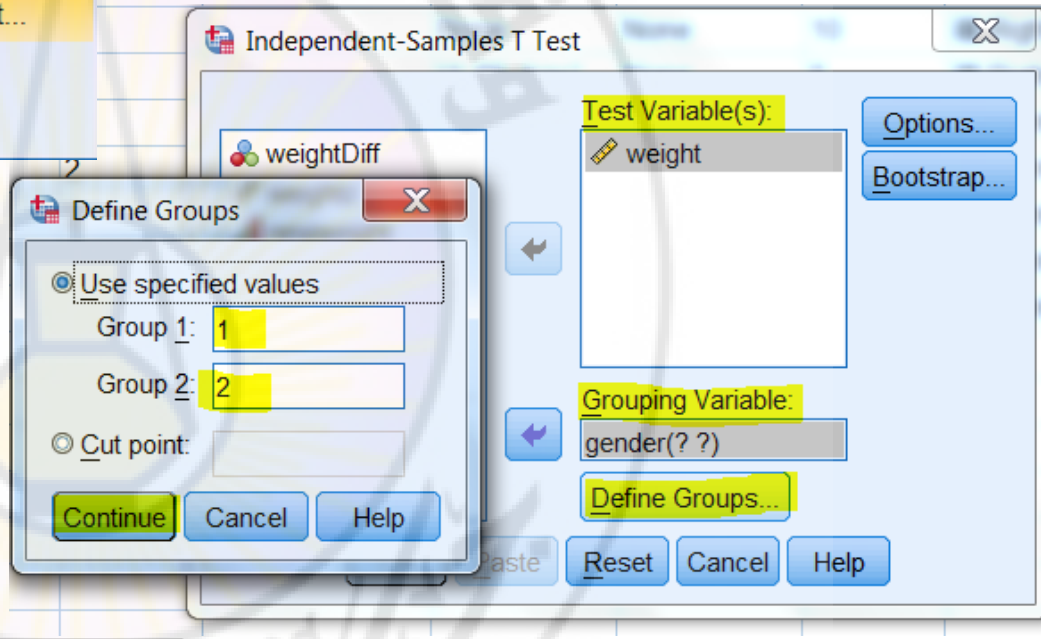


Chart Builder
Variables:
Chart preview uses example data
weight
weightDiff
weightz
gender
newweight
country
weightM4
weightP10
Zscore(weight) [...
No categories (scale variable)
weight
boy
girl
gender
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes
Choose from:
Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot
Histogram
High-Low
Boxplot
Dual Axes
Simple Boxplot
OK
Paste
Reset
Cancel


60.00
50.00
40.00
30.00
20.00
10.00
.00
weight
boy
girl
gender





Group Statistics

	gender	N	Mean	Std. Deviation	Std. Error Mean
weight	boy	122	40.4344	5.93211	.53707
	girl	87	38.5402	5.33702	.57219

$$\overline{x}_1 = 40.43\ ,\ \overline{x}_2 = 38.54$$

$$s_1 = 5.93, s_2 = 5.34$$

$$\left(\begin{array}{l} \overline{x} = 40.43\ ,\ \overline{y} = 38.54 \\ s_x = 5.93, s_y = 5.34 \end{array}\right)$$

اختبار حول مجتمعين مستقلين

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
weight	Equal variances assumed	1.707	.193	2.371	207	.019	1.89420	.79879	.31939	3.46900
	Equal variances not assumed			2.414	196.102	.017	1.89420	.78476	.34655	3.44184

العينات من حجم كبير لذا اختبار التجانس راجع لذوق ورغبة الباحث، ونلاحظ أنّ التجانس محقق كون $sig = .193 > \alpha = .05$ وبالتالي يمكن قراءة نتائج اختبار t-ستيودنت من السطر الأول (الملون بالأزرق) أو من السطر الثاني (الملون بالأحمر).

$H_0 : \mu_1 = \mu_2 \quad v.s. \quad H_1 : \mu_1 \neq \mu_2$

$sig = .019 < \alpha = .05$

رفض الفرض الصفري و قبول الفرض البديل.

$H_0 : \mu_1 = \mu_2 \quad v.s. \quad H_1 : \mu_1 \neq \mu_2$

$sig = .017 < \alpha = .05$

رفض الفرض الصفري و قبول الفرض البديل.

اختبار حول مجتمعين مستقلين

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
weight	Equal variances assumed	1.707	.193	2.371	207	.019	1.89420	.79879	.31939	3.46900
	Equal variances not assumed			2.414	196.102	.017	1.89420	.78476	.34655	3.44184

إنّ 95% مجال ثقة للفرق بين متوسطي المجتمعين هو $\mu_1 - \mu_2 \in [+,+]$

أي بثقة 95% إنّ الصفر غير واقع في هذا المجال أي $\mu_1 - \mu_2 > 0$

ولذا بثقة 95% إنّ $\mu_1 > \mu_2$

اختبار حول مجتمعين مرتبطين

اختبار t للعينتين المرتبطتين *paired samples t test*:

(correlated pairs t test or dependent samples t test)

وهي طريقة وسيطية لاختبار أنّ متوسطي مجتمعين (مرتبطين، أي أنّ المشاهدة موجودة في كل من المجتمعين) متساويان بالقيمة أي أننا نختبر الفرضية الإحصائية التالية

$$H_0 : \mu_1 = \mu_2 \quad v.s. \quad H_1 : \mu_1 \neq \mu_2$$

أو يتم صياغة الاختبار كالتالي:

أي يمكن إرجاعه لاختبار حول العينة الواحدة للتساوي مع الصفر

$$H_0 : \mu_1 - \mu_2 = 0 \quad v.s. \quad H_1 : \mu_1 - \mu_2 \neq 0$$

ويتم إنجاز الاختبار باعتماد التوزيع الاحتمالي t-ستيودنت (t-student) إذا تحققت شروط إنجازه و إلا نستخدم للبديل اللاوسيطي المناسب.

اختبار حول مجتمعين مرتبطين

شروط إنجاز اختبار t للعينتين المرتبطتين:

لأجل عينة من حجم كبير أي $N \geqslant 30$ يجب توفر الشروط التالية:

1) المتحولين المدروسين مستمرين ومرتبطين (أي الحالة التجريبية ذاتها موجودة في كل منهما) وبمشاهدات مستقلة (أي المشاهدات داخل كل متحول مستقلة عن المشاهدات الأخرى)
2) خلو كل منهما من النقط القاصية extreme values (فإذا وجدت القيم القاصية وتمت معالجتها وبقي حجم العينة المدروسة كبيراً، نستطيع استخدام اختبار t)

أمّا إذا كانت العينة من حجم صغير أي $N < 30$ فالشروط هي:

1) المتحولين المدروسين مستمرين ومرتبطين وبمشاهدات مستقلة
2) المتحولين المدروسين يتوزعان طبيعياً

فإذا اختل شرط مما سبق، نستخدم البديل اللاوسيطي: اختبار ويلكوكسن للرتب المؤشرة أو اختبار الإشارة (يشترطان أن البيانات رتبية على الأقل) أواختبار ماكنيمار (بشرط أن البيانات اسمية)

اختبار حول مجتمعين مرتبطين

مثال: يحتوي الملف **deer.sav** بيانات لطول القوائم الخلفية وطول القوائم الأمامية (مقاسة بالسنتيمتر) لعشرة غزلان في غابة معينة. عند $\alpha = .10$ ، هل تعتقد بأنّ القوائم الأمامية أطول من القوائم الخلفية لدى كل غزال؟

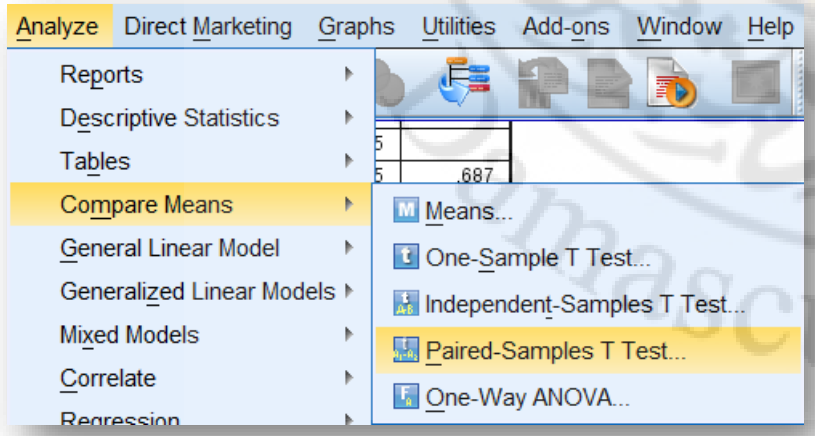
الحل: نريد اختبار

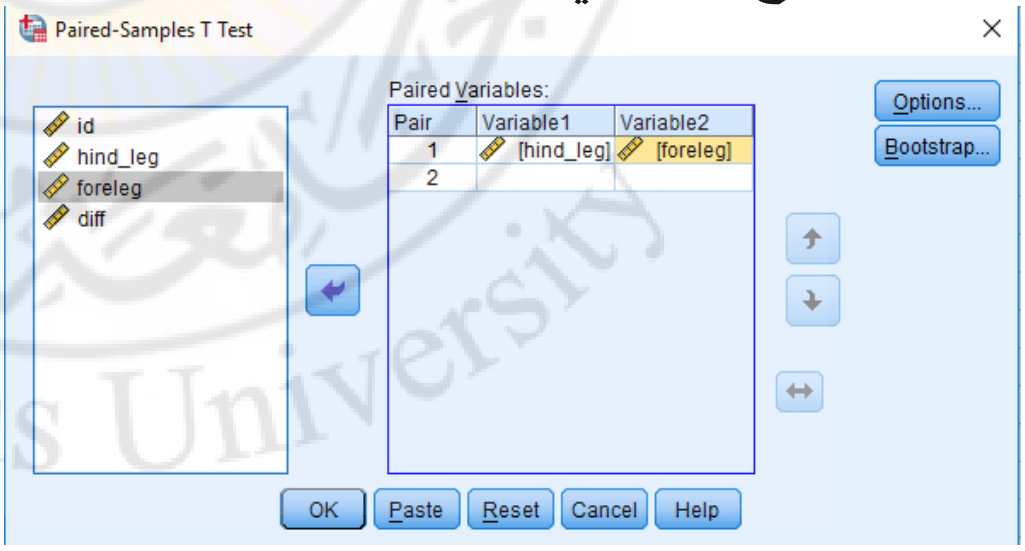
$$H_0: \mu_{\text{hind\_leg}} = \mu_{\text{foreleg}} \quad v.s. \quad H_1: \mu_{\text{hind\_leg}} \neq \mu_{\text{foreleg}}$$

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
hind_leg	.186	10	.200*	.926	10	.408
foreleg	.154	10	.200*	.931	10	.454

العينة من حجم صغير $N = 10$ لذا نتحقق من التوزع الطبيعي لكل متحول





Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	hind_leg	144.7000	10	3.40098	1.07548
	foreleg	141.4000	10	4.03320	1.27541

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	hind_leg & foreleg	10	.674	.033

$$r = .674$$

$$H_0: \rho = 0 \;\; v.s. \; H_1: \rho \neq 0$$

$$sig = .674 > \alpha = .10$$

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	80% Confidence Interval of the Difference Lower	Upper			
Pair 1	hind_leg - foreleg	3.30000	3.05687	.96667	1.96307	4.63693	3.414	9	.008

$$H_0: \mu_{\text{hind\_leg}} = \mu_{\text{foreleg}}$$

$$H_1: \mu_{\text{hind\_leg}} \neq \mu_{\text{foreleg}}$$

$$sig = .008 < \alpha = .10$$

$$\mu_{\text{hind\_leg}} - \mu_{\text{foreleg}} \in [+, +]$$

$$H_1: \mu_{\text{hind\_leg}} > \mu_{\text{foreleg}}$$

اختبار حول أكثر من مجتمعين مستقلين

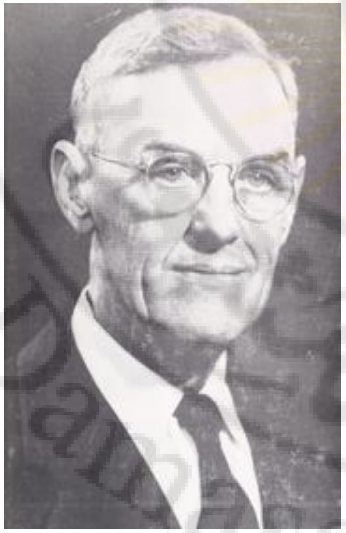
اختبار ANOVA وحيد الاتجاه *1-way ANOVA test* :

وهي طريقة وسيطية لاختبار أن متوسطات k مجتمع مستقل (k عينة مستقلة) متساوية، أي أننا نختبر الفرضية الإحصائية التالية عند المعنوية α

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

مقابل الفرض البديل:

H_1 : (يوجد اختلاف بين متوسطين على الأقل)



George Snedecor 1881 -1974

يتم إنجاز الاختبار باعتماد توزيع احتمالي يسمى توزيع F-فيشر (F-Fisher) الذي قام بتعريفه العالم الإحصائي Snedecor و أسماه إكراماً لعالم الإحصاء و عالم الوراثة Fisher، بتوزيع فيشر أو Fisher–Snedecor distribution

اختبار حول أكثر من مجتمعين مستقلين

شروط إنجاز اختبار ANOVA لأجل k عينة مستقلة:

لأجل عينات من حجم كبير أي $N_k \geqslant 30, \cdots, N_2 \geqslant 30, N_1 \geqslant 30$:

1) المتحولات مستمرة و مستقلة
2) خلو المتحولات من النقط القاصية extreme values (فإذا وجدت القيم القاصية في متحول ما وتمت معالجتها وبقي حجم العينة كبيراً، نستطيع استخدام اختبار ANOVA)

أمّا إذا كانت العينات (أو بعضاً منها) من حجم صغير

$$N_k < 30 , \cdots , N_2 < 30 , N_1 < 30$$

1) المتحولات مستمرة و مستقلة
2) التوزع طبيعياً للمتحولات (أو لبعضها و ذات الحجم الصغير)
3) التجانس أي أن يتساوى تباين مجتمعات المتحولات homogeneity of variance (or homoscedasticity)

فإذا اختل شرط مما سبق، نستخدم البديل اللاوسيطي:اختبار كروسكال-والاس

اختبار حول أكثر من مجتمعين مستقلين

التجانس هو قبول الفرض الصفري في الاختبار (اختبار لفين Levene)

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

$$v.s. \quad H_1$$:(يوجد اختلاف بين تباينين على الأقل)

ملاحظة: إذا كان المتحولات تتوزع طبيعياً و كانت $30 > N_1 \geqslant 12$ و $30 > N_2 \geqslant 12$ و ... و $30 > N_k \geqslant 12$ فإن التجانس محقق.

وهذا ينطبق أيضاً في حال كانت فقط بعض العينات من حجم صغير، أي إذا كانت بعض العينات من حجم كبير (تتوزع طبيعياً أو لا) والبعض الآخر من حجم صغير (يزيد عن 12) وتتوزع طبيعياً فإن التجانس موجود.

أي أن كل من اختبارات t و ANOVA مقاومان ومتينان مقابل خرق فرضية التجانس (فرضية تساوي التباينات)

T tests and ANOVA are fairly robust to deviations from homogeneity of variances assumption.

اختبار حول أكثر من مجتمعين مستقلين

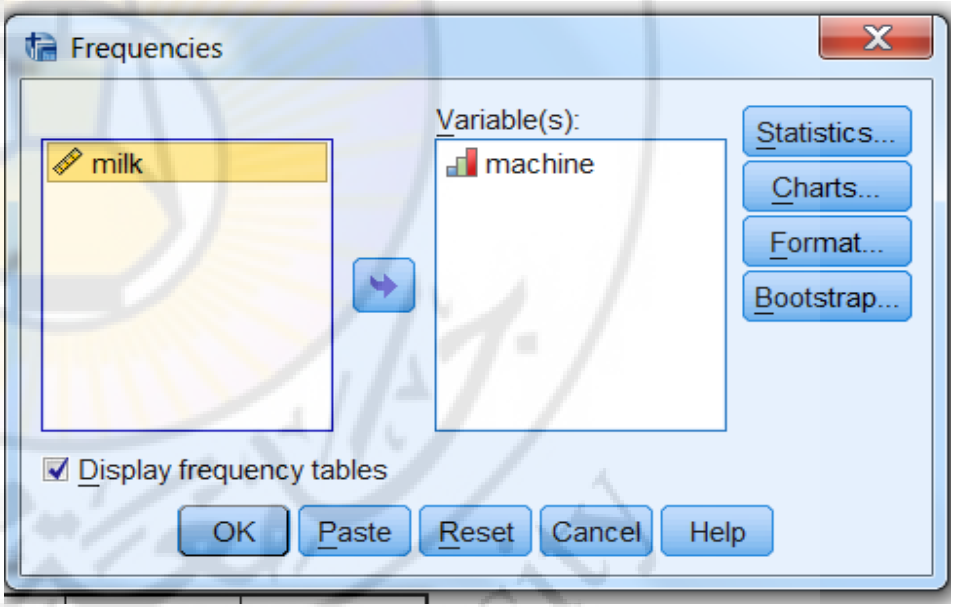
مثال: milk weight.sav اختبر أنّ متوسط أوزان زجاجات الحليب (بالكيلوغرام) لا تختلف اختلافاً معنوياً عن بعضها تبعاً لآلة التعبئة. $\alpha = .05$

$$H_0: \mu_1 = \mu_2 = \mu_3$$

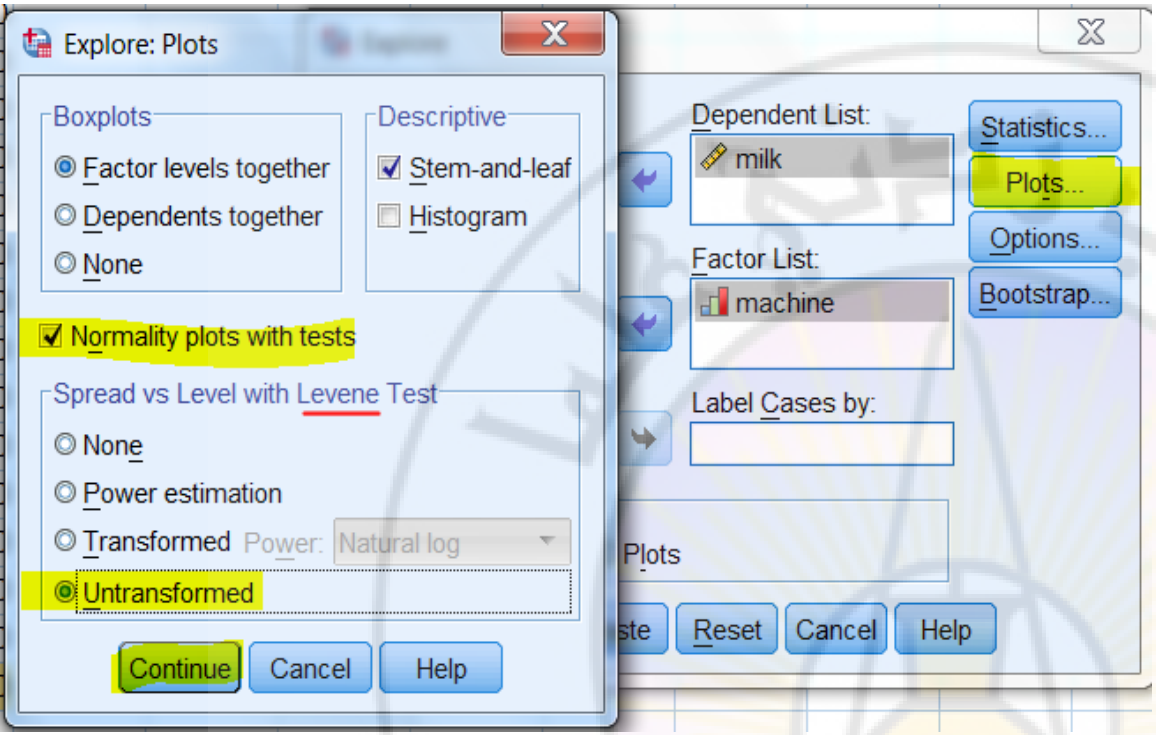
$$v.s.\ H_1:$$ (يوجد اختلاف بين متوسطين على الأقل)

اصطلاح:
Cases- Subjects- Experimental units

	milk	machine
1	3.00	1.00
2	2.00	1.00
3	1.00	1.00
4	1.00	1.00
5	4.00	1.00
6	5.00	2.00
7	2.00	2.00
8	4.00	2.00
9	2.00	2.00
10	3.00	2.00
11	7.00	3.00
12	4.00	3.00
13	5.00	3.00
14	3.00	3.00
15	6.00	3.00



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	5	33.3	33.3	33.3
	2.00	5	33.3	33.3	66.7
	3.00	5	33.3	33.3	100.0
	Total	15	100.0	100.0	



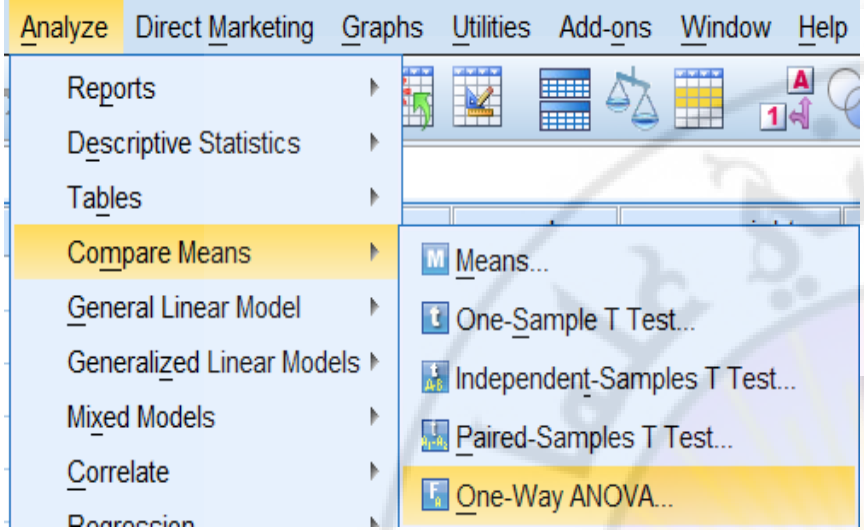
Tests of Normality

	machine	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
milk	1.00	.221	5	.200*	.902	5	.421
	2.00	.221	5	.200*	.902	5	.421
	3.00	.136	5	.200*	.987	5	.967

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
milk	Based on Mean	.092	2	12	.913
	Based on Median	.118	2	12	.890
	Based on Median and with adjusted df	.118	2	11.677	.890
	Based on trimmed mean	.097	2	12	.908

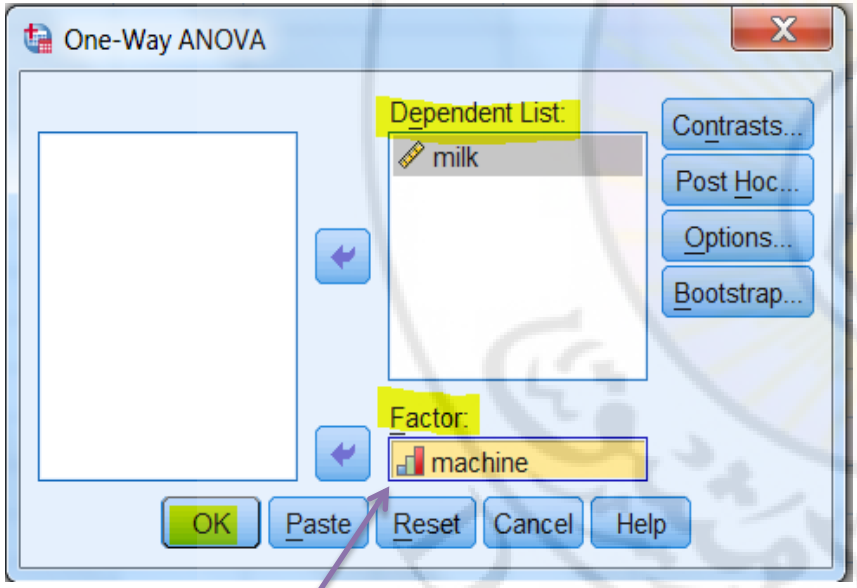
إنّ شرط التوزع الطبيعي لمجتمع كل عينة محقق، وإنّ شرط التجانس (تساوي تباينات المجتمعات) أيضاً محقق ولذا نتابع نحو ANOVA.



ANOVA

milk

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	20.133	2	10.067	5.119	.025
Within Groups	23.600	12	1.967		
Total	43.733	14			



العامل

مجموع المربعات بين المعالَجات

SSB=between groups sum of squares

SSB=20.133

مجموع المربعات ضمن المعالجات

SSW=sum of squares within groups

SSW=23.6

مجموع المربعات الكلي

SST=total sum of squares

SST=SSB+SSW=43.733

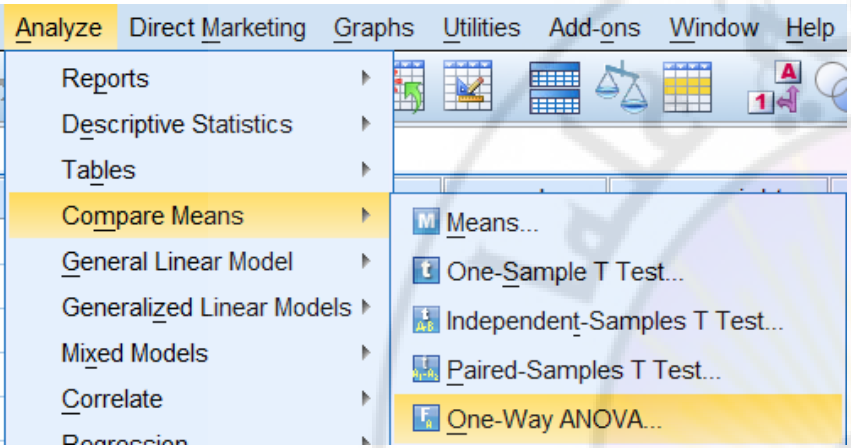
$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$v.s. \quad H_1 : \text{(يوجد اختلاف بين متوسطين على الأقل)}$$

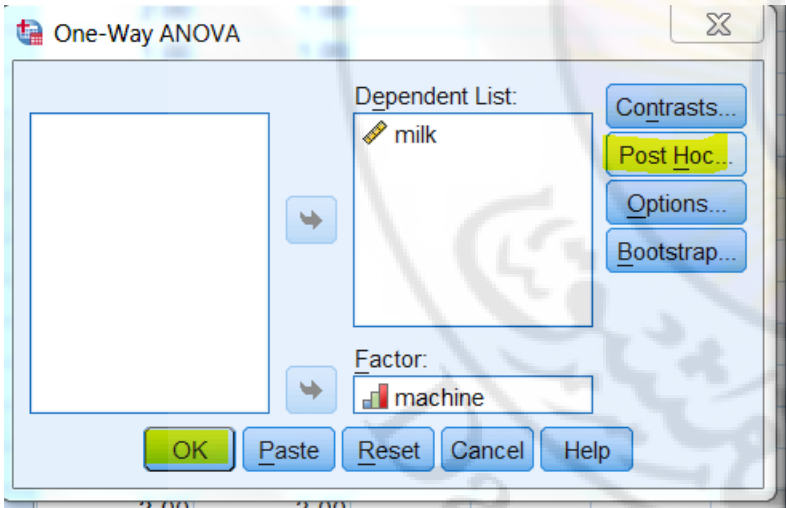
اختبار حول أكثر من مجتمعين مستقلين

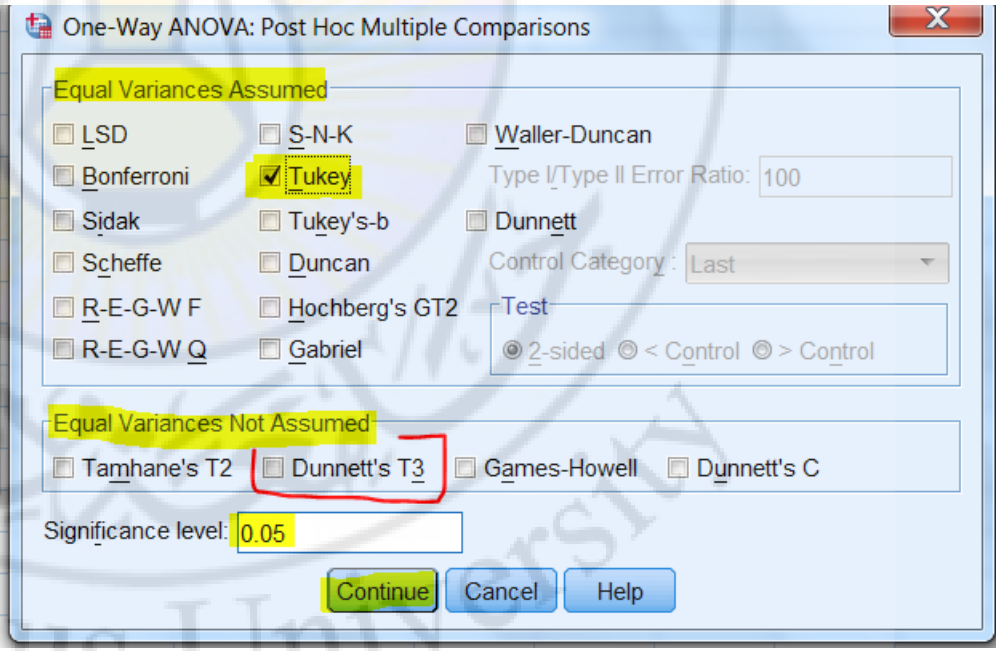
توجد فروق هامة بين متوسطات أوزان الحليب المعبأة وفقاً للآلة لأن $sig = .025 < \alpha = .05$

والآن علينا أن نكتشف أين تقع هذه الفروق.



post hoc means





إنّ الاختبارات في post hoc هي ثنائية (مثنى مثنى)

اختبار توكي Tukey: اختبارات ثنائية لاكتشاف الفروق الهامة بين المتوسطات بعد رفض الفرضية الصفرية في ANOVA.

Post Hoc Tests

Multiple Comparisons

Dependent Variable: milk

Tukey HSD

(I) machine	(J) machine	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	-1.00000	.88694	.516	-3.3662	1.3662
	3.00	-2.80000*	.88694	.021	-5.1662	-.4338
2.00	1.00	1.00000	.88694	.516	-1.3662	3.3662
	3.00	-1.80000	.88694	.147	-4.1662	.5662
3.00	1.00	2.80000*	.88694	.021	.4338	5.1662
	2.00	1.80000	.88694	.147	-.5662	4.1662

*. The mean difference is significant at the 0.05 level.

Homogeneous Subsets

milk

Tukey HSD[a]

machine	N	Subset for alpha = 0.05	
		1	2
1.00	5	2.2000	
2.00	5	3.2000	3.2000
3.00	5		5.0000
Sig.		.516	.147

$H_0: \mu_1 = \mu_2 \ \ v.s. \ \ H_1: \mu_1 \neq \mu_2$

$sig = .516 > \alpha = .05$

$\mu_1 - \mu_2 \in [-, +]$

$H_0: \mu_1 = \mu_3 \ v.s. \ H_1: \mu_1 \neq \mu_3$

$sig = .021 < \alpha = .05$

$\mu_1 - \mu_3 \in [-, -] \quad \mu_1 - \mu_3 < 0$

$\mu_1 < \mu_3$

$H_0: \mu_2 = \mu_3 \ v.s. \ H_2: \mu_2 \neq \mu_3$

$sig = .147 > \alpha = .05$

$\mu_2 - \mu_3 \in [-, +] \ \mu_2 - \mu_3 = 0$

$\mu_2 = \mu_3$

الإجراء الوسيطي	الإجراء اللاوسيطي
بيرسون	سبيرمان؛ كندال تاو(ب)
t (ستيودنت) لعينة واحدة	ويلكوكسن للرتب المؤشرة
t لعينتين مستقلتين	مان-وتني؛ كولموغوروف-سميرنوف z
t لعينتين مرتبطتين	ويلكوكسن؛ الإشارة؛ ماكنيمار
1-way ANOVA المستقل	كروسكال-والاس

الإحصاء الحيوي

لطلاب السنة الثانية والثالثة - كلية العلوم الصحية

جامعة دمشق

" الجلسة الثالثة "

مدرس المقرر: أ. سلام الجرّاح

إعداد: د. ياسر الزعيم

المدرس في قسم الإحصاء الرياضي

كلية العلوم - جامعة دمشق

نيسان– 2023

نتكلم اليوم عن:

- تذكرة بما تحدثنا عنه
- **البحث الثاني: التمثيل البياني للمشاهدات**
- التمثيل البياني في SPSS

تذكرة بما تحدثنا عنه

$$\sum f_i = N = 209$$

ندعوا i بالدليل السفلي (subscript (or index.

الحدود الفعلية هي $[a,b]$ $[a-0.5,b+0.5[$.

		Frequency
Valid	23.00 - 27.00	4
	28.00 - 32.00	17
	33.00 - 37.00	58
	38.00 - 42.00	64
	43.00 - 47.00	47
	48.00+	19
	Total	209

البحث 2: التمثيل البياني للمشاهدات

الرسم البياني (graph) أو المخطط البياني (chart)
عبارة عن تمثيل تصويري مرئي للمشاهدات بهدف فهمها بيانياً.

التوزيع التكراري يساعد على فهم المشاهدات عددياً في جداول،
أمّا **الرسم البياني** فيساعد على الفهم بطريقة ترسيمية.

البحث 2: التمثيل البياني للمشاهدات

أولاً مخطط الدائرة (الفطيرة) Pie chart

رسم بياني على شكل شرائح (slices) لتمثيل المتغيرات المنفصلة سواء كانت اسمية أم ترتيبية.

(ندعو المتغيرات الترتيبية أيضاً بالفئوية categorical variables).

kids weight_extended.sav

Data View | Variable View

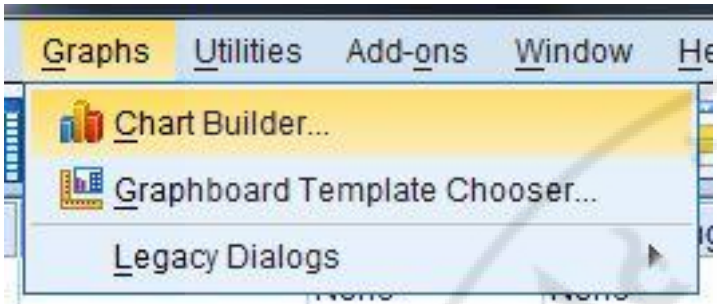

Graphs
Utilities
Add-ons
Window
Chart Builder...
Graphboard Template Chooser...
Legacy Dialogs


Chart Builder
Before you use this dialog, measurement
variable in your chart. In addition, if your ch
value labels should be defined for each c
Press OK to define your chart.
Press Define Variable Properties to set m
labels for chart variables.
Don't show this dialog again
OK
Defi

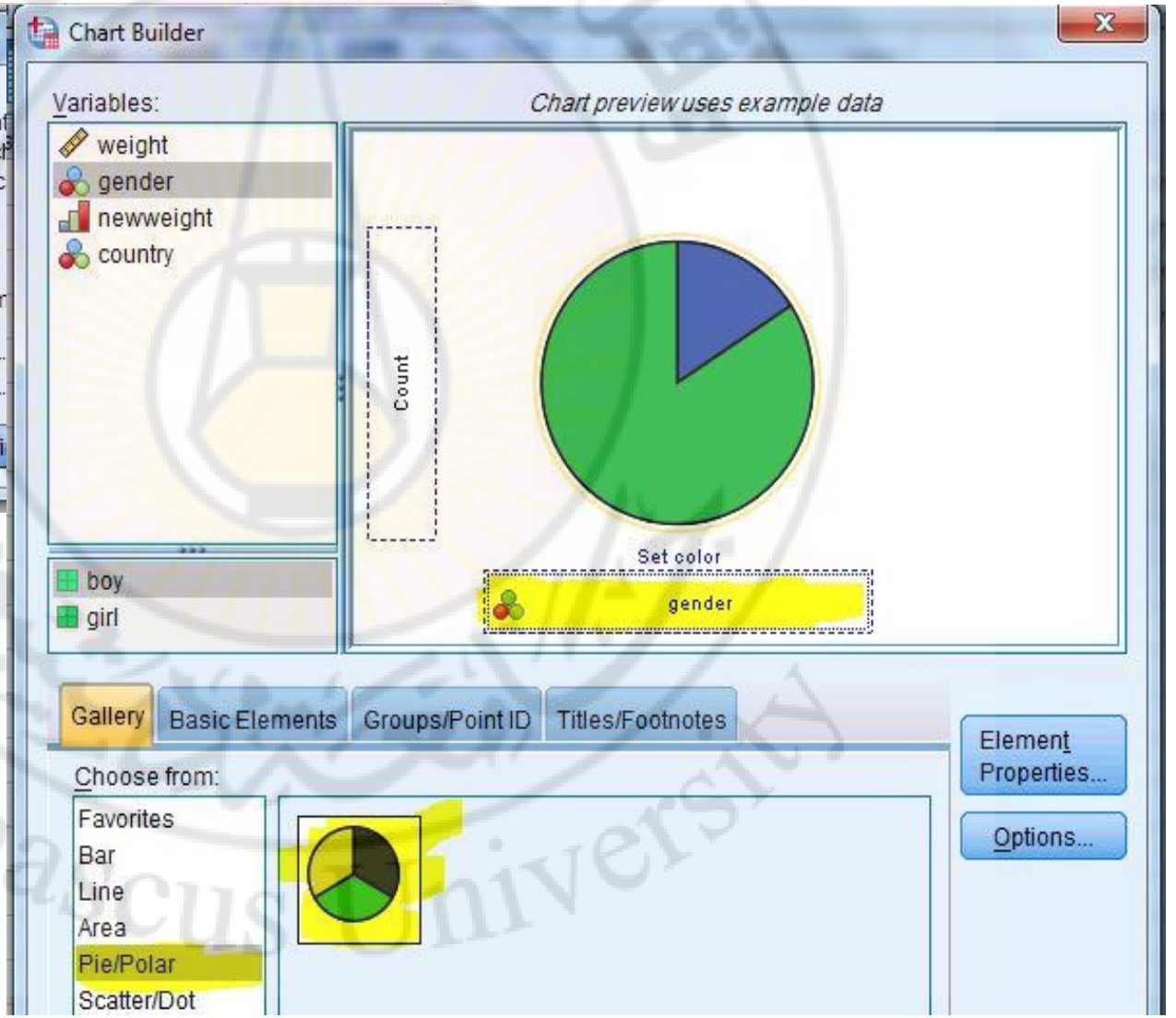
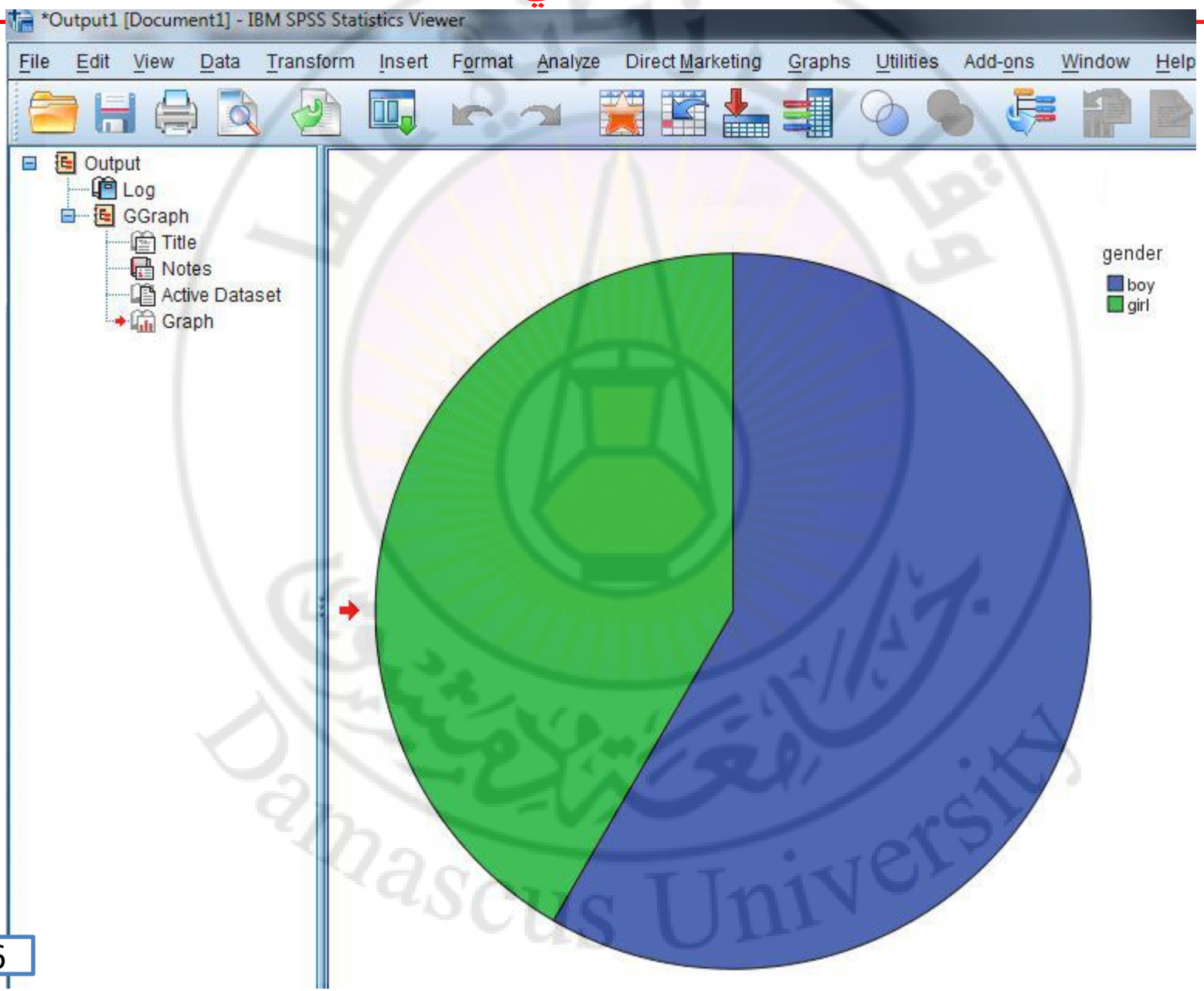


Chart Builder
Variables:
Chart preview uses example data
weight
gender
newweight
country
boy
girl
Count
Set color
gender
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes
Element Properties...
Options...
Choose from:
Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot

البحث 2: التمثيل البياني للمشاهدات



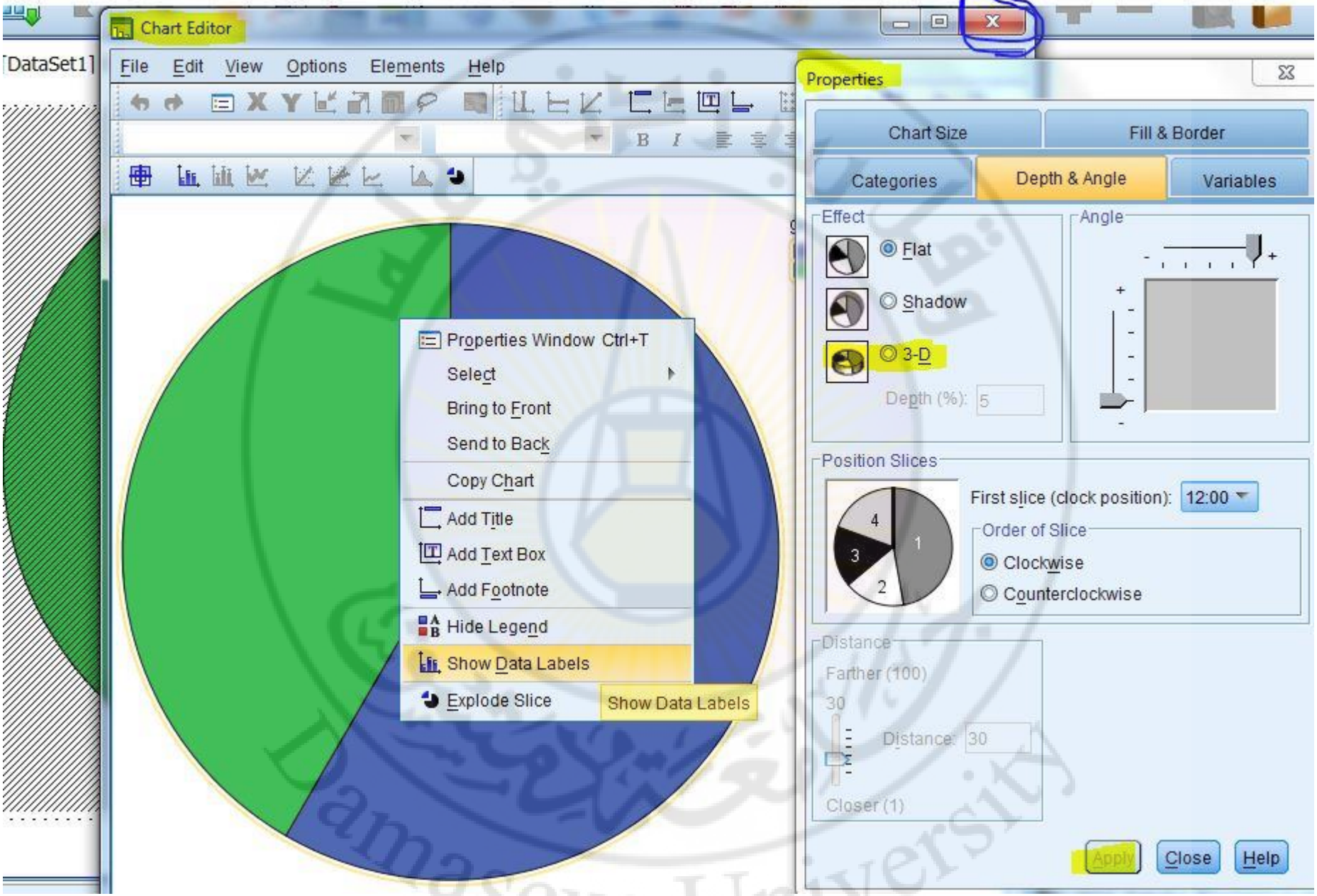
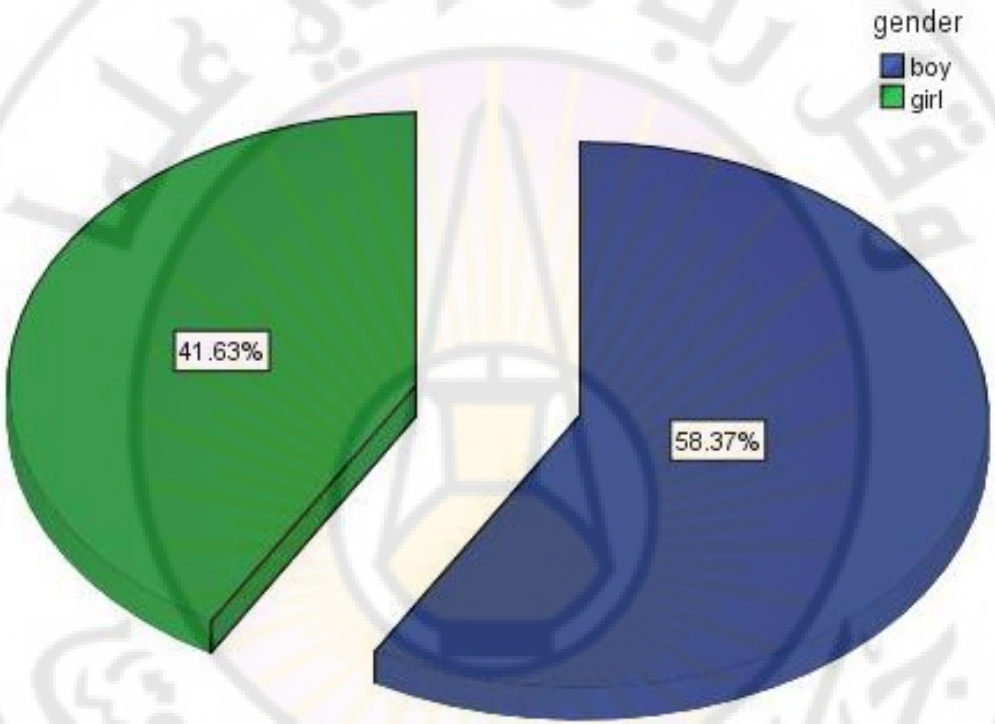


Chart Editor
File Edit View Options Elements Help
Properties Window Ctrl+T
Select
Bring to Front
Send to Back
Copy Chart
Add Title
Add Text Box
Add Footnote
Hide Legend
Show Data Labels
Explode Slice
Show Data Labels
Properties
Chart Size
Fill & Border
Categories
Depth & Angle
Variables
Effect
Flat
Shadow
3-D
Depth (%): 5
Angle
Position Slices
First slice (clock position): 12:00
Order of Slice
Clockwise
Counterclockwise
Distance
Farther (100)
30
Distance: 30
Closer (1)
Apply
Close
Help

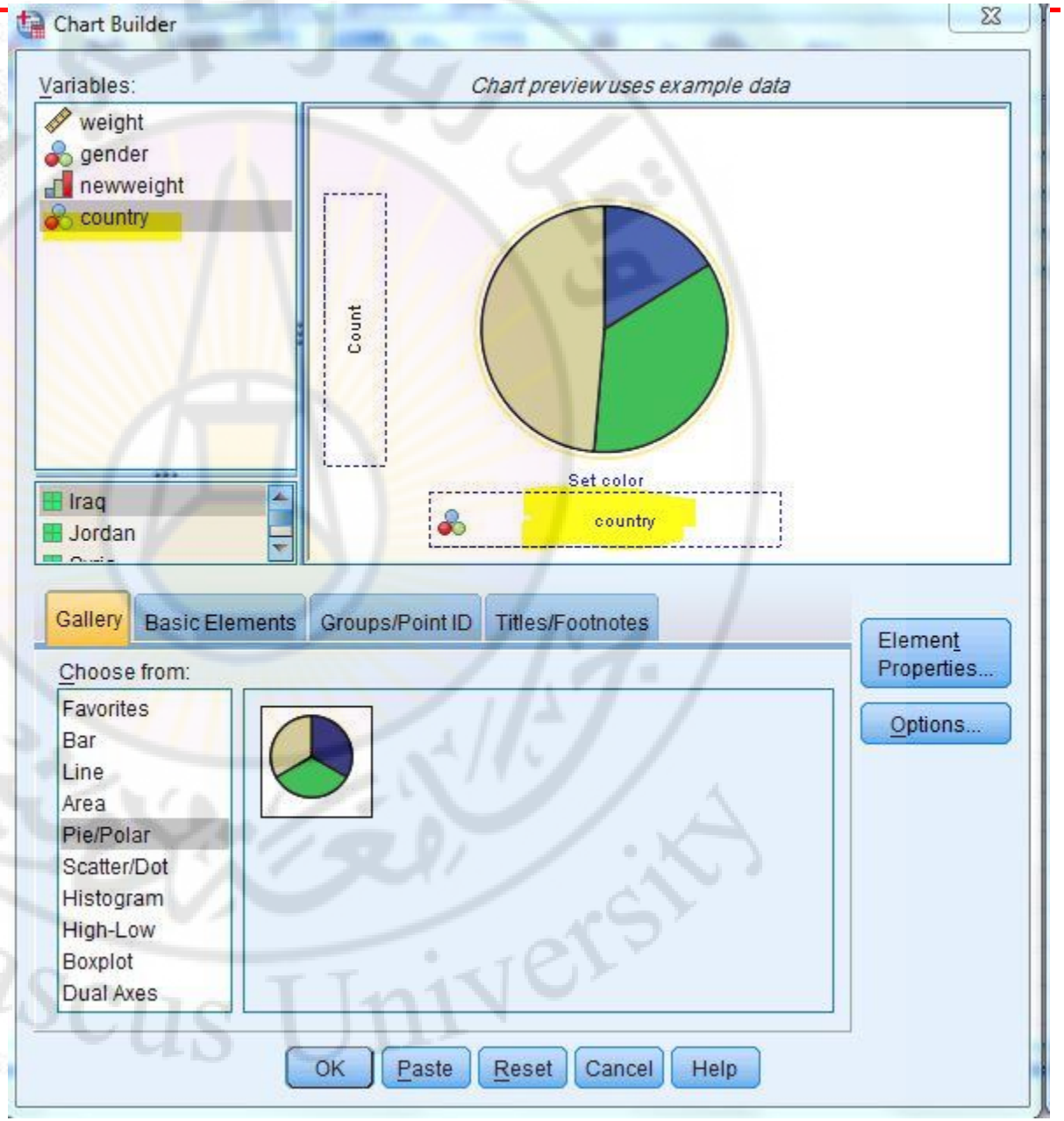
البحث 2: التمثيل البياني للمشاهدات



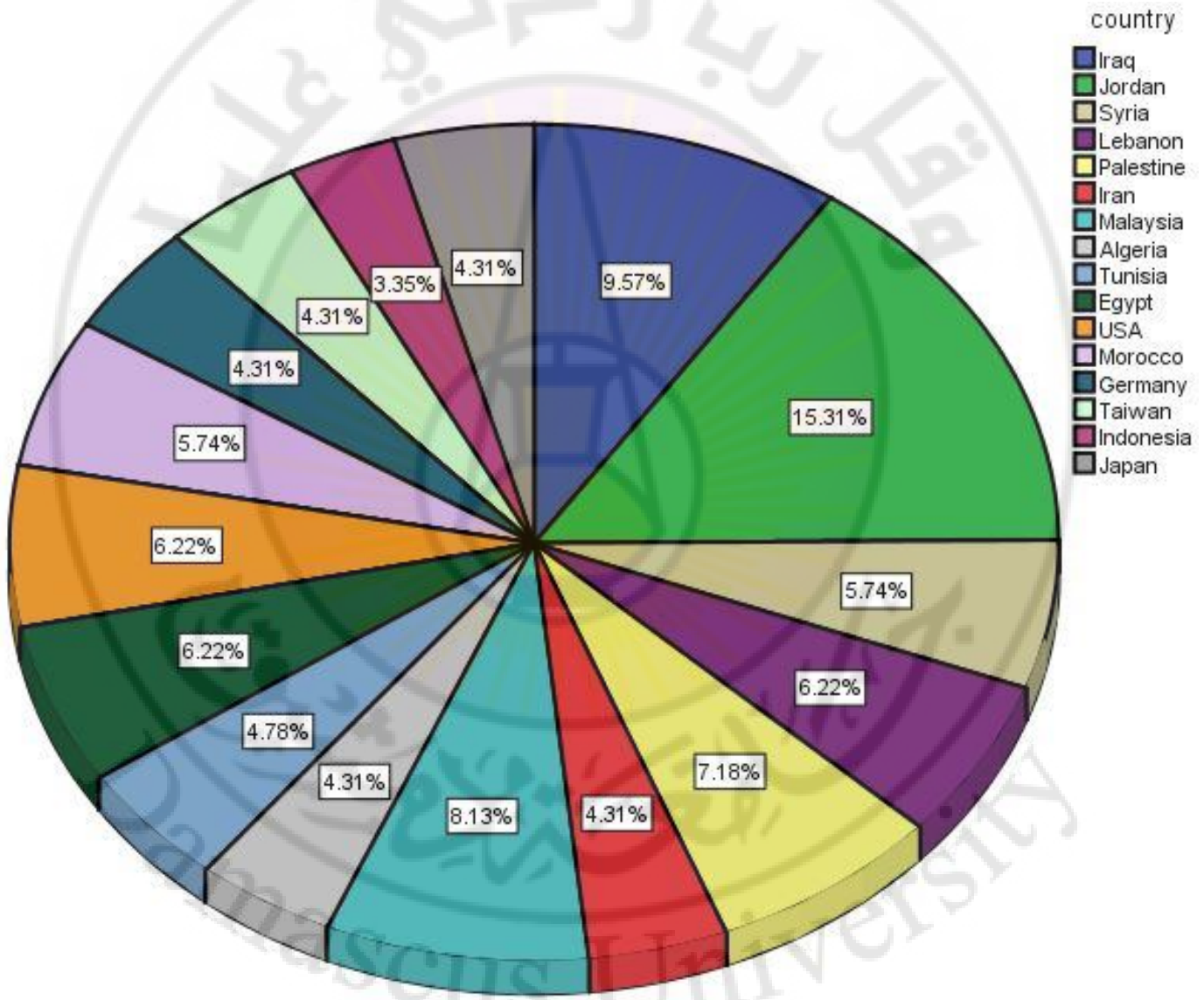
رسم الفطيرة مناسب للمتغير الفئوي مع فئات لا يزيد عددها عن ست فئات.

البحث 2: التمثيل البياني للمشاهدات

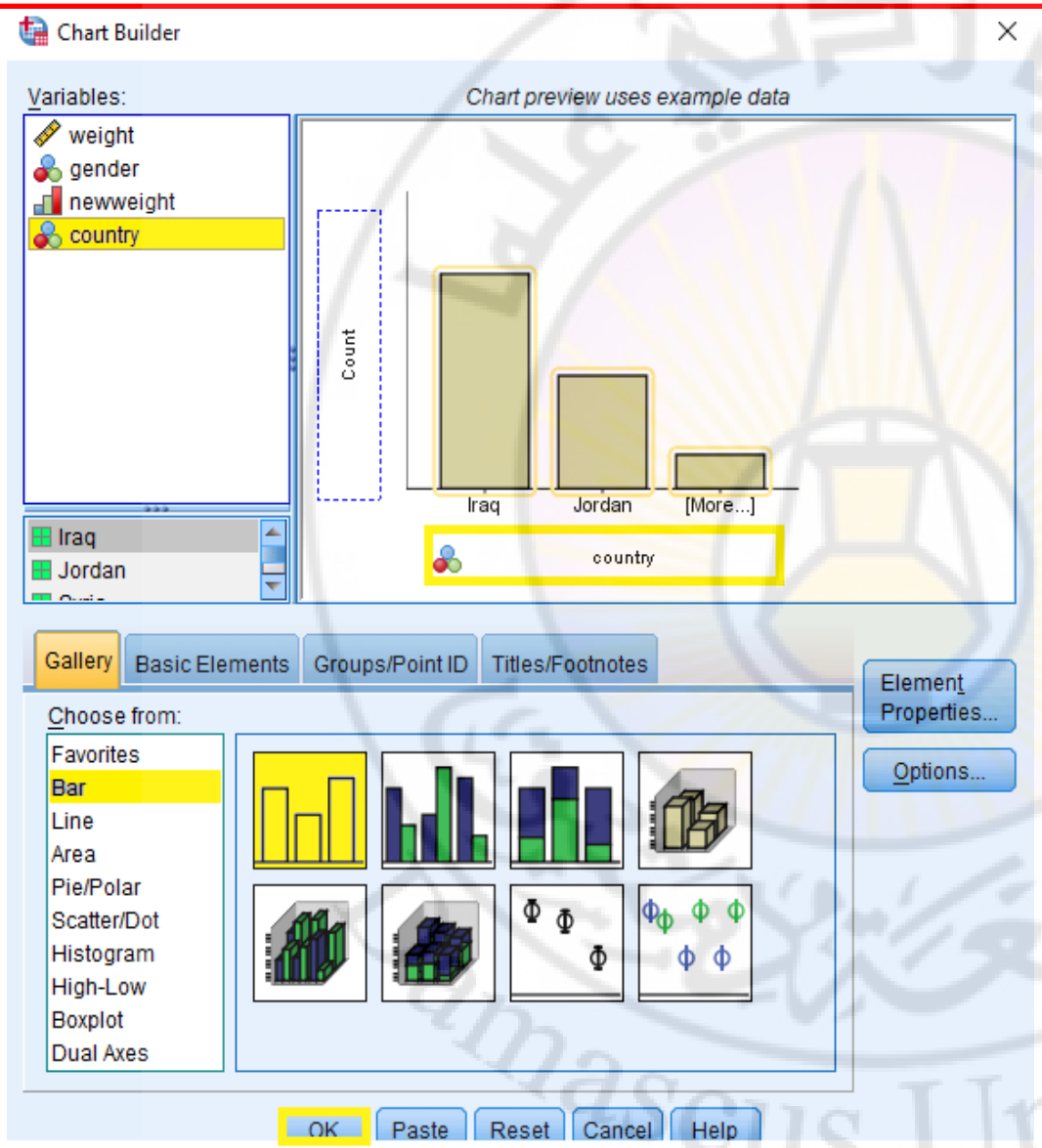
	country	Valid N	I
weight	Iraq	20	
	Jordan	32	
	Syria	12	
	Lebanon	13	
	Palestine	15	
	Iran	9	
	Malaysia	17	
	Algeria	9	
	Tunisia	10	
	Egypt	13	
	USA	13	
	Morocco	12	
	Germany	9	
	Taiwan	9	
	Indonesia	7	
	Japan	9	



البحث 2: التمثيل البياني للمشاهدات



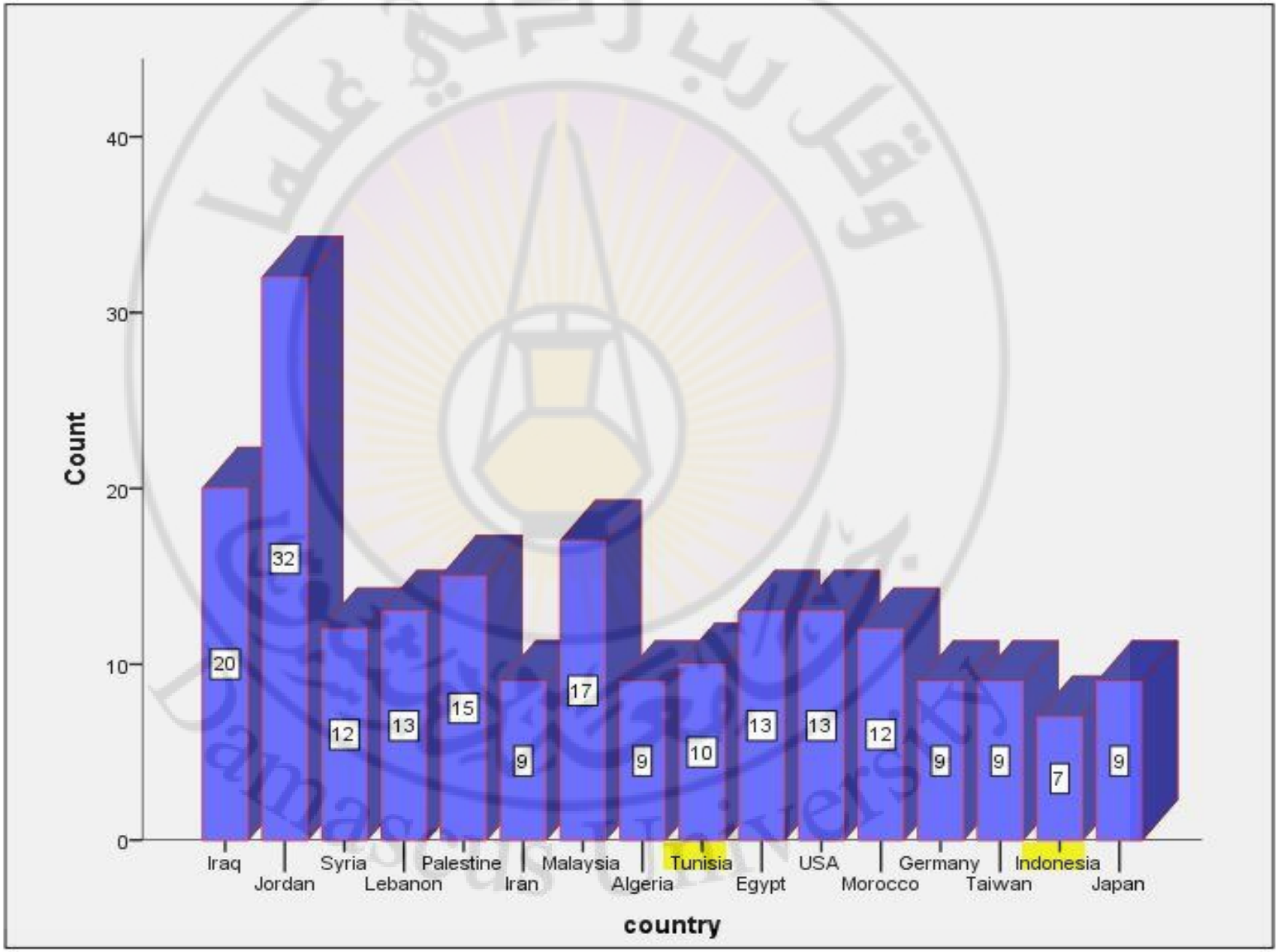
البحث 2: التمثيل البياني للمشاهدات



ثانياً مخطط (رسم) الأعمدة Bar chart:

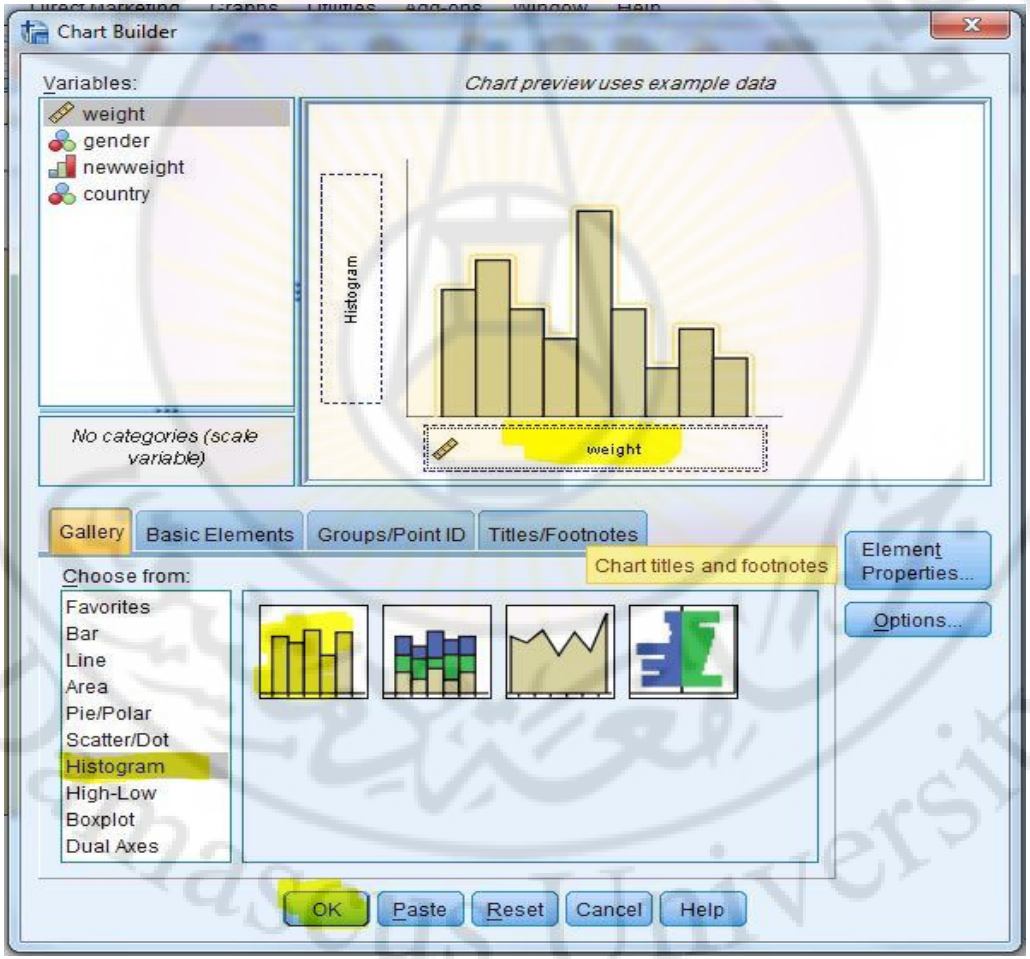
رسم بياني على شكل أعمدة (bars or bins) متباعدة فيما بينها بمسافات متساوية، وذلك لتمثيل المتغيرات الفئوية (اسمية أم ترتيبية) مهما كان عدد فئاتها.

البحث 2: التمثيل البياني للمشاهدات

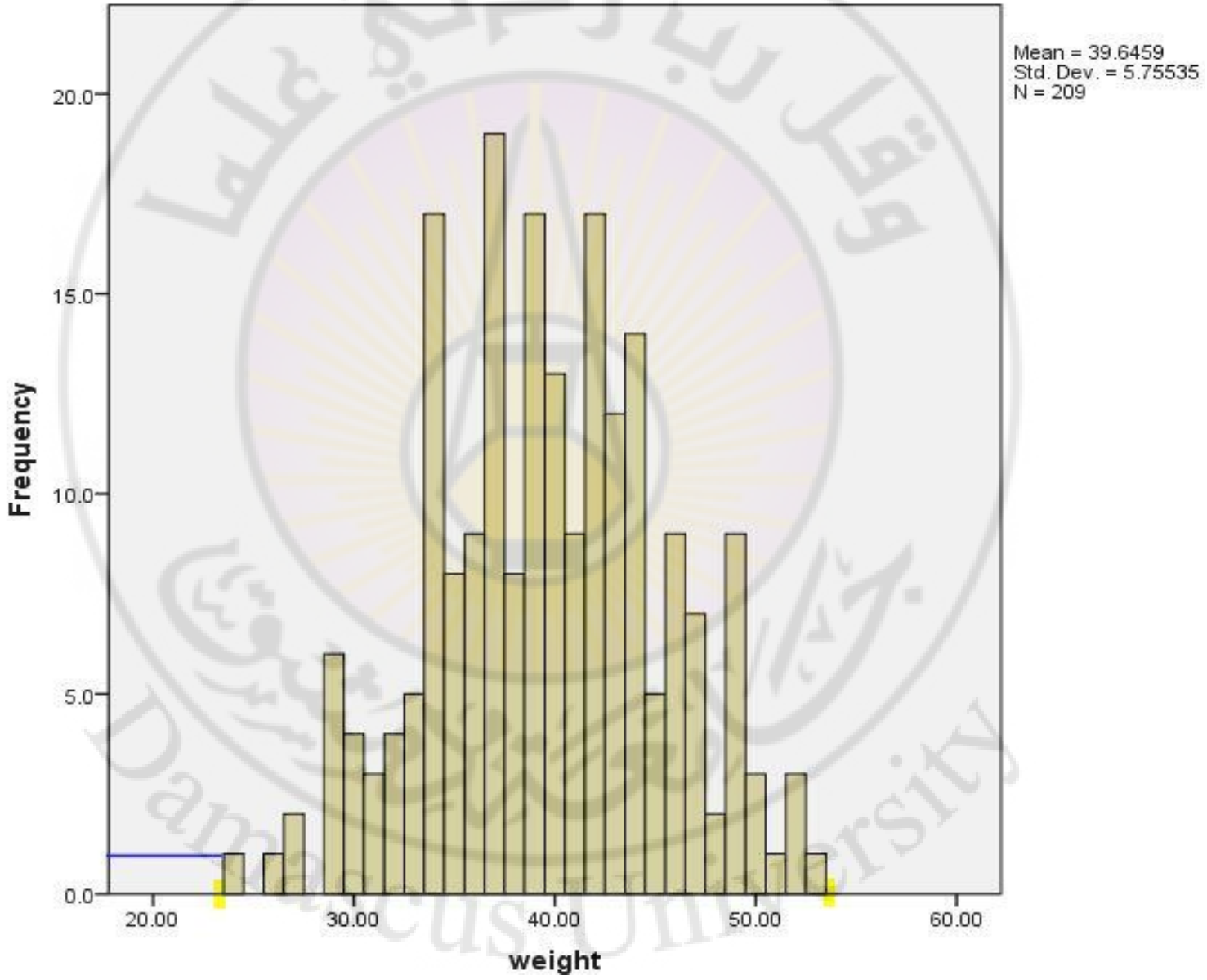


ثالثاً المدرّج التكراري Histogram

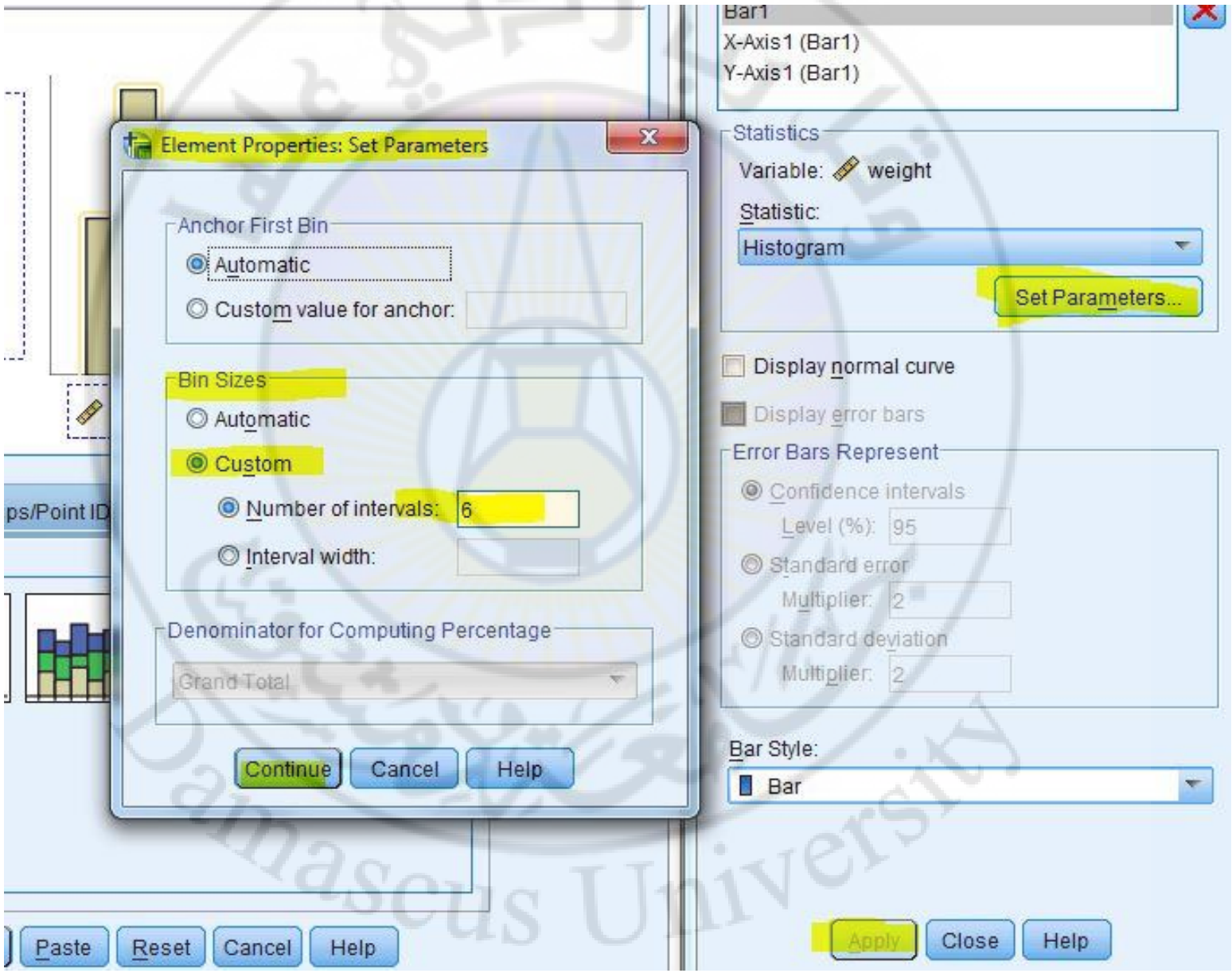
وهو الرسم البياني المستخدم لتمثيل مشاهدات المتغير المستمر بيانياً.



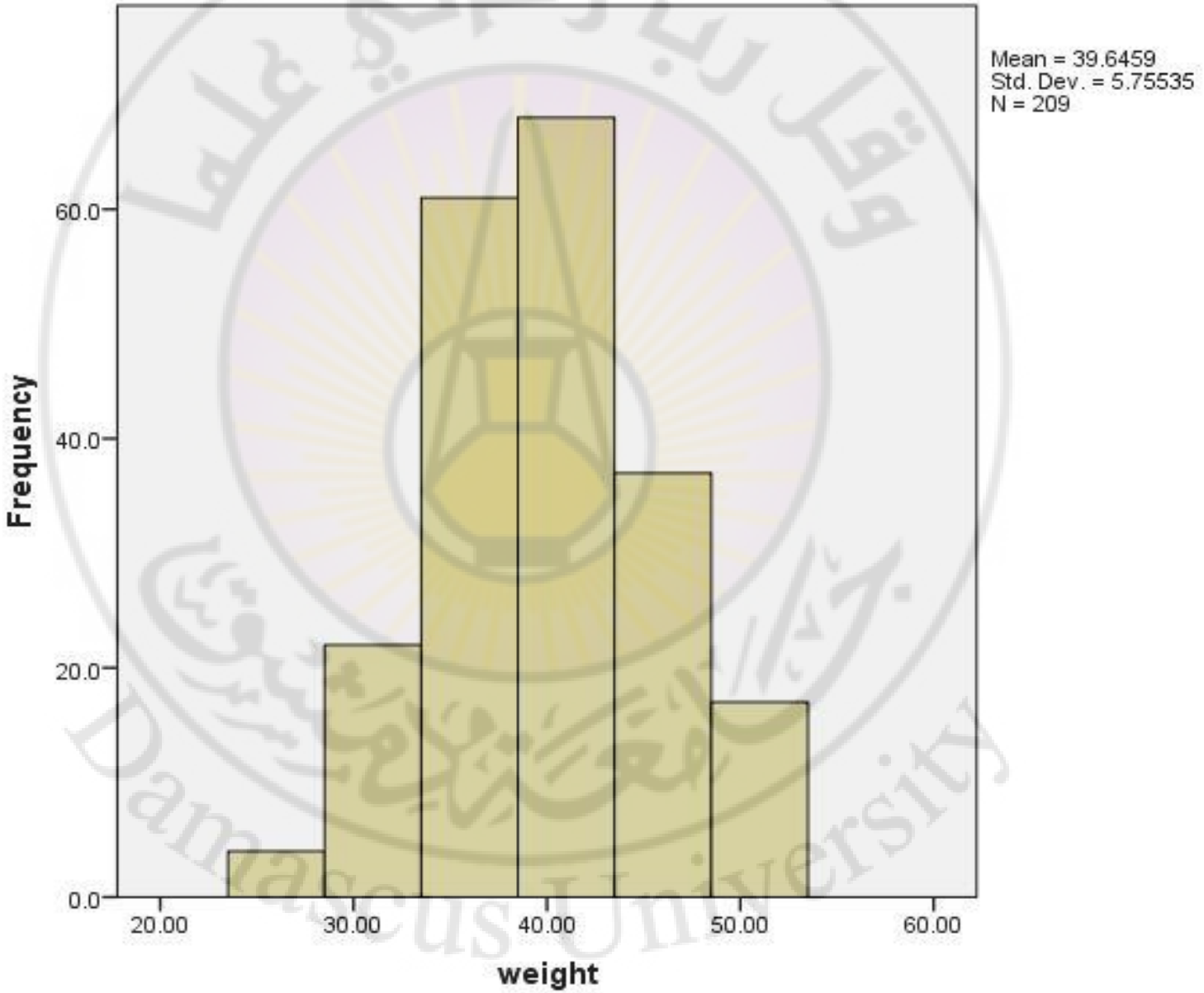
البحث 2: التمثيل البياني للمشاهدات



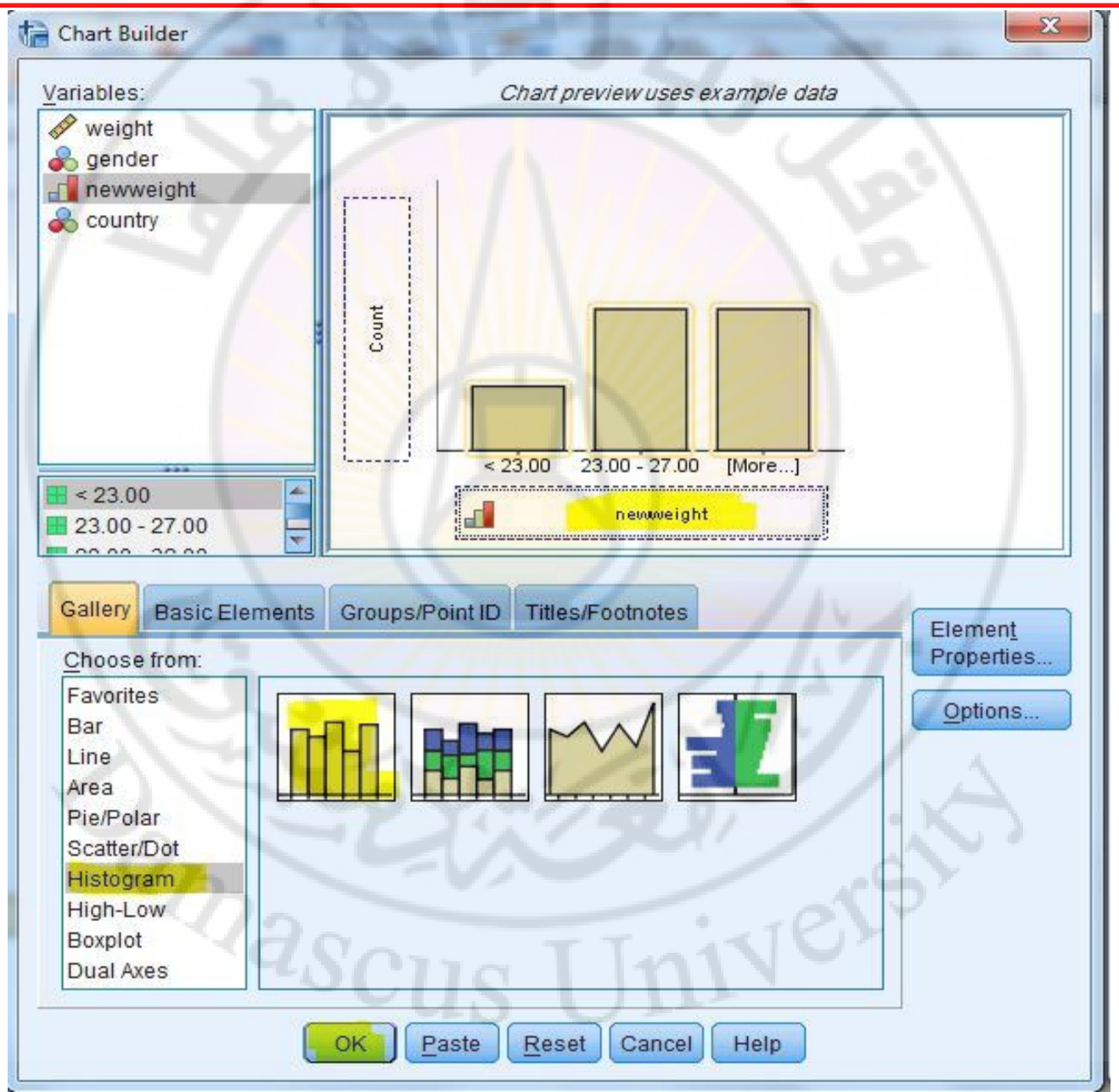
البحث 2: التمثيل البياني للمشاهدات



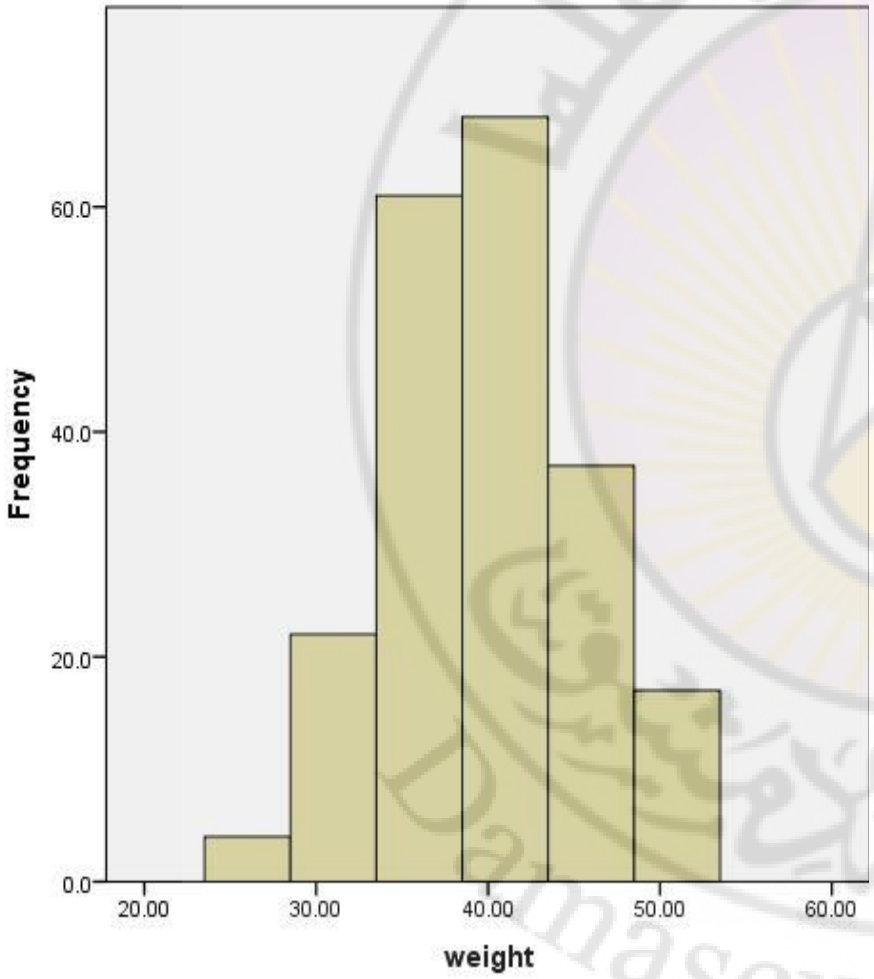
البحث 2: التمثيل البياني للمشاهدات

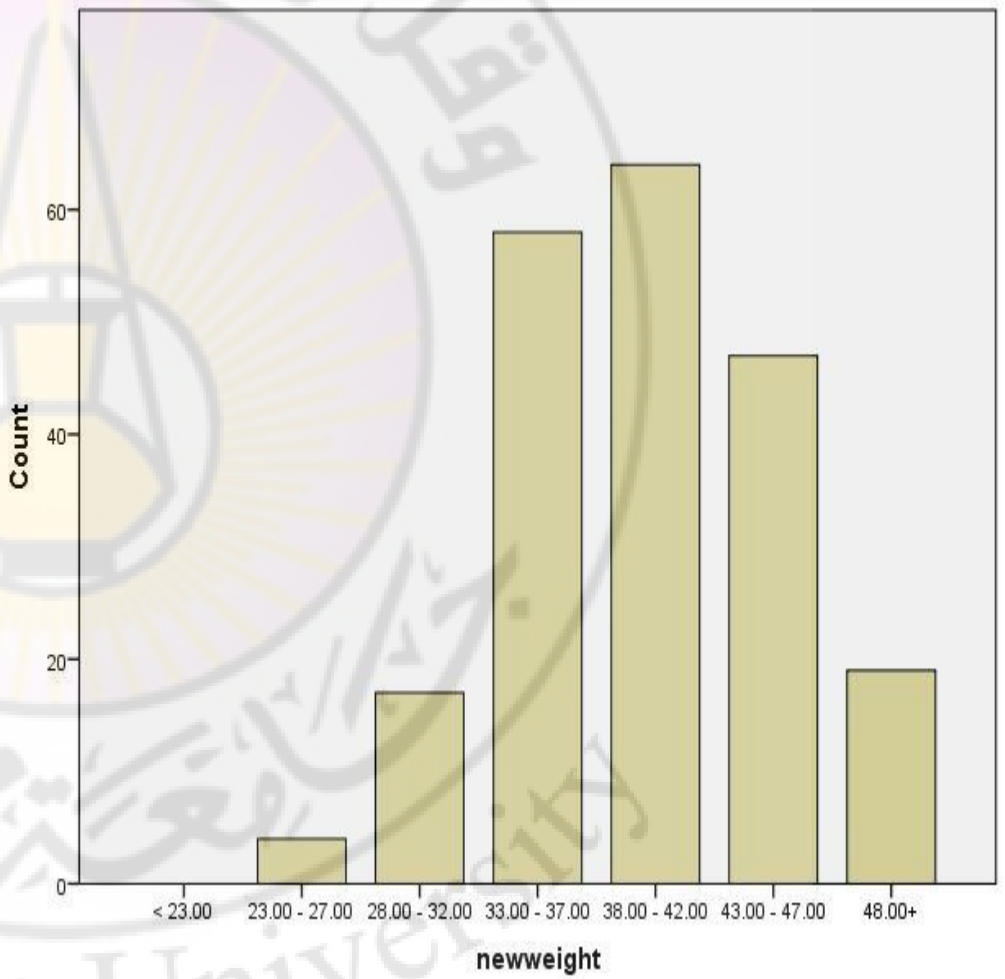


البحث 2: التمثيل البياني للمشاهدات



البحث 2: التمثيل البياني للمشاهدات





رابعاً المضلع التكراري polygon

وهو الرسم البياني المستخدم لتمثيل مشاهدات المتغير المستمر بيانياً بحيث تتصل مراكز الفئات (المجالات) ببعضها بخطوط مستقيمة (عادة يتم رسم المضلع التكراري على المدرج التكراري).

مركز الفئة midpoint هو $\frac{a+b}{2}$

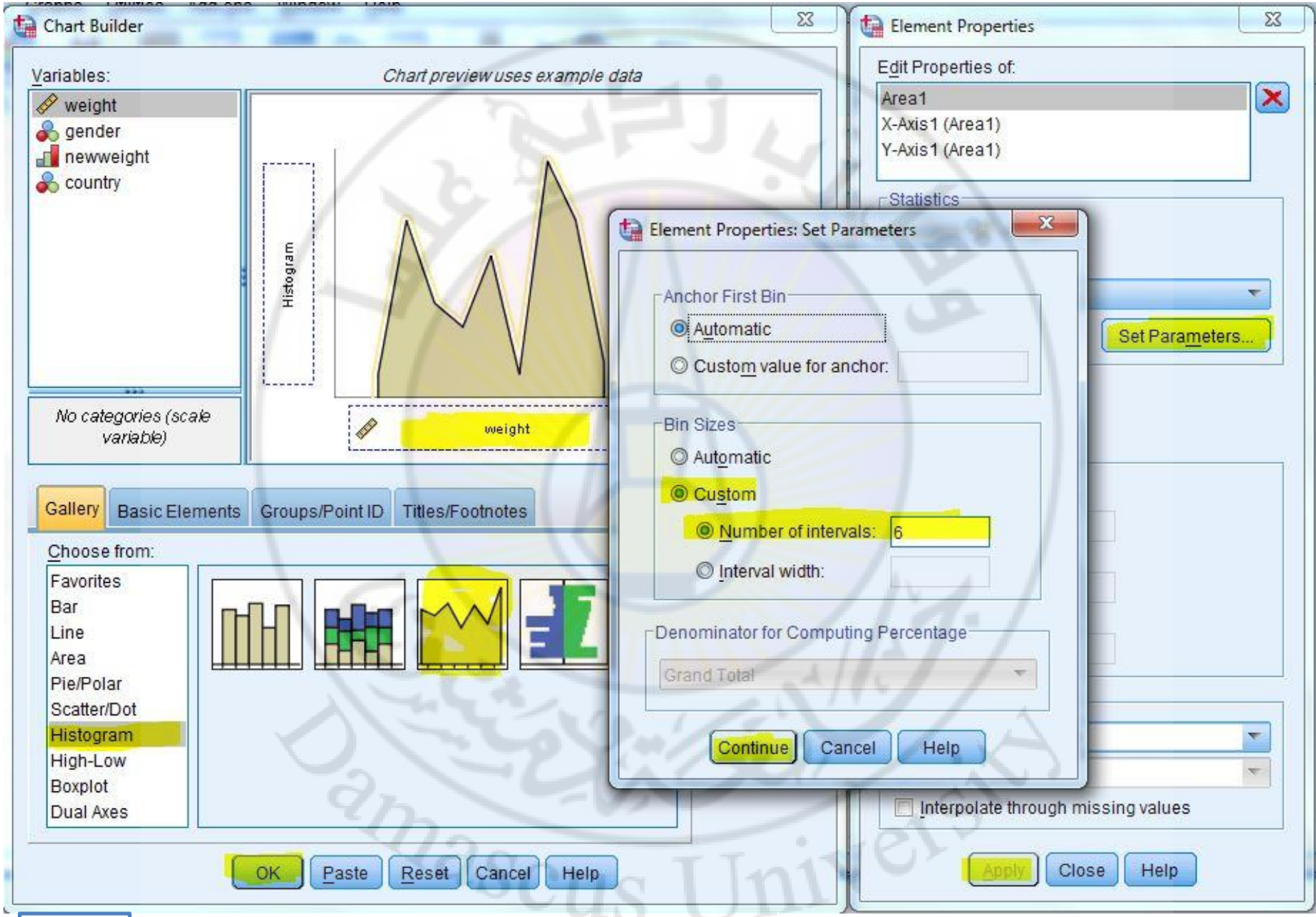
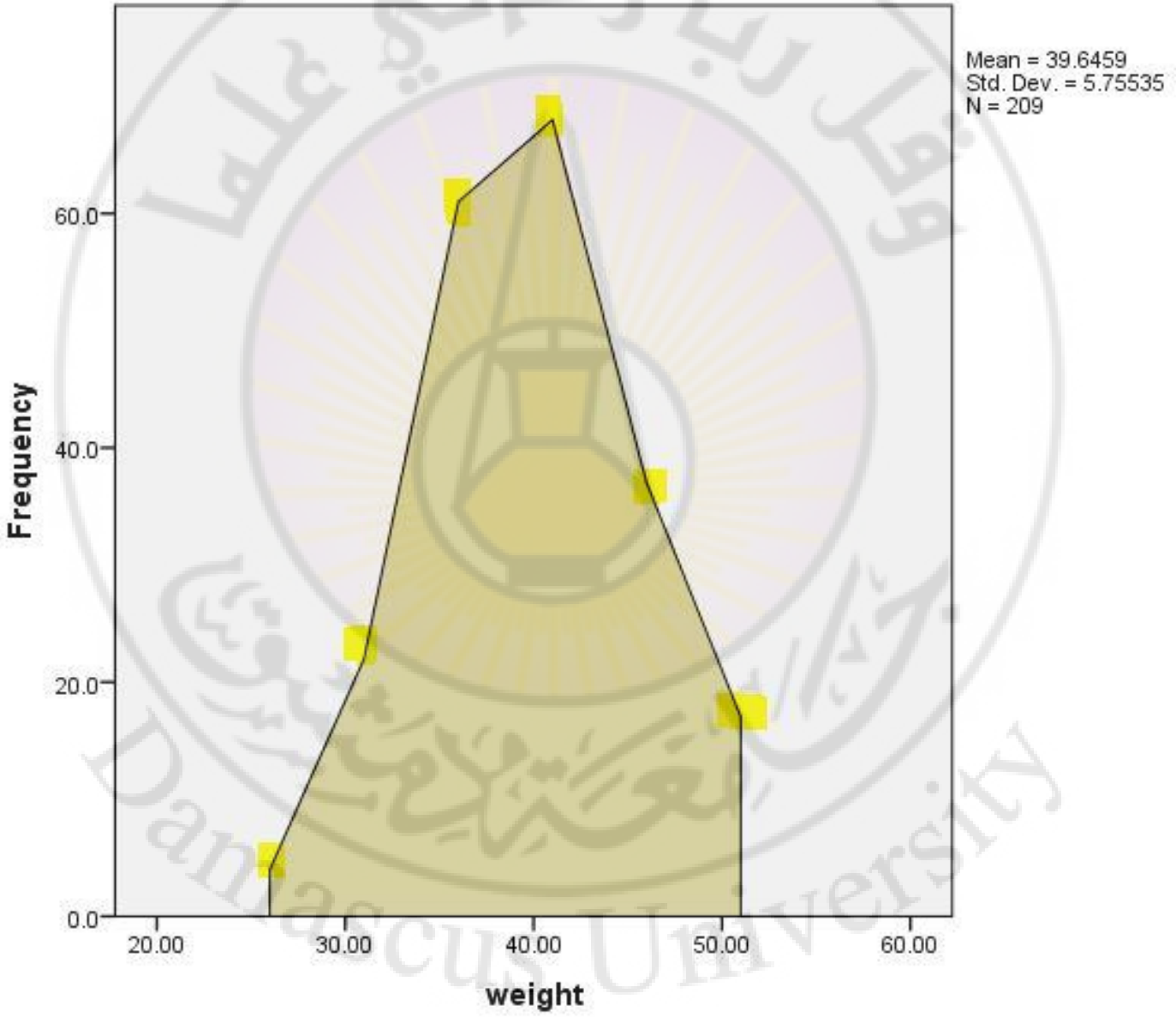


Chart Builder
Variables:
weight
gender
newweight
country
Chart preview uses example data
Histogram
weight
No categories (scale variable)
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes
Choose from:
Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot
Histogram
High-Low
Boxplot
Dual Axes
OK
Paste
Reset
Cancel
Help
Element Properties
Edit Properties of:
Area1
X-Axis1 (Area1)
Y-Axis1 (Area1)
Statistics
Set Parameters...
Interpolate through missing values
Apply
Close
Help
Element Properties: Set Parameters
Anchor First Bin
Automatic
Custom value for anchor:
Bin Sizes
Automatic
Custom
Number of intervals: 6
Interval width:
Denominator for Computing Percentage
Grand Total
Continue
Cancel
Help

البحث 2: التمثيل البياني للمشاهدات



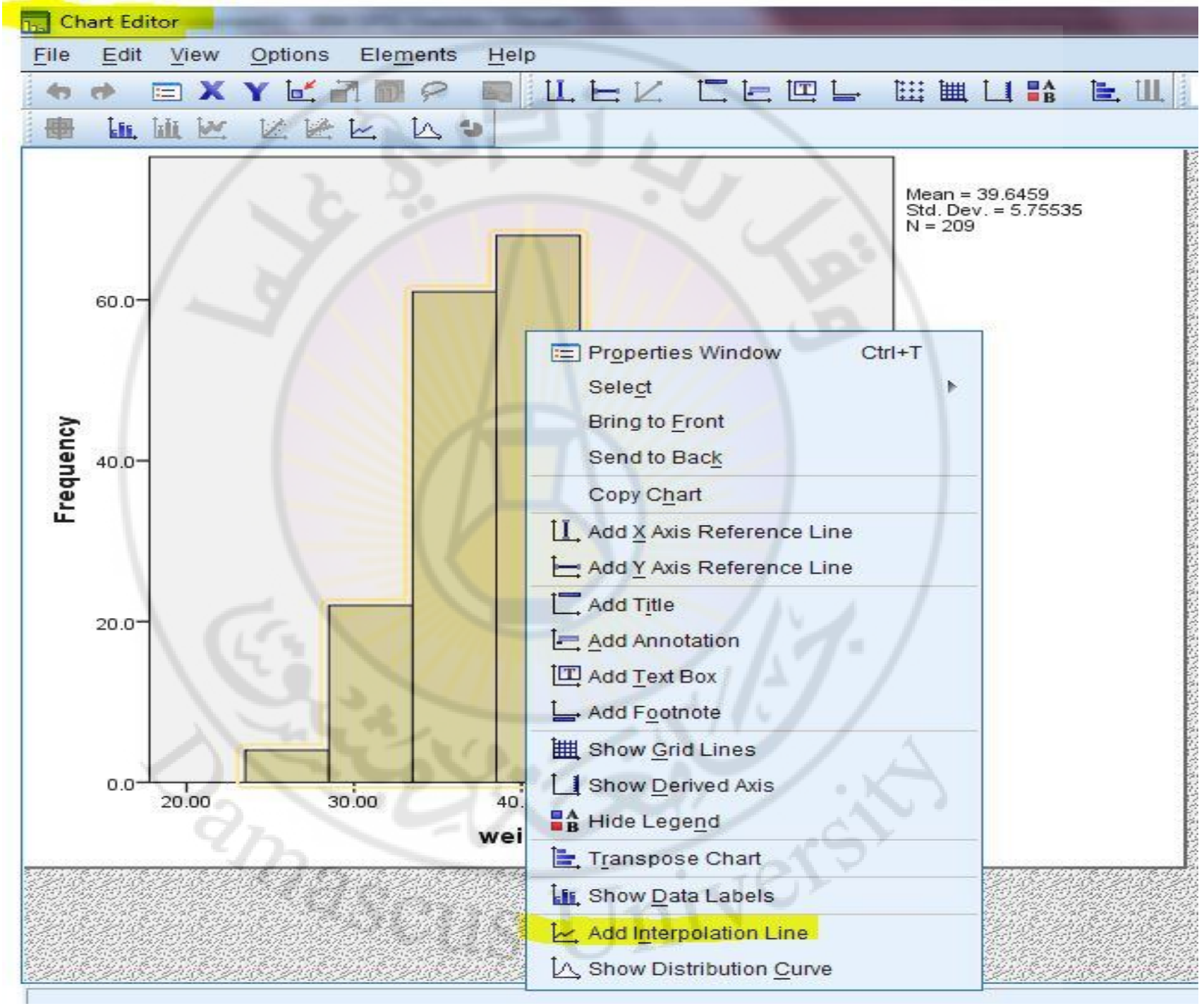
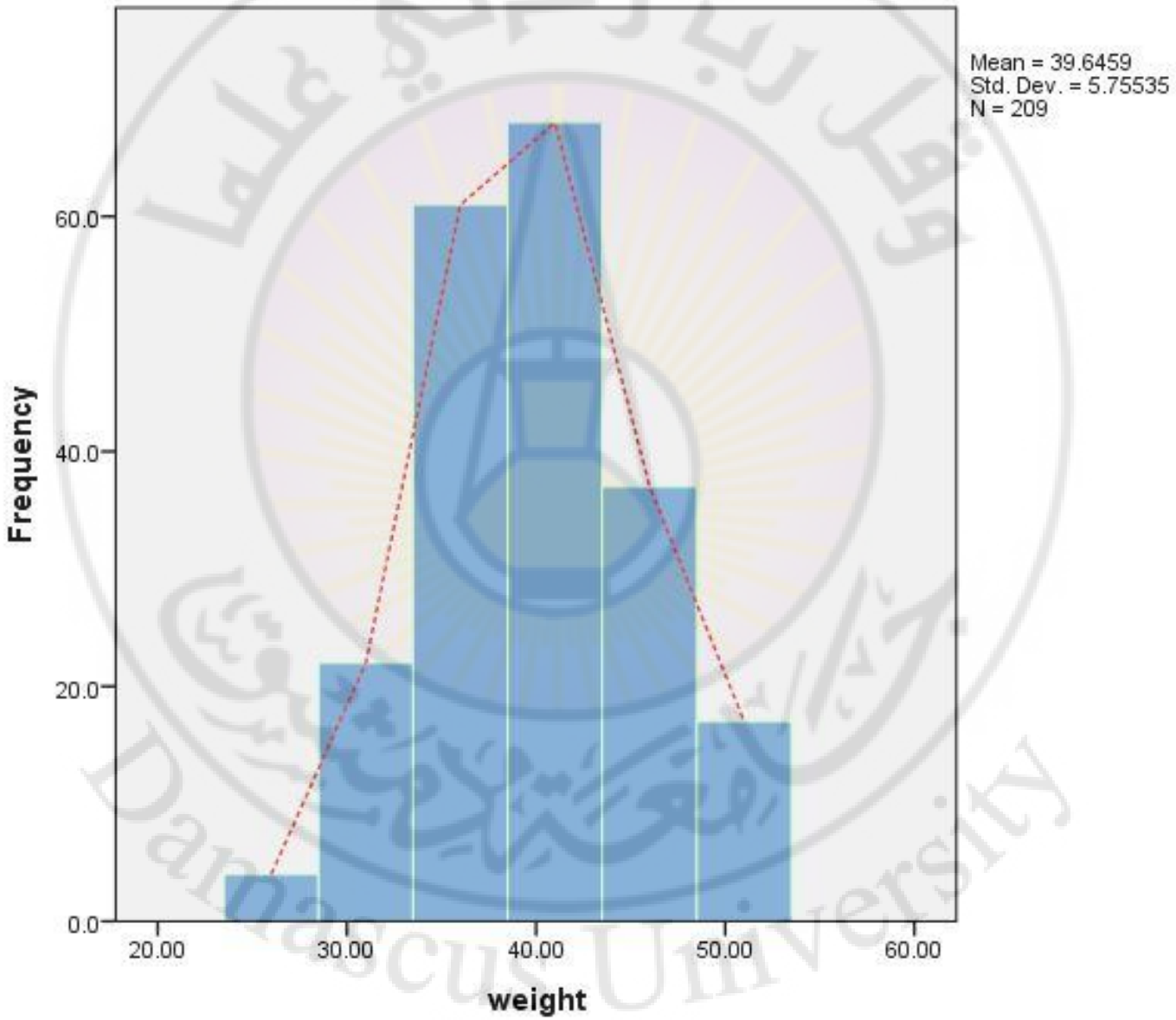


Chart Editor
File Edit View Options Elements Help
Mean = 39.6459
Std. Dev. = 5.75535
N = 209
Frequency
60.0
40.0
20.0
0.0
20.00
30.00
Properties Window Ctrl+T
Select
Bring to Front
Send to Back
Copy Chart
Add X Axis Reference Line
Add Y Axis Reference Line
Add Title
Add Annotation
Add Text Box
Add Footnote
Show Grid Lines
Show Derived Axis
Hide Legend
Transpose Chart
Show Data Labels
Add Interpolation Line
Show Distribution Curve

البحث 2: التمثيل البياني للمشاهدات



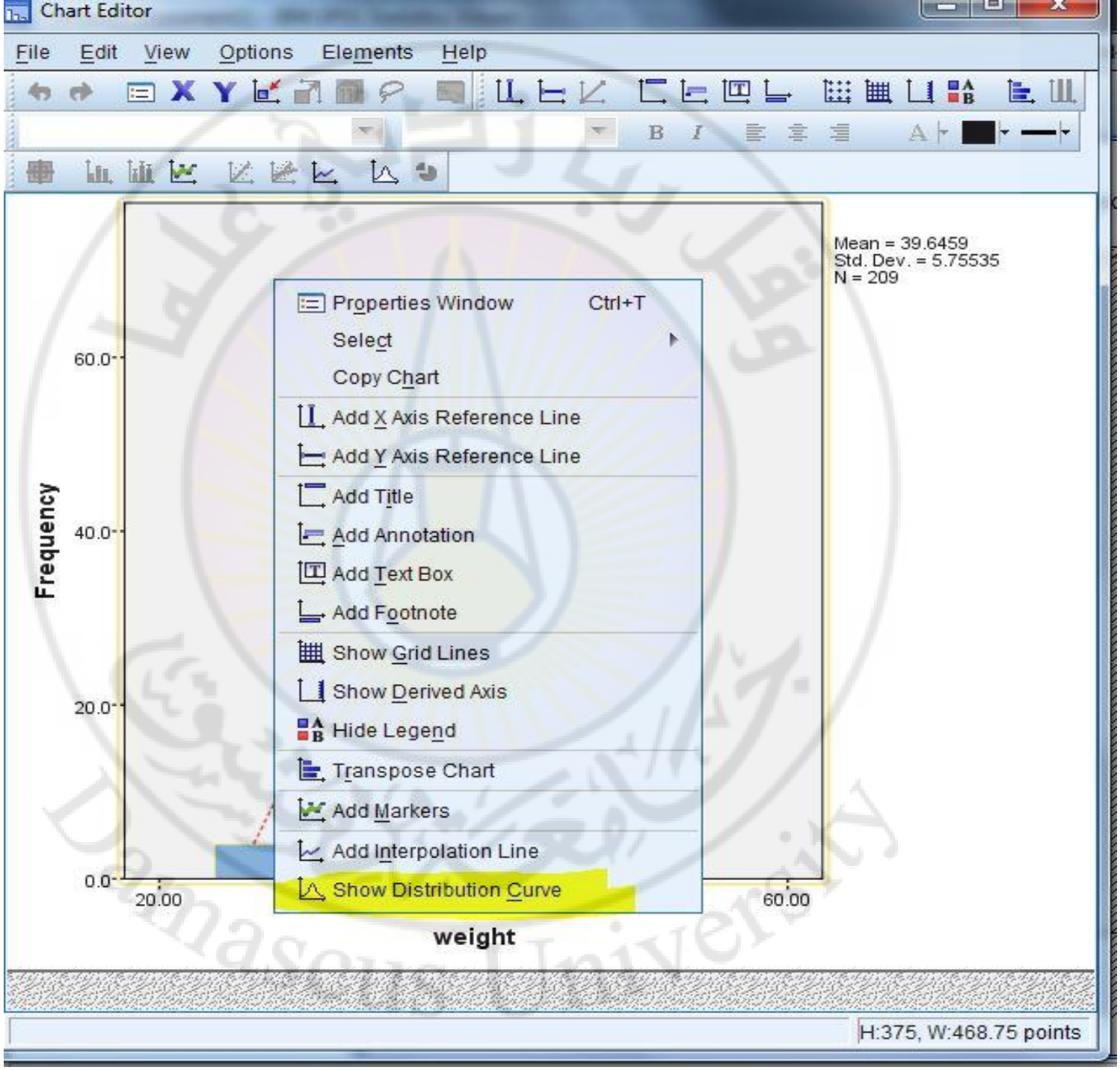
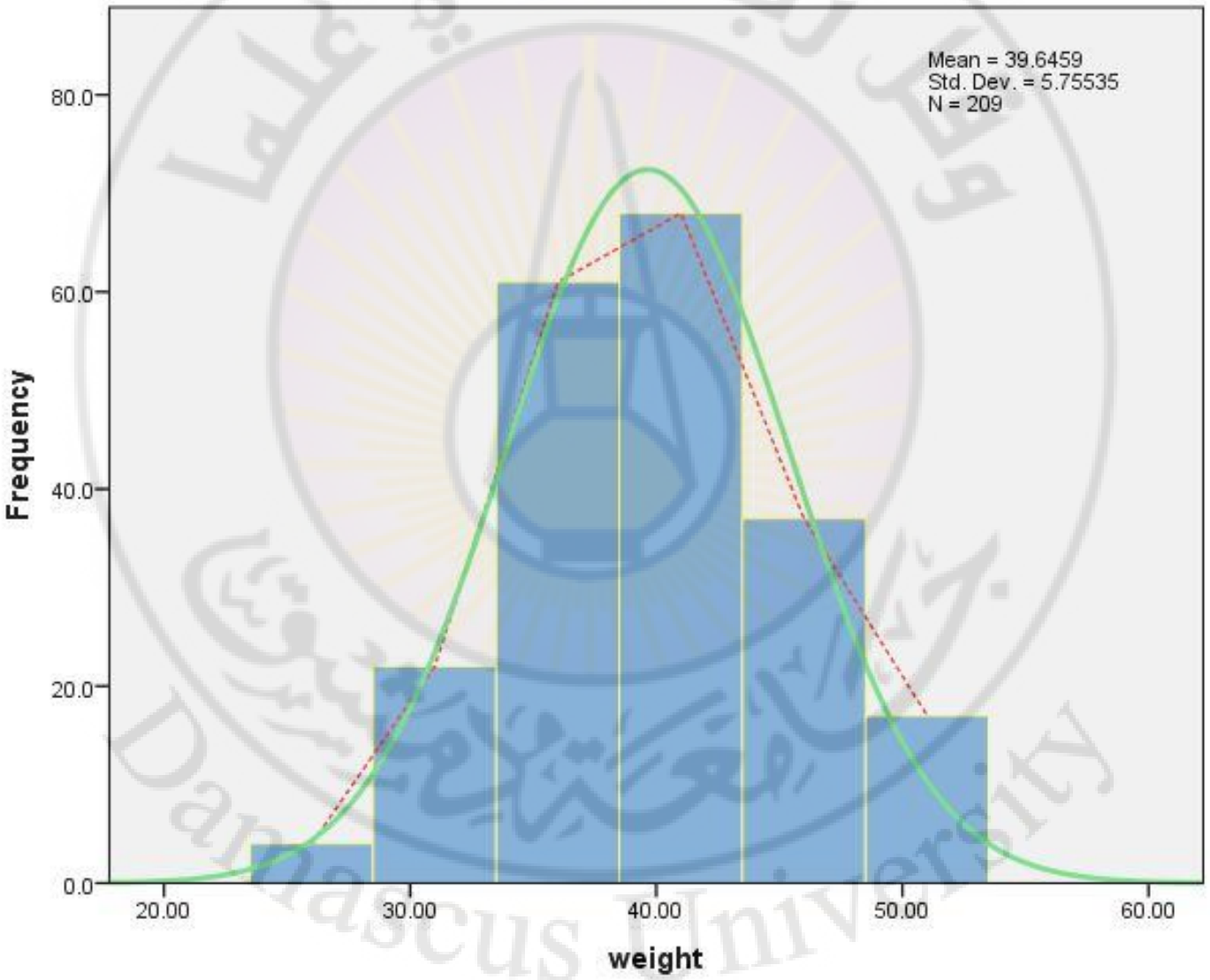


Chart Editor
File Edit View Options Elements Help
Mean = 39.6459
Std. Dev. = 5.75535
N = 209
Properties Window Ctrl+T
Select
Copy Chart
Add X Axis Reference Line
Add Y Axis Reference Line
Add Title
Add Annotation
Add Text Box
Add Footnote
Show Grid Lines
Show Derived Axis
Hide Legend
Transpose Chart
Add Markers
Add Interpolation Line
Show Distribution Curve
Frequency
60.0
40.0
20.0
0.0
20.00
60.00
weight
H:375, W:468.75 points

البحث 2: التمثيل البياني للمشاهدات



الإحصاء الحيوي

لطلاب السنة الثانية والثالثة - كلية العلوم الصحية

جامعة دمشق

”الجلسة السابعة“

مدرس المقرر: أ. سلام الجرّاح

إعداد: د. ياسر الزعيم

المدرس في قسم الإحصاء الرياضي

كلية العلوم - جامعة دمشق

آيار– 2023

نتكلم اليوم عن:

- **البحث الخامس: التوزيع الطبيعي**
- مقدمة في اختبار الفرضيات الإحصائية
- اختبار التوزع الطبيعي في SPSS

التوزيع الطبيعي (التوزيع الغاوسي)

Normal distribution (Gaussian)

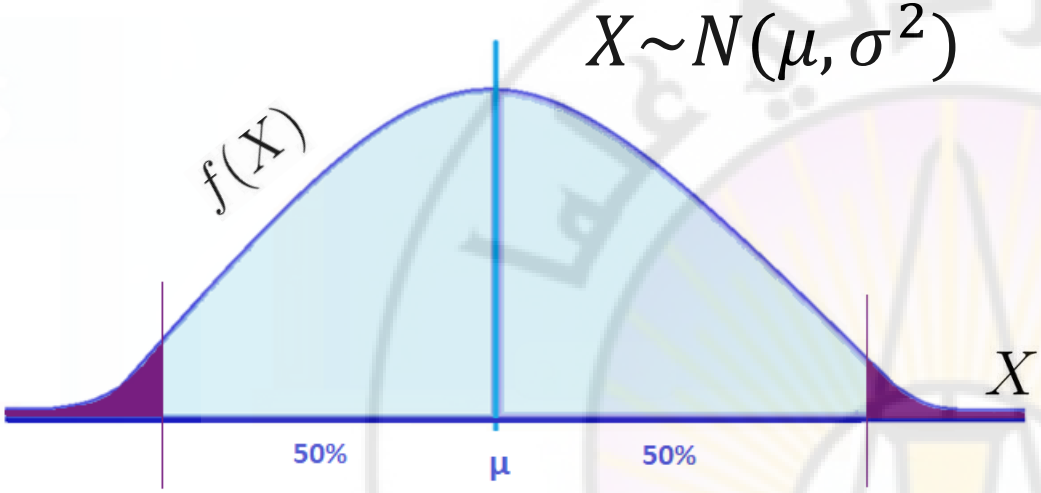
إنّ للمتحول العشوائي X توزيعياً طبيعياً بمتوسط μ وتباين σ^2، ونكتب اصطلاحاً $X \sim N(\mu,\sigma^2)$ إذا كانت معادلة المنحني التكراري له هي (أي إذا كان للمتحول تابع الاحتمال) المعادلة:

$$f(X_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} ; X_i \in]-\infty, +\infty[$$

$e = 2.718$ (Euler's number)

$\pi = 3.14159..(pi)$

البحث الخامس: التوزيع الطبيعي



بعض خصائص هذا التوزيع:

1) إنّ $f(X_i) \geq 0$

2) إنّ $\int_{X_i=-\infty}^{\infty} f(X_i)\,\mathrm{d}X_i = 1$

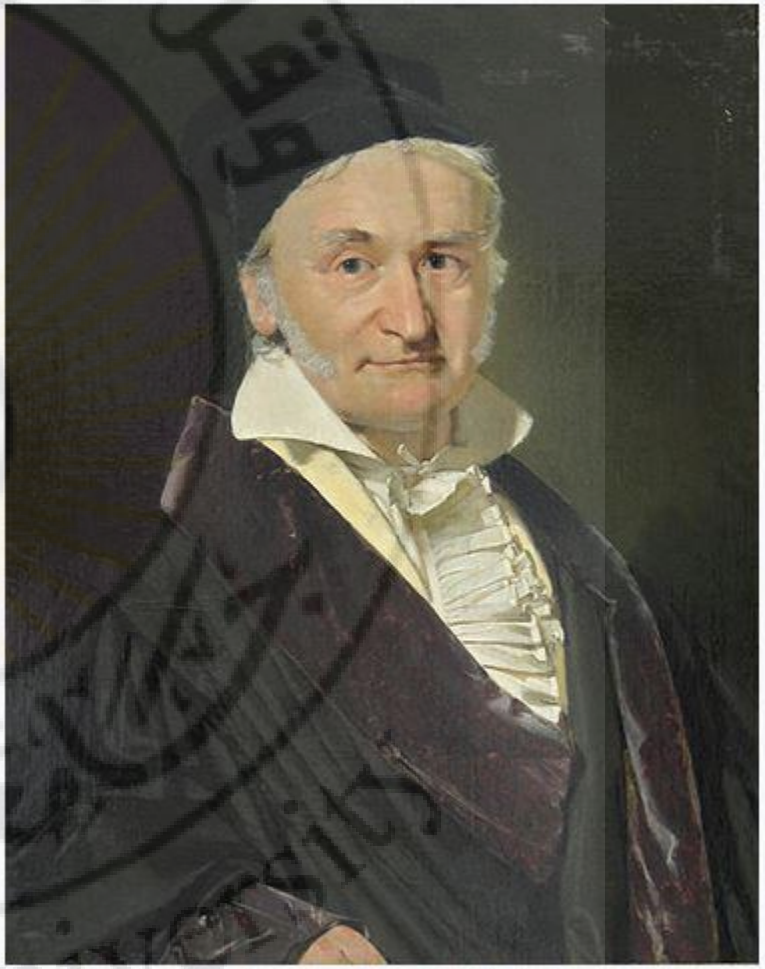
3) التوزيع متناظر حول المتوسط μ

4) إنّ $\mu = \text{mode} = median$

5) **القاعدة التجريبية**

البحث الخامس: التوزيع الطبيعي

Normal (or Gaussian) distribution التوزيع الغاوصي (الغاوسي)



Carl Friedrich Gauß (1777–1855)

إذا كان $X \sim N(\mu,\sigma^2)$ فإنّ $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ أي أنّ:

$$f(Z_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}} \qquad ; z_i \in]-\infty, +\infty[$$

* **غالباً معالم المجتمع مجهولة لذا نقدر μ بـ $\overline{X}$ ، ونقدر σ^2 بـ s^2**

ونكتب $\hat{\mu} = \overline{X}$ (mu hat) و $\hat{\sigma^2} = s^2$ (sigma square hat)

فإذا كان $X \sim N(\bar{X},s^2)$ فإنّ:

$$Z = \frac{X - \bar{X}}{s} \sim N(0,1)$$

القاعدة التجريبية The empirical rule (The 68, 95, 99.7 rule)

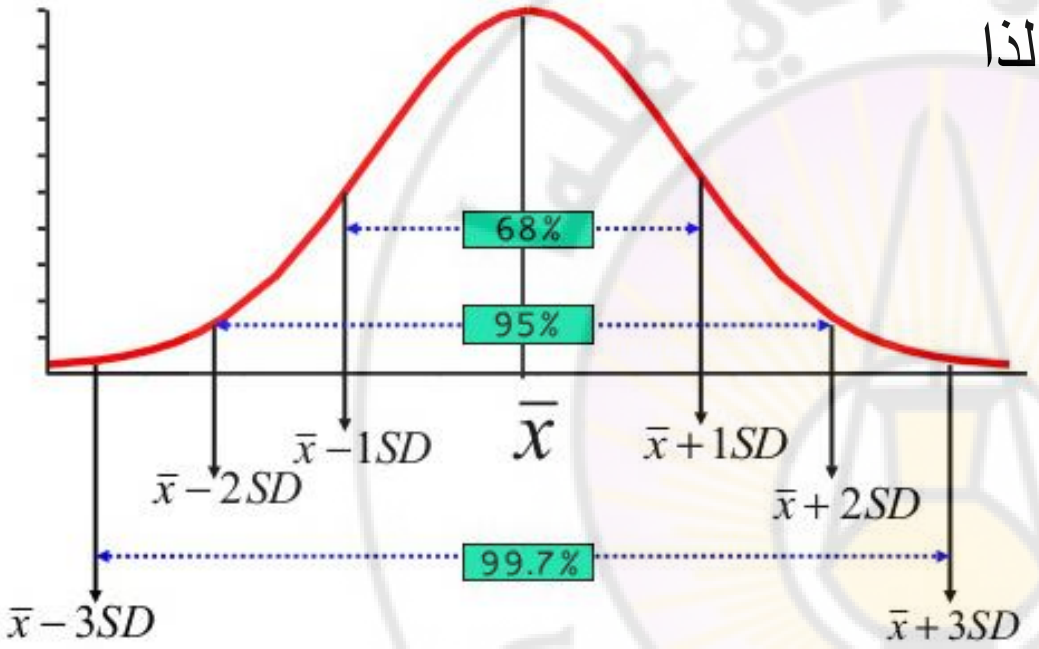
$X_1, X_2, \ldots, X_N$ عينة لـ X وكان $X \sim N(\mu, \sigma^2)$ فإنّ:

68% من المشاهدات تتوضع بين $\mu - \sigma$ و $\mu + \sigma$.

95% من المشاهدات تتوضع بين $\mu - 2\sigma$ و $\mu + 2\sigma$.

99.7% من المشاهدات تتوضع بين $\mu - 3\sigma$ و $\mu + 3\sigma$.

البحث الخامس: التوزيع الطبيعي



وكون غالباً معالم المجتمع مجهولة لذا

$$\widehat{\mu} = \overline{X} \quad \text{و} \quad \widehat{\sigma^2} = s^2$$

وبالتالي القاعدة التجريبية تصبح:

68% من المشاهدات تتوضع بين $\bar{X} - s$ و $\bar{X} + s$.

95% من المشاهدات تتوضع بين $\bar{X} - 2s$ و $\bar{X} + 2s$.

99.7% من المشاهدات تتوضع بين $\bar{X} - 3s$ و $\bar{X} + 3s$.

أسلوب اختبار الفرضيات الإحصائي

مقدمة

عينة أولى

$$X : 20, 40, 35, 60, 55 \rightarrow \hat{\mu} = \bar{x} = 42 \quad \& \quad s_X = 16.05$$

عينة ثانية

$$Y : 100, 80, 90, 75, 82 \rightarrow \hat{\mu} = \bar{y} = 85.4 \quad \& \quad s_Y = 9.8$$

$$H_0 : \mu = a \qquad \text{versus} \qquad H_1 : \mu \neq a$$

مصطلحات في اختبار الفرضيات

اختبار الفروض الإحصائي statistical hypotheses testing:

هو علم الوصول إلى استدلالات واستقراءات إحصائية حول معالم ووسطاء المجتمع من خلال مشاهدات عينة مسحوبة ومنتقاة من ذلك المجتمع الإحصائي (بالرغم من أنّ المشاهدات معرضة للخطأ).

H_0

وهي الفرضية الصفرية (الفرضية الابتدائية) null hypothesis وفيها نفترض صحة ادعاء ما (أحياناً يرمز لها بـ H_O)

H_1

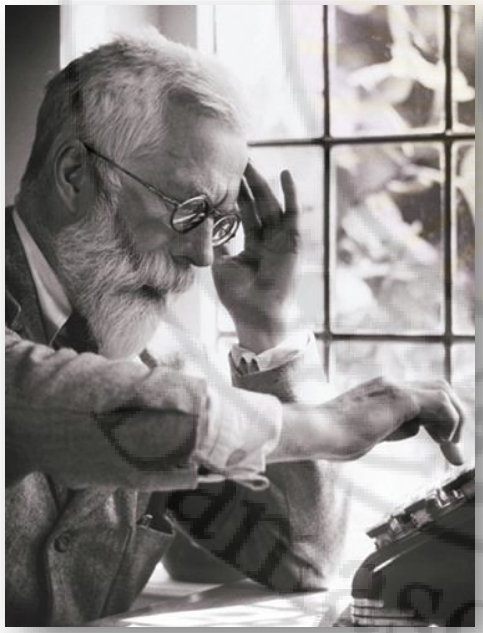
وهي الفرضية البديلة (الفرض واحد) alternative hypothesis وفيها نفترض عدم صحة الادعاء في الفرضية الصفرية (يرمز لها H_a)

مصطلحات في اختبار الفرضيات

α هو مستوى الأهمية النظري significance level:

ونحدد قيمته قبل إجراء الاختبار (ويسمى كذلك بمستوى المعنوية أو مستوى الدلالة). ويتم **تعريفه:**

بأنّه حجم الخطأ من النوع الأول (type one error)، حيث أنّ هذا الخطأ هو رفض الفرضية H_0 (علماً بأنها صحيحة) وقبول الفرض البديل H_1 .



Ronald Fisher 1890 – 1962

بمعنى آخر

إنّ α هو الخطأ الذي نقبل به عند رفض H_0 وهي صحيحة.

ومن أشهر القيم:

$$\alpha = .01,\ \alpha = .05,\ \alpha = .10$$

Statistical Methods for Research Workers (1925)

مصطلحات في اختبار الفرضيات

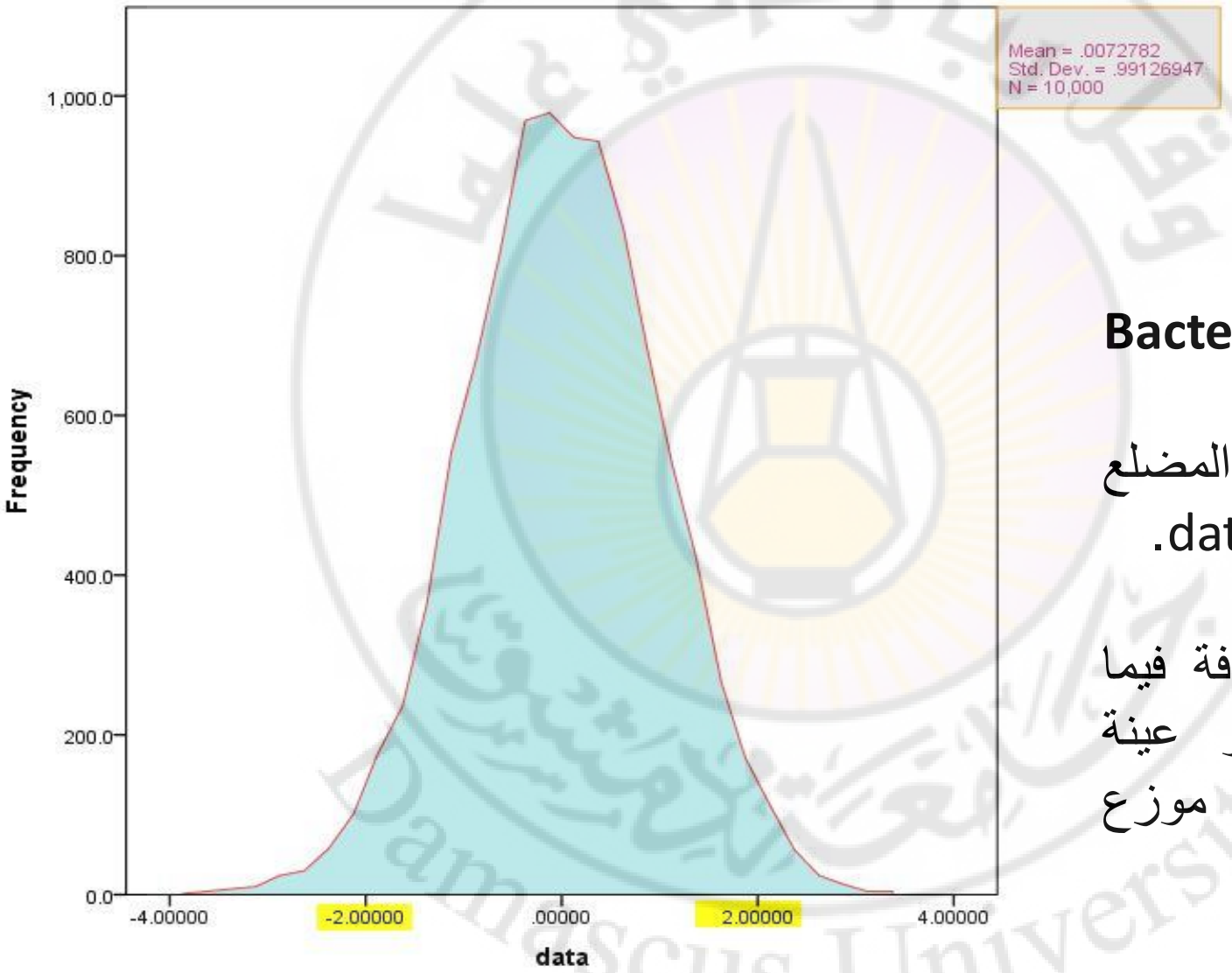
sig وهو مستوى الأهمية الفعلي **significance level**
وقيمته محسوبة من الاختبار (يرمز له أيضاً $p.value$ أو $p.v.$).
وهو الخطأ الحقيقي المرتكب عند رفض H_0 وهي صحيحة.

القاعدة الثابتة لاتخاذ القرار في اختبار الفرضيات:

إذا كان $sig \geq \propto$ فإننا نقبل H_0 و نرفض H_1

أمّا إذا كان $sig < \propto$ فإننا نرفض H_0 و نقبل H_1

اختبار التوزع الطبيعي



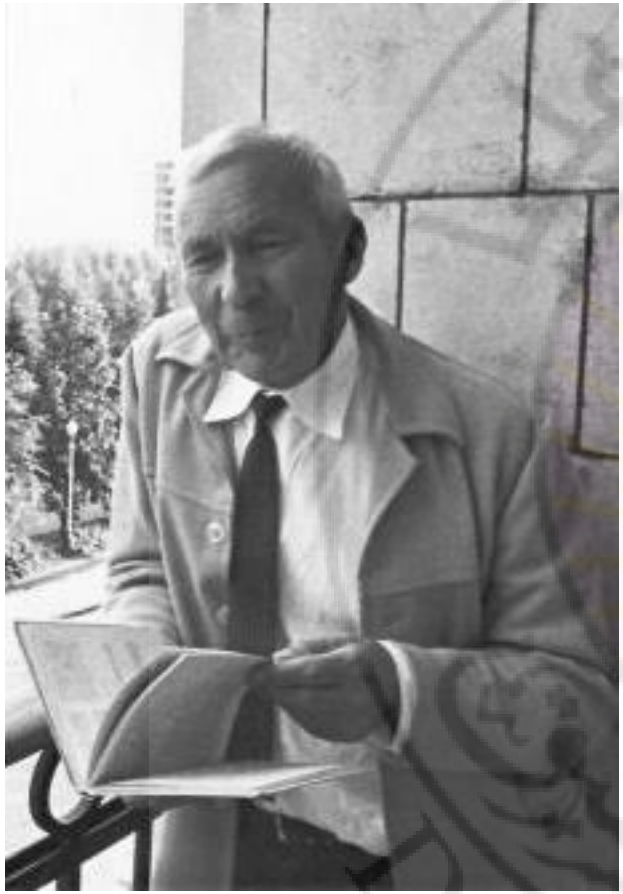
Bacteria.sav

الشكل المجاور هو المضلع التكراري للمتحول data.

الهدف اختبار ومعرفة فيما إذا كان data هو عينة عشوائية من مجتمع موزع طبيعياً؟

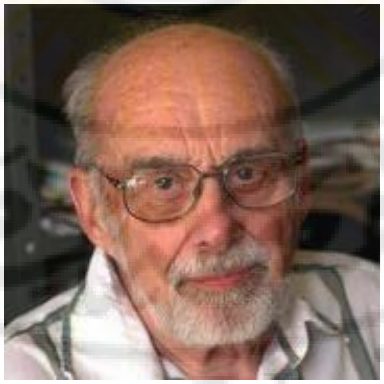
نريد اختبار $\text{data} \sim N(\mu,\sigma^2)$ مقابل $\text{data} \nsim N(\mu,\sigma^2)$

عند مستوى الأهمية α.



Andrey Kolmogorov 1903-1987

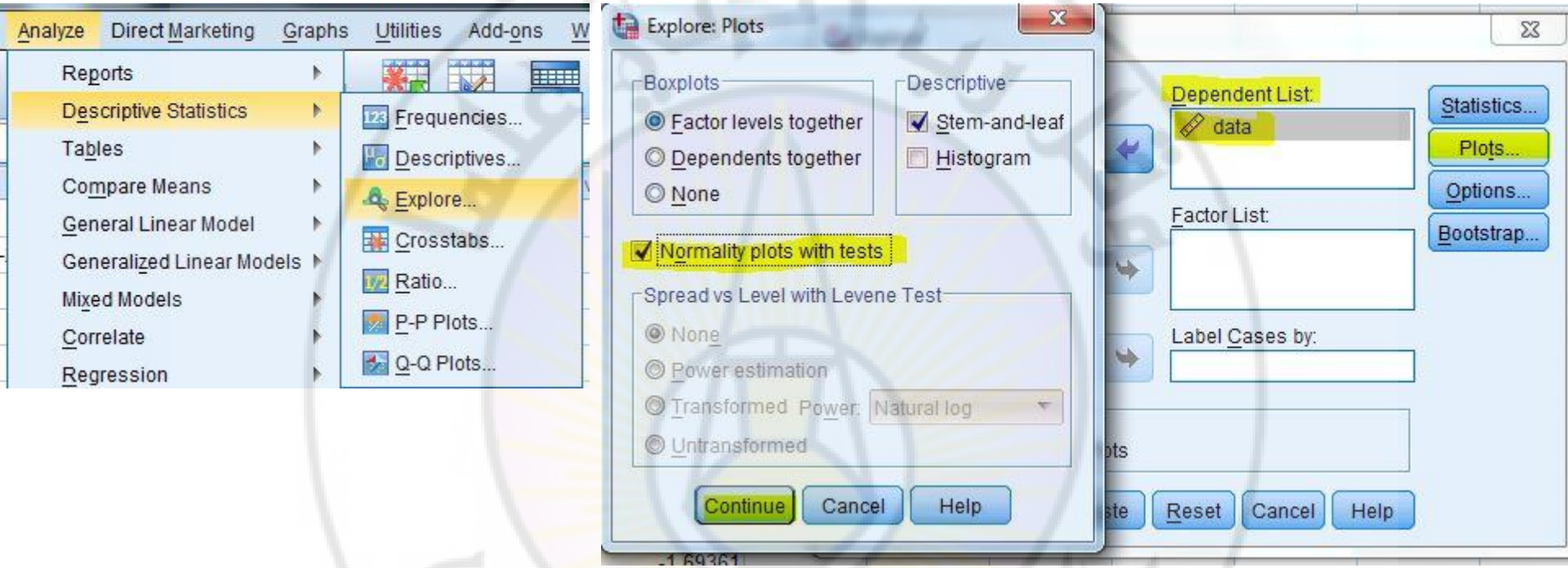
Nikolai Smirnov 1900-1966



Samuel Shapiro 1930-
Florida International University

Martin Wilk 1915-2013

اختبار التوزع الطبيعي في SPSS



Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
data	.006	10000	.200*

$$sig = .2 > \alpha = .05$$

إذاً نقبل الفرض الصفري:

$H_0 : \text{data} \sim N\left(\mu = .007, \sigma^2 = .983\right)$ ونرفض الفرض البديل.

الإحصاء الحيوي

لطلاب السنة الثانية والثالثة - كلية العلوم الصحية

جامعة دمشق

”الجلسة التاسعة“

مدرس المقرر: أ. سلام الجرّاح

إعداد: د. ياسر الزعيم

المدرس في قسم الإحصاء الرياضي

كلية العلوم - جامعة دمشق

حزيران– 2023

نتكلم اليوم عن:

- مخطط الصندوق لمتحول
- تذكرة بحساب معامل بيرسون في SPSS واختبار معنويته
- مصفوفة الارتباط الخطي (لأكثر من متحولين)
- تتمة في تحليل الانحدار الخطي
- ما بين كولموغوروف-سميرنوف و شابيرو-ويلك

رسم الصندوق Box Plot

بفرض $x_1, x_2, \ldots, x_N$ عينة من المتحول X ونريد رسم الصندوق لهذا المتحول. يتم رسم الصندوق بعد حساب المقادير التالية
Q_1 , Q_2 , Q_3 , IQR و min, max خاصة بالرسم هذا.

Q_1: الربيعي الأول first quartile و هو المئوي الـ 25 (25^{th} percentile) وهو القيمة التي على يسارها 25% من المشاهدات و على يمينها 75%.

Q_2 : الربيعي الثاني second quartile وهو المئوي الـ 50 (50^{th} percentile) وهو القيمة التي على يسارها 50% من المشاهدات و على يمينها 50% أي إن Q_2 هو الوسط (لذلك إن Q_1هو وسط ما تحت الوسط)

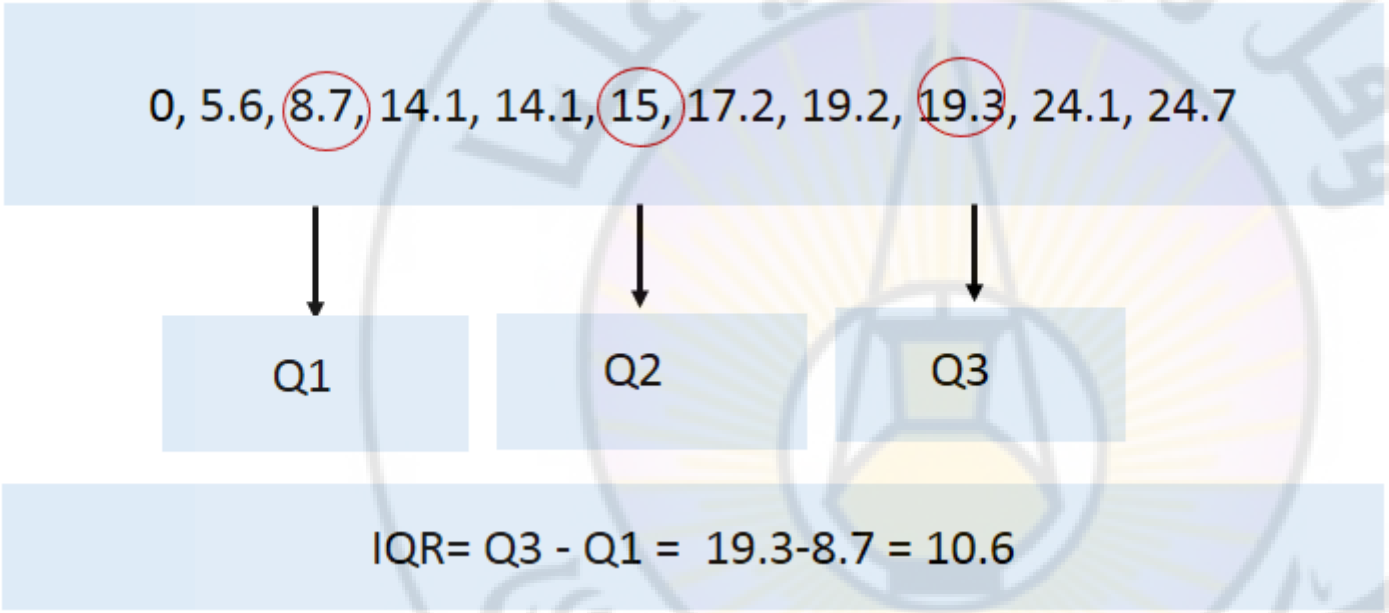
Q_3 : الربيعي الثالث third quartile وهو المئوي الـ 75(75^{th} percentile) وهو القيمة التي على يسارها 75% من المشاهدات و على يمينها 25% أي إن Q_3 هو وسط ما فوق الوسط.

رسم الصندوق Box Plot

IQR هو المدى الربيعي interquartile range وهو يحسب كالتالي

$$IQR = Q_3 - Q_1$$

$$N = 11$$



القيم الشاذة (المنعزلة) و القيم القاصية: إنّ المشاهدات المنعزلة outliers هي المقادير الواقعة ضمن المجال $I_1 = [\mathbf{Q_1 - 3(IQR)}, \mathbf{Q_1 - 1.5(IQR)}]$ أو ضمن المجال $I_2 = [\mathbf{Q_3 + 1.5(IQR)}, \mathbf{Q_3 + 3(IQR)}]$.

ولذلك كل مشاهدة تزيد عن القيمة $\mathbf{Q_3 + 3(IQR)}$ هي مشاهدة قاصية **extreme** و كل مشاهدة تقل عن $\mathbf{Q_1 - 3(IQR)}$ هي قاصية.

رسم الصندوق Box Plot

أمّا الـ min, max الخاصة بالرسم فهي محسوبة من المشاهدات المحصورة بين و خارج المجالين $I_1 = [Q_1 - 3(IQR), Q_1 - 1.5(IQR)]$ و $I_2 = [Q_3 + 1.5(IQR), Q_3 + 3(IQR)]$، أي محسوبة من المشاهدات التالية

$$Q_1 - 1.5(IQR) < x_i < Q_3 + 1.5(IQR); \quad i = 1, 2, \ldots, N$$

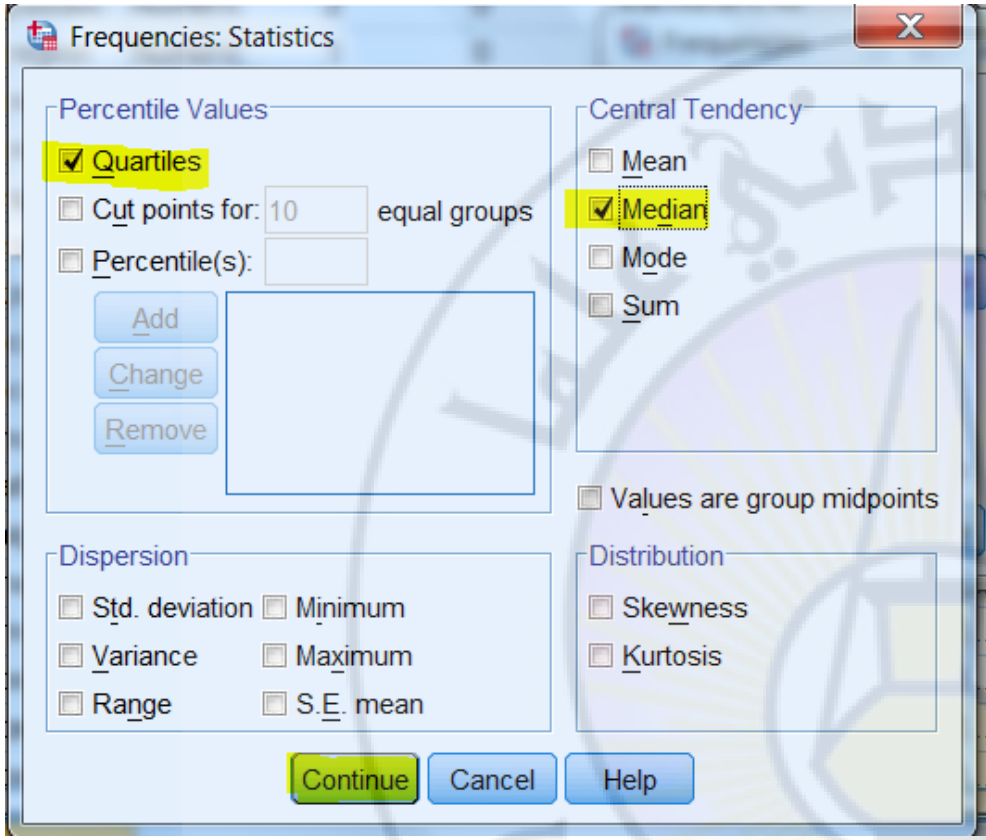
0, 5.6, 8.7, 14.1, 14.1, 15, 17.2, 19.2, 19.3, 24.1, 24.7

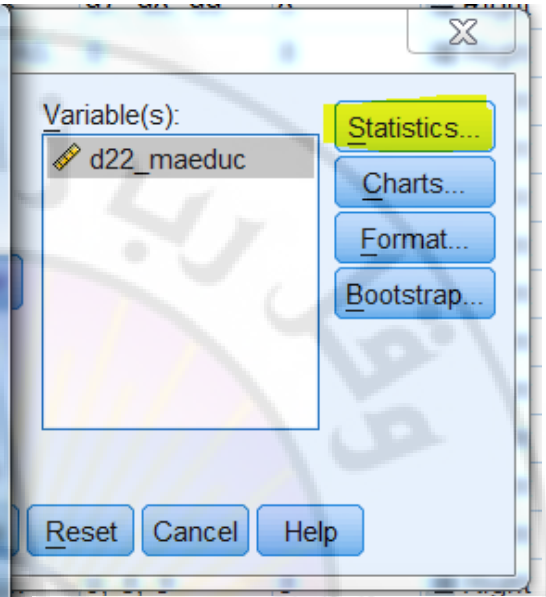
Q1 Q2 Q3

IQR= Q3 - Q1 = 19.3-8.7 = 10.6

لا توجد نقط منعزلة و لا قاصية في هذا المثال كون المجالات هي كالتالي

$$I_1 = [-23.1, -7.2] \quad , \quad I_2 = [35.2, 51.1]$$





Statistics

MOTHER'S HIGHEST YEAR SCHOOL COMPLETED

N	Valid	2288
	Missing	250
Median		12.00
Percentiles	25	10.00
	50	12.00
	75	14.00

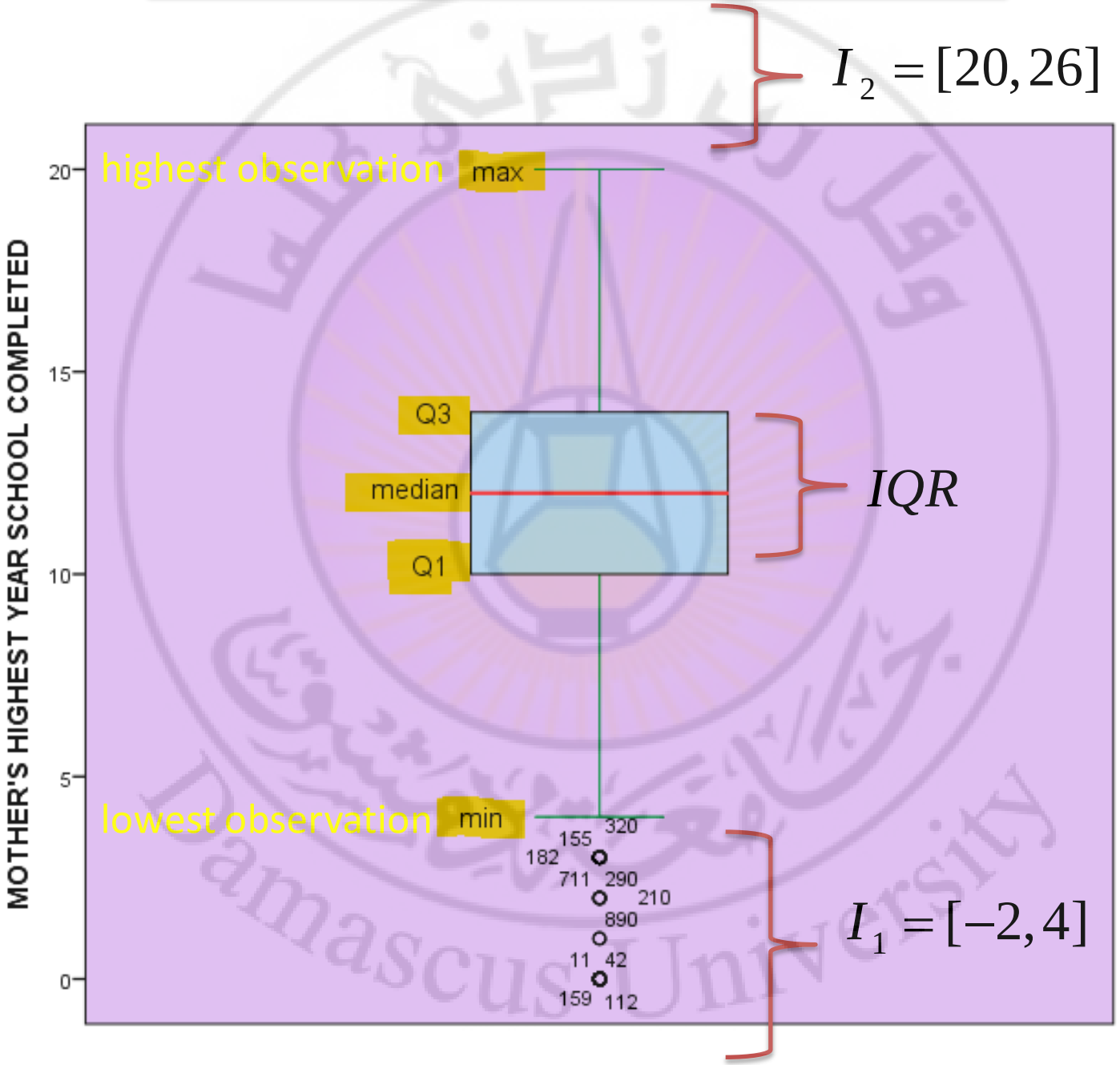
$$IQR = 14 - 10 = 4$$

$$I_1 = [Q_1 - 3(IQR), Q_1 - 1.5(IQR)]$$

$$= [10 - 3*4, 10 - 1.5*4] = [-2, 4]$$

$$I_2 = [Q_3 + 1.5(IQR), Q_3 + 3(IQR)] = [14 + 1.5*4, 14 + 3*4]$$

$$= [20, 26]$$


$I_2 = [20, 26]$
highest observation
max
Q3
median
Q1
IQR
lowest observation
min
$I_1 = [-2, 4]$
MOTHER'S HIGHEST YEAR SCHOOL COMPLETED
20
15
10
5
0
320
155
182
711
290
210
890
11
42
159
112

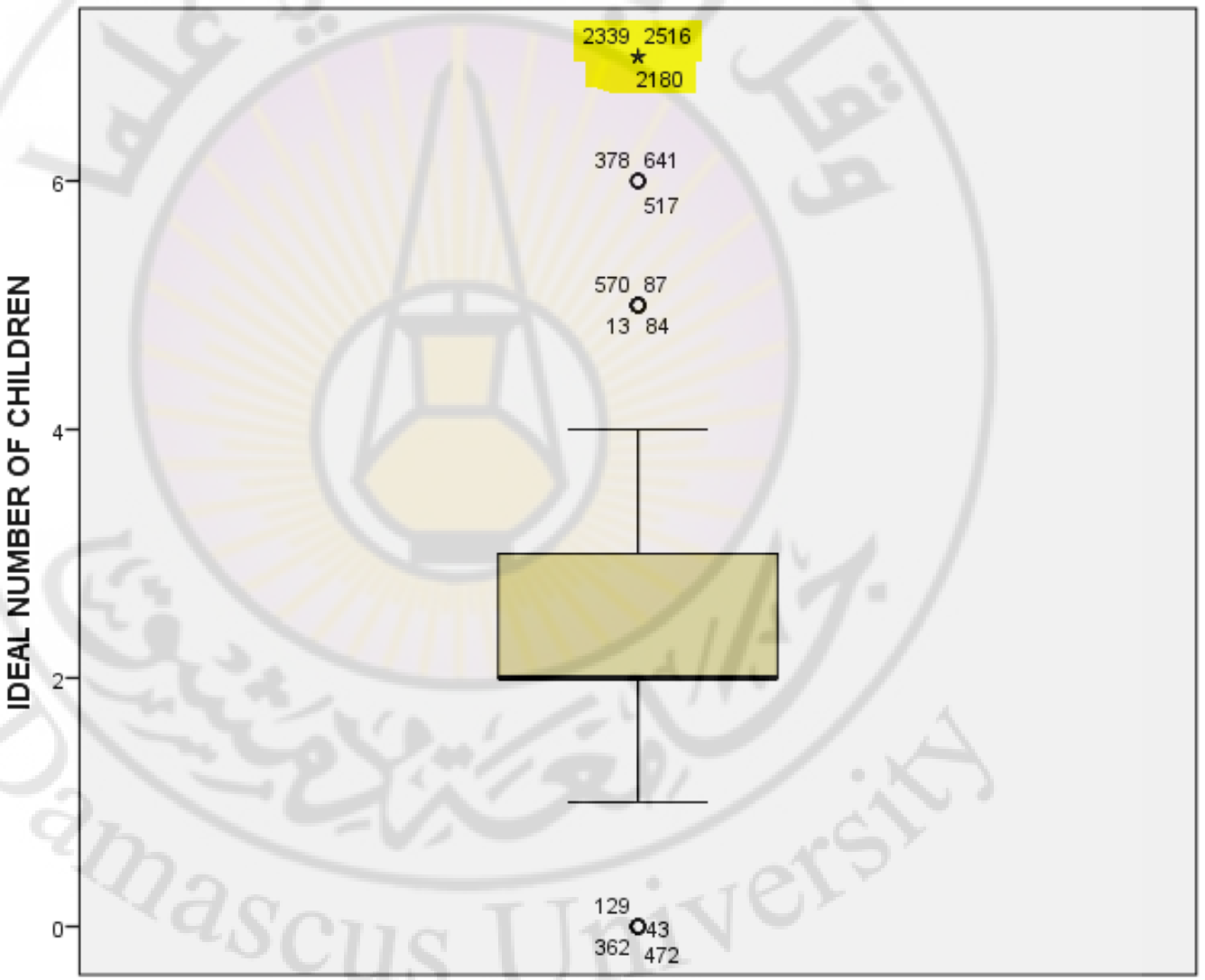
رسم الصندوق Box Plot

d13_chldidel

extreme values in cases: 2339, 2516, 2180

	d13_chldidel
2339	SEVEN+
2340	3

	d13_chldidel
2516	SEVEN+
2517	3



تذكرة بما تحدثنا عنه

متغيرين X , Y مستمرين، نقيس شدة العلاقة الخطية بينهما من خلال معامل الارتباط الخطي بيرسون Pearson correlation coefficient.

$$X_1, X_2, \ldots, X_{Total}$$
$$Y_1, Y_2, \ldots, Y_{Total}$$
$$\Rightarrow \quad \rho_{XY} = \frac{\sum_{i=1}^{Total} (X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{Total} (X_i - \mu_X)^2} \sqrt{\sum_{i=1}^{N} (Y_i - \mu_Y)^2}}$$

$$x_1, x_2, \ldots, x_N$$
$$y_1, y_2, \ldots, y_N$$
$$\Rightarrow \quad \hat{\rho}_{XY} = r_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}}$$

أن تكون البيانات ازدواجية related pairs و خلو كل من المتحولين من المشاهدات القاصية (أو أن يتوزع كل من المتحولين طبيعياً)

البحث السادس: تحليل الارتباط الخطي

Correlations

		FATHER'S HIGHEST YEAR SCHOOL COMPLETED	MOTHER'S HIGHEST YEAR SCHOOL COMPLETED
FATHER'S HIGHEST YEAR SCHOOL COMPLETED	Pearson Correlation	1	.706**
	Sig. (2-tailed)		.000
	N	1903	1796
MOTHER'S HIGHEST YEAR SCHOOL COMPLETED	Pearson Correlation	.706**	1
	Sig. (2-tailed)	.000	
	N	1796	2288

**. Correlation is significant at the 0.01 level (2-tailed).

$$H_0 : \rho_{XY} = 0 \ \ v.s. \ \ H_1 : \rho_{XY} \neq 0$$

$$sig = .000 < \alpha = .05$$

لذا نرفض الفرض الصفري و نقبل الفرض البديل أي أنه توجد علاقة خطية حقيقية (معنوية، ذات دلالة) موجبة (طردية) قوية بين مجتمعي هذين المتحولين عند مستوى الأهمية ألفا.

اصطلاح: بعض الكتب ترمز لمعامل الارتباط الخطي بيرسون بهذه الرموز

$$Cor(x,y) \ , \ Corr(x,y) \ , \ r(x,y) \ , \ r_{xy}$$

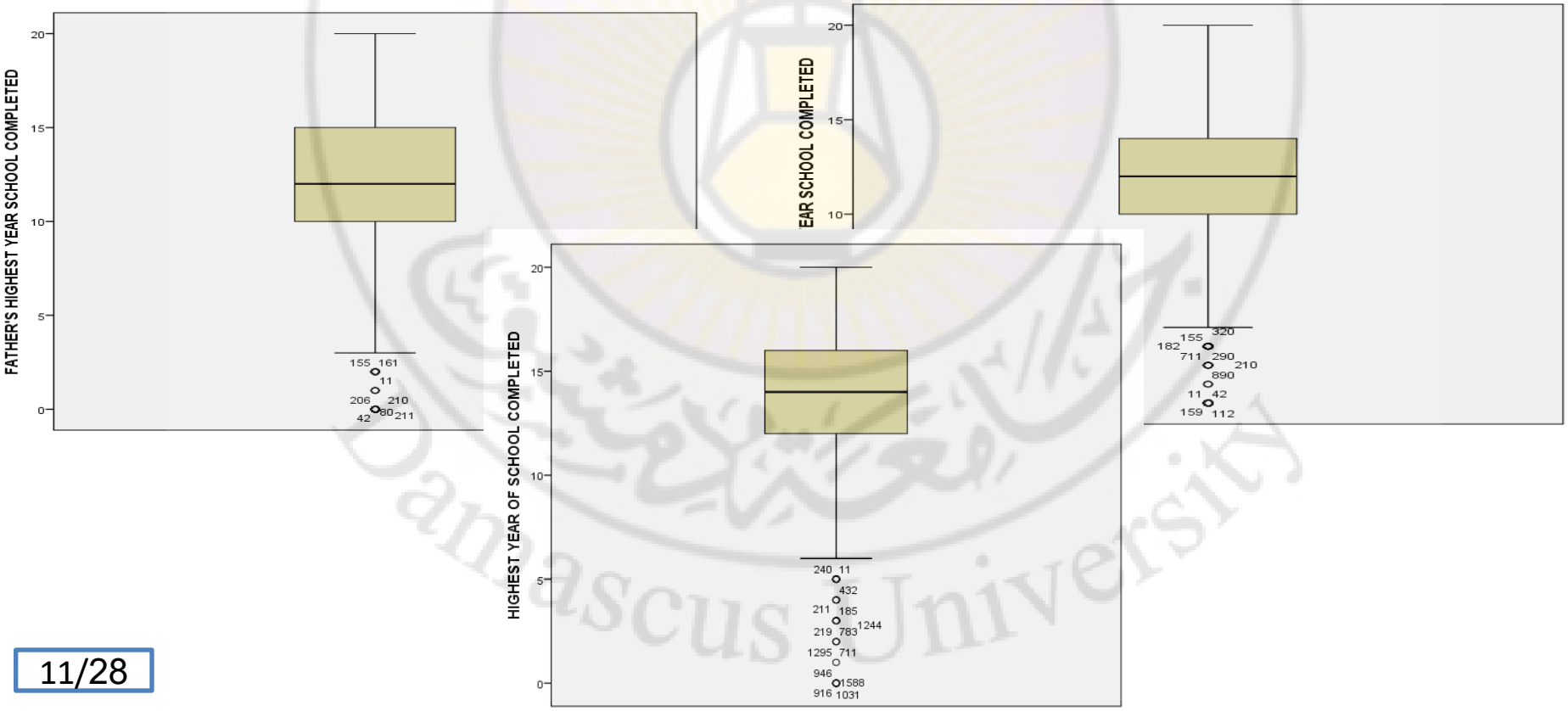
البحث السادس: تحليل الارتباط الخطي

مصفوفة الارتباط الخطي (لأكثر من متحولين) correlation matrix

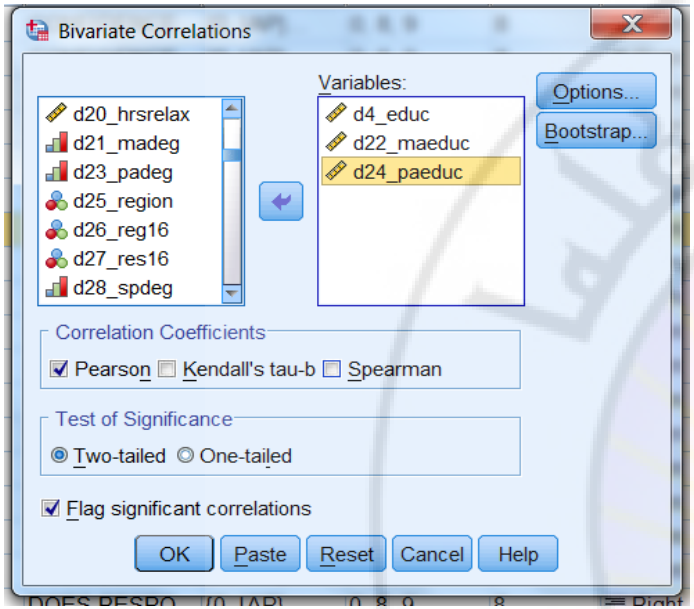
z=d4_educ HIGHEST YEAR OF SCHOOL COMPLETED

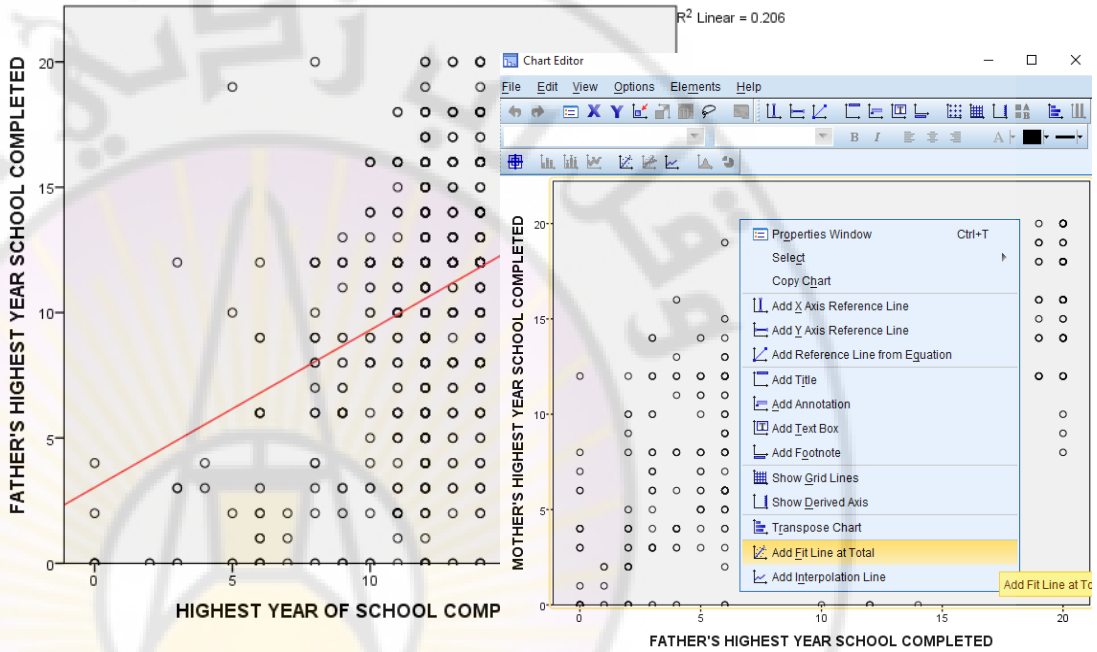
x=d22_maeduc MOTHER'S HIGHEST YEAR SCHOOL COMPLETED

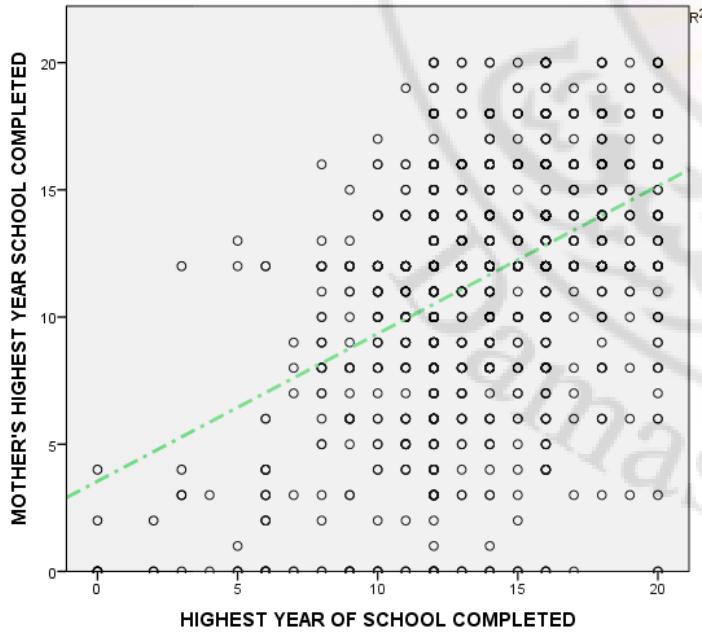
y=d24_paeduc FATHER'S HIGHEST YEAR SCHOOL COMPLETED

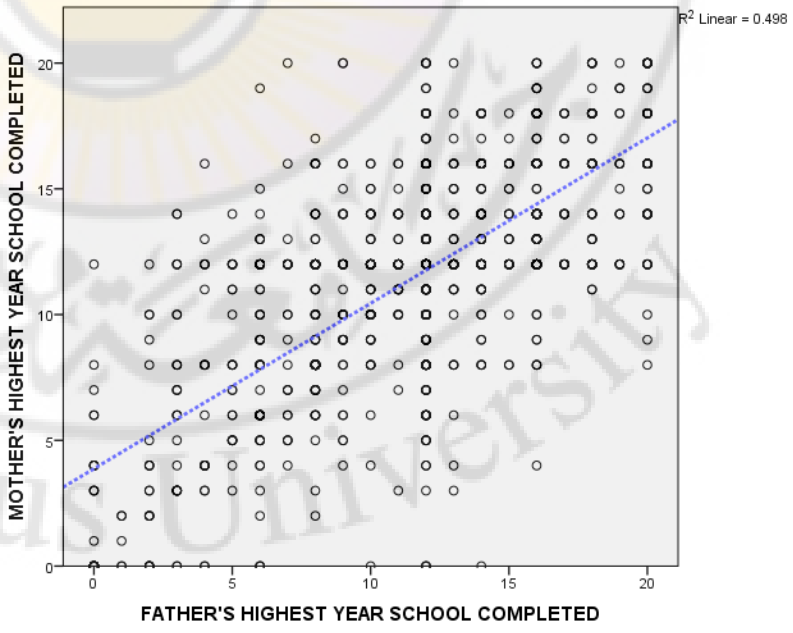


مصفوفة الارتباط الخطي (لأكثر من متحولين) correlation matrix









Correlations

		HIGHEST YEAR OF SCHOOL COMPLETED	MOTHER'S HIGHEST YEAR SCHOOL COMPLETED	FATHER'S HIGHEST YEAR SCHOOL COMPLETED
HIGHEST YEAR OF SCHOOL COMPLETED	Pearson Correlation	1	.438**	.454**
	Sig. (2-tailed)		.000	.000
	N	2537	2288	1903
MOTHER'S HIGHEST YEAR SCHOOL COMPLETED	Pearson Correlation	.438**	1	.706**
	Sig. (2-tailed)	.000		.000
	N	2288	2288	1796
FATHER'S HIGHEST YEAR SCHOOL COMPLETED	Pearson Correlation	.454**	.706**	1
	Sig. (2-tailed)	.000	.000	
	N	1903	1796	1903

**. Correlation is significant at the 0.01 level (2-tailed).

z=d4_educ

x=d22_maeduc

y=d24_paeduc

$$r_{zx} = .438 \quad , \quad sig_{zx} = .000 < \alpha = .05$$

$$H_0 : \rho_{ZX} = 0 \;\; v.s. \;\; H_1 : \rho_{ZX} \neq 0$$

$$r_{zy} = .454 \quad , \quad sig_{zy} = .000 < \alpha = .05$$

$$H_0 : \rho_{ZY} = 0 \;\; v.s. \;\; H_1 : \rho_{ZY} \neq 0$$

$$r_{xy} = .706 \quad , \quad sig_{XY} = .000 < \alpha = .05$$

$$H_0 : \rho_{XY} = 0 \;\; v.s. \;\; H_1 : \rho_{XY} \neq 0$$

		HIGHEST YEAR OF SCHOOL COMPLETED	MOTHER'S HIGHEST YEAR SCHOOL COMPLETED	FATHER'S HIGHEST YEAR SCHOOL COMPLETED
HIGHEST YEAR OF SCHOOL COMPLETED	Pearson Correlation	1	.438**	.454**
	Sig. (2-tailed)		.000	.000
	N	2537	2288	1903
MOTHER'S HIGHEST YEAR SCHOOL COMPLETED	Pearson Correlation	.438**	1	.706**
	Sig. (2-tailed)	.000		.000
	N	2288	2288	1796
FATHER'S HIGHEST YEAR SCHOOL COMPLETED	Pearson Correlation	.454**	.706**	1
	Sig. (2-tailed)	.000	.000	
	N	1903	1796	1903

**. Correlation is significant at the 0.01 level (2-tailed).

$$R_{zxy} = \begin{pmatrix} 1 & .438 & .454 \\ .438 & 1 & .706 \\ .454 & .706 & 1 \end{pmatrix}$$

first column ↓ z

العمود الثالث ↓ y

x

$$R_{zxy} = \begin{pmatrix} 1 & .438 & .454 \\ .438 & 1 & .706 \\ .454 & .706 & 1 \end{pmatrix}$$

z ← first row

x ←

y ← السطر الثالث

مصفوفة الارتباط الخطي

تمرين:

في الملف question.sav لدينا المتحولات الثلاث x , y , u

و المطلوب:

1) ارسم المدرج و عليه المضلع التكراري للمتحولات (سترجس)
2) ارسم مخطط الانتشار لكل متحولين ثم ارسم عليه الاتجاه المتوقع للارتباط الخطي بينهما.
3) تأكد من خلو المتحولات من القيم القاصية extremes ثم أوجد مصفوفة الارتباط للمتحولات. فسر (كتابياً) النتائج و اختبر الفرضيات الإحصائية المصاحبة عند مستوى الأهمية (الدلالة) $\alpha = .05$.

إنّ نموذج الانحدار الخطي البسيط simple linear regression model يتناول بناء المعادلة:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

بهدف التنبؤ بقيم مستقبلية للمتحول Y من خلال معرفة قيم المتحول X فقط.

مثال: لنفرض أن العلاقة (معادلة الانحدار الخطي البسيط) بين حجم المبيعات اليومية (بالكيلوغرام) لنوع من الشيبس و السعر (بالدينار) هي $\widehat{y} = -1 + 2.3x$

عندئذ بفرض $x = 30$ فإننا نتوقع حجم المبيعات اليومي هو

$$\widehat{y} = -1 + 2.3x = -1 + 2.3(30) = 68$$

بعض الشروط الواجب تحققها بعد بناء نموذج الانحدار الخطي البسيط:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad ; \quad i = 1, 2, \ldots, N$$

الخطأ العشوائي ε في معادلة الانحدار هو مجهول و يتم تقديره من خلال حساب الرواسب residuals أي أن عند كل مشاهدة x_i إنّ:

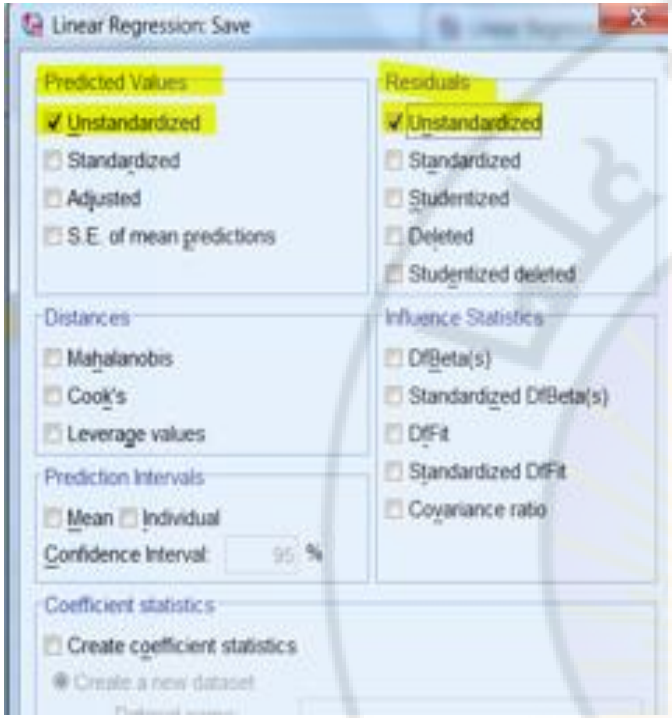
$$\hat{\varepsilon}_i = e_i = y_i - \hat{y}_i$$

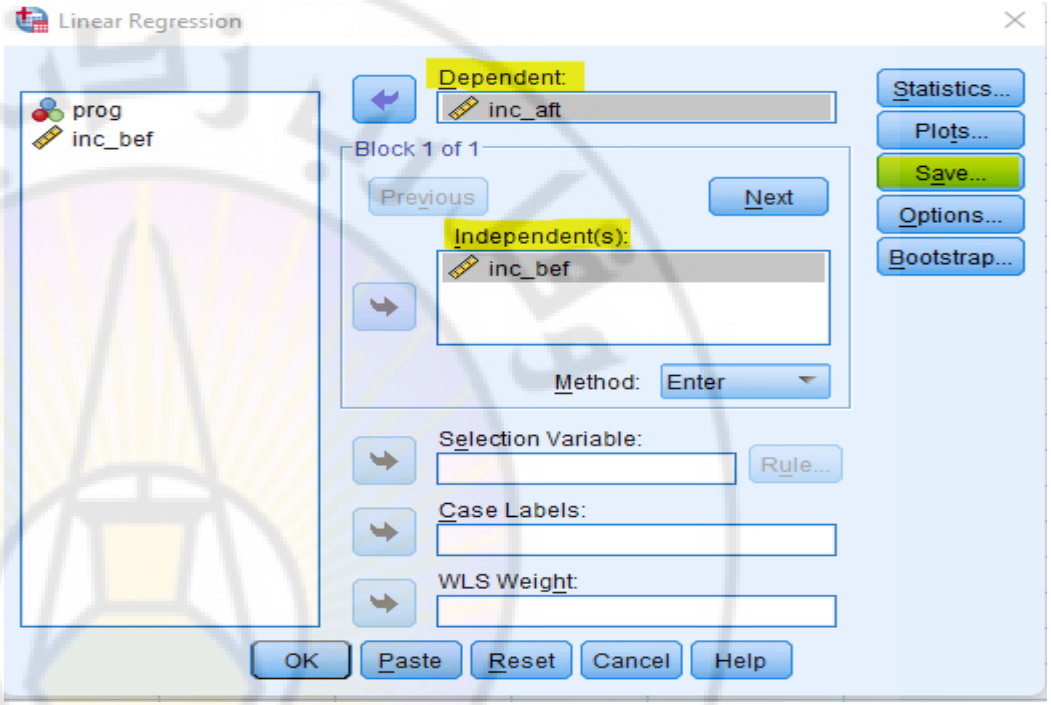
SPSS يقوم بحساب البواقي والقيم المتوقعة و يحفظها في متحولين جديدين و عندئذ يجب التحقق من الشرطين التاليين:

1) توزع البواقي (الرواسب) توزعاً طبيعياً أي $\varepsilon \sim N(\mu, \sigma^2)$

2) استقلال البواقي عن القيم المتوقعة استقلالاً خطياً أي $\rho_{\varepsilon \hat{Y}} = 0$

مثال: training.sav





	inc_aft	PRE_1
1	12.00	14.99000
2	10.00	14.99000
3	11.00	14.99000
4	18.00	16.67029

PRE_1 = $\widehat{inc\_after} = \hat{y} = 1.548 + 1.680x$

Predicted

	inc_aft	PRE_1	RES_1
1	12.00	14.99000	-2.99000
2	10.00	14.99000	-4.99000
3	11.00	14.99000	-3.99000
4	18.00	16.67029	1.32971
5	12.00	13.30970	-1.30970
6	15.00	14.99000	.01000
7	13.00	14.99000	-1.99000
8	22.00	16.67029	5.32971

residuals

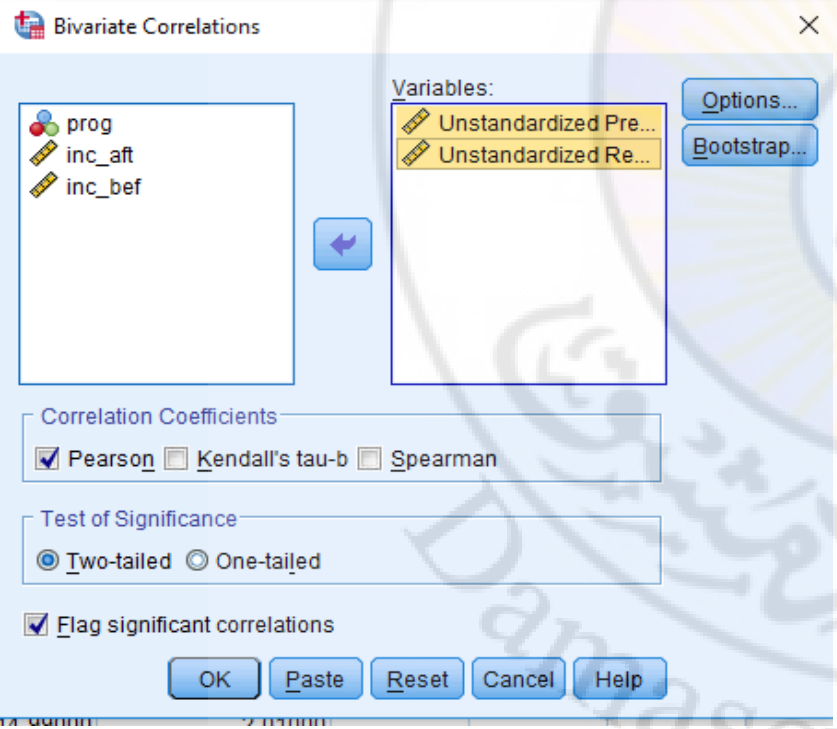
البحث السادس: تحليل الارتباط الخطي

التوزع الطبيعي للرواسب

إنّ رواسب النموذج لا تتوزع طبيعياً وهذا ليس جيداً.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	.051	1000	.000	.987	1000	.000



Correlations

		Unstandardized Predicted Value	Unstandardized Residual
Unstandardized Predicted Value	Pearson Correlation	1	.000
	Sig. (2-tailed)		1.000
	N	1000	1000
Unstandardized Residual	Pearson Correlation	.000	1
	Sig. (2-tailed)	1.000	
	N	1000	1000

الأخطاء العشوائية مستقلة خطياً عن القيم المتوقعة و هذا جيد. لاحظ الاختبار الإحصائي و قيمة معامل بيرسون.

$$r_{e\hat{y}} = .000 \quad sig_{\varepsilon\hat{Y}} = 1 > \alpha = .05$$

$$H_0: \rho_{\varepsilon\hat{Y}} = 0 \quad v.s. \quad H_1: \rho_{\varepsilon\hat{Y}} \neq 0$$

البحث السادس: تحليل الانحدار الخطي

تمرين: PEARSON_LinReg.sav

y=Son x=Father

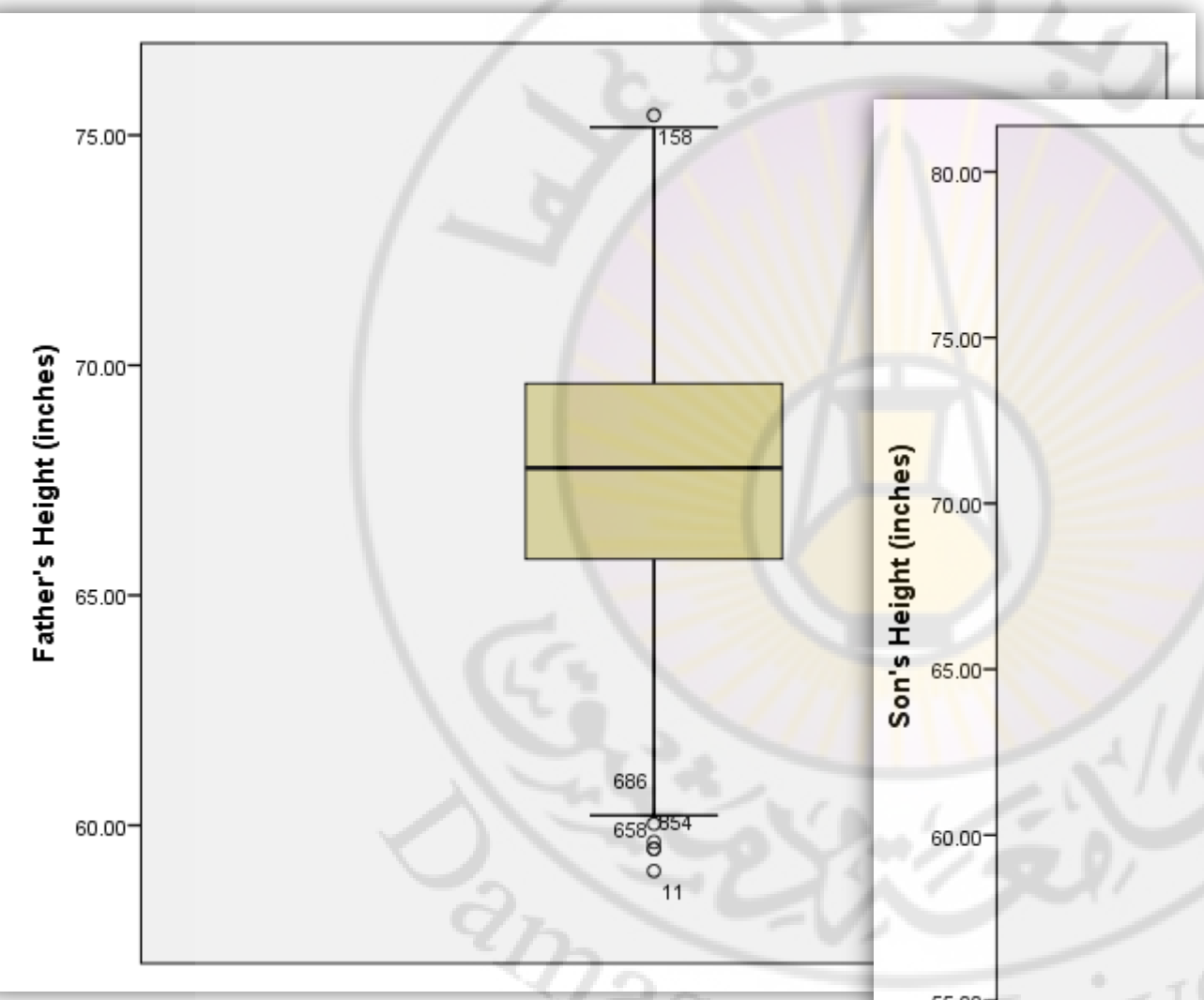
طول قامة الابن بالإنش طول قامة الأب بالإنش

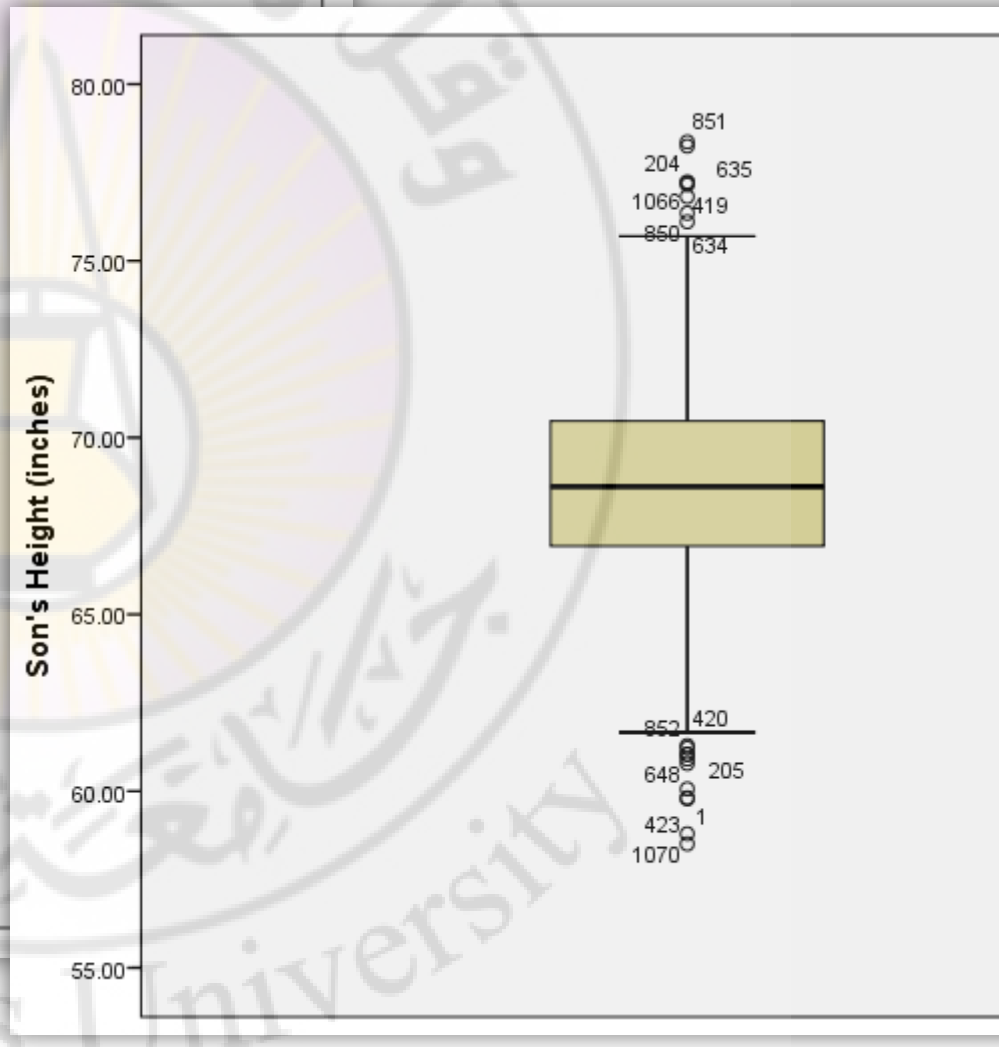
قم ببناء نموذج انحدار خطي بسيط للتنبؤ بطول قامة الأبناء من خلال معرفة طول قامة الآباء مع الانتباه للخطوات المطلوب إنجازها

(رسم الصندوق- الارتباط الخطي المستقيم الحقيقي بين متحول التنبؤ و المتحول التابع- التوزع الطبيعي للرواسب- الاستقلال الخطي للرواسب عن القيم المتوقعة)

إذا كان طول أب هو x=80 كم تتوقع طول ابنه بالإنش؟

البحث السادس: تحليل الانحدار الخطي

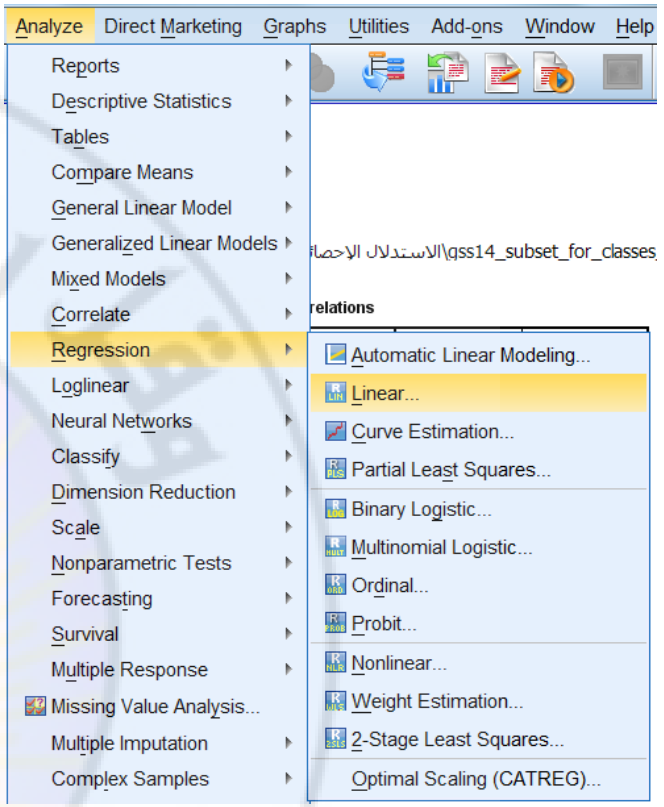


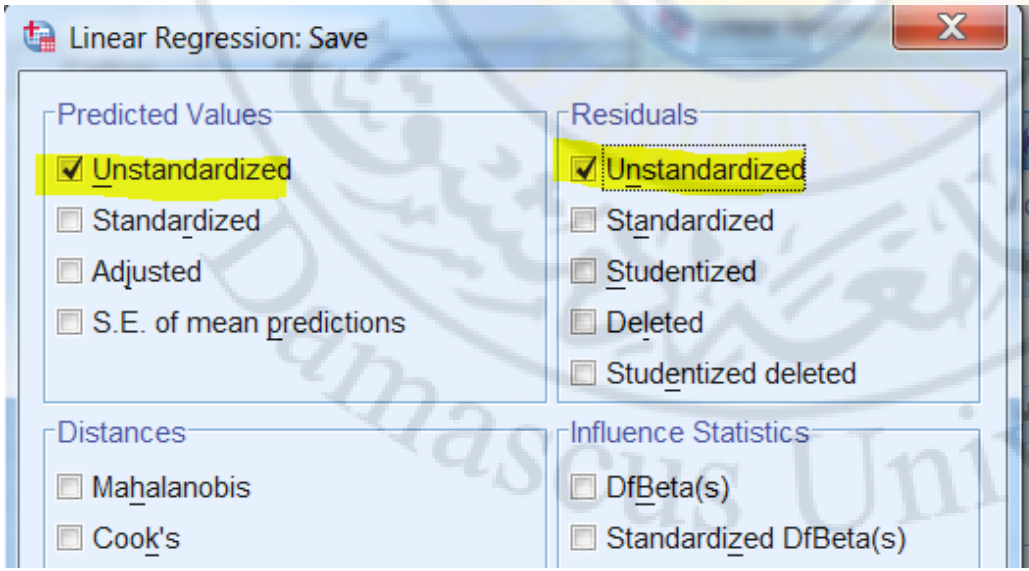


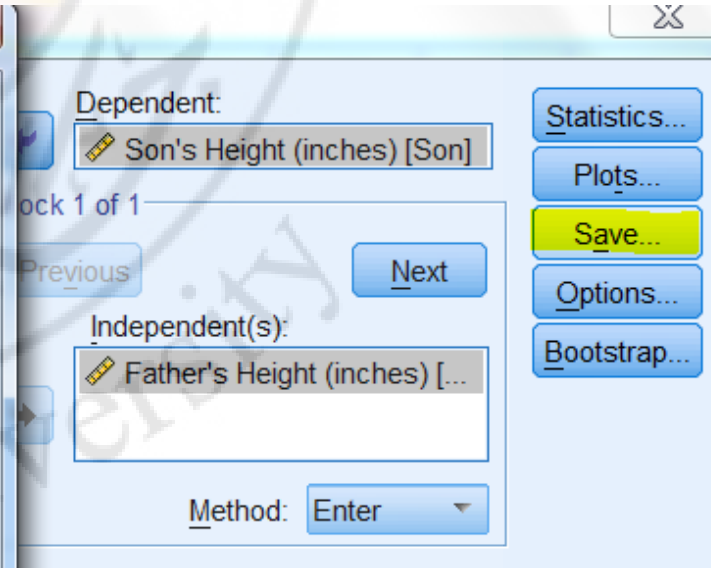
Correlations

		Son's Height (inches)	Father's Height (inches)
Son's Height (inches)	Pearson Correlation	1	.501**
	Sig. (2-tailed)		.000
	N	1078	1078
Father's Height (inches)	Pearson Correlation	.501**	1
	Sig. (2-tailed)	.000	
	N	1078	1078

**. Correlation is significant at the 0.01 level (2-tailed).







البحث السادس: تحليل الانحدار الخطي

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.501[a]	.251	.251	2.43656

a. Predictors: (Constant), Father's Height (inches)

b. Dependent Variable: Son's Height (inches)

$$R = r_{xy} = .501$$

$$R^2 = (R)^2 = .251$$

25.1% من التغير في المتحول التابع مسؤول عنه المتحول المستقل

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2144.580	1	2144.580	361.235	.000[b]
	Residual	6388.001	1076	5.937		
	Total	8532.581	1077			

a. Dependent Variable: Son's Height (inches)

b. Predictors: (Constant), Father's Height (inches)

SSR=2144.580

SSE=6388.001

SST=8532.581

sig=.000$<\alpha = .05$

Coefficients[a]

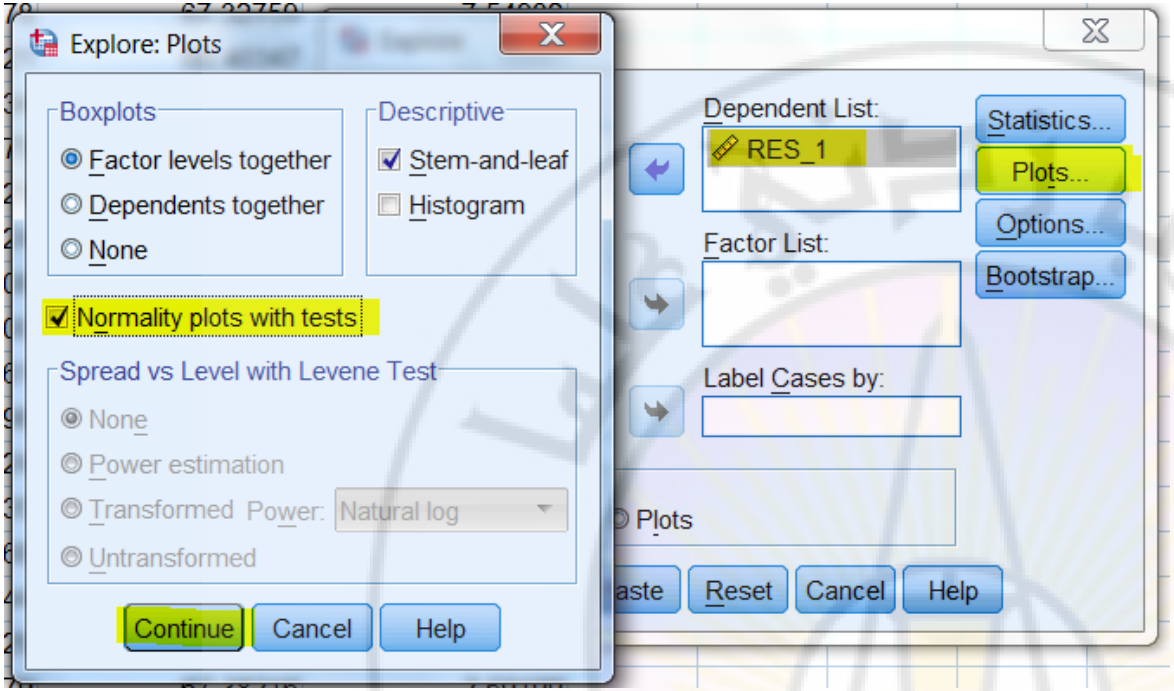
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	33.887	1.832		18.493	.000
	Father's Height (inches)	.514	.027	.501	19.006	.000

a. Dependent Variable: Son's Height (inches)

$$\hat{y} = 33.887 + .514x$$

$H_0 : \beta_1 = 0 \; v.s. \; H_1 : \beta_1 \neq 0$

$sig = .000 < \alpha = .05$



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	.022	1078	.200*	.993	1078	.000

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

$$H_0 : \varepsilon \sim N(\mu, \sigma^2) \quad v.s. \quad H_1 : \varepsilon - N(\mu, \sigma^2)$$

$sig_{KS} = .200 > \alpha = .05$ ← نقبل الفرض الصفري

$sig_{SW} = .000 < \alpha = .05$ ← نرفض الفرض الصفري

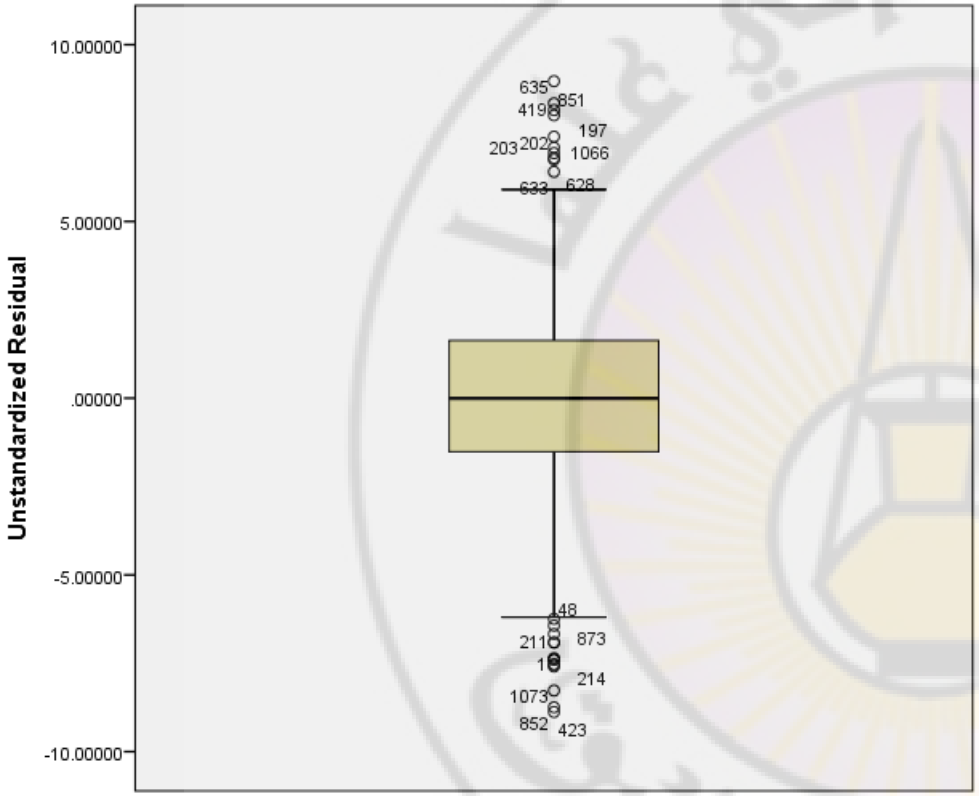
ما بين كولموغوروف-سميرنوف و شابيرو-ويلك

بشكل عام إن اختبار شابيرو-ويلك أقوى من اختبار كولموغوروف-سميرنوف.
غالباً كلا الاختبارين يقودان إلى نفس القرار أي القرار برفض (أو قبول) الفرض الصفري، و لكن إذا حدث تناقض في القرار (كما هو الحال في مثالنا) فعندئذ يجب الانتباه إلى هاتين النقطتين:

1) وجود نقط منعزلة outliers أو قاصية extremes: إن اختبار كولموغوروف-سميرنوف حساس جداً لوجود النقط المنعزلة و حساس أيضاً لوجود النقط القاصية. فإذا وجدت نقط منعزلة فالأفضل الابتعاد عن قراءة نتائج هذا الاختبار و اعتماد اختبار شابيرو-ويلك إلا إذا كانت:

2) توجد الربطات ties: الربطة تعني تكرار مشاهدة عدة مرات.
حيث أن اختبار شابيرو-ويلك يتأثر بوجود الربطات.

ما بين كولموغوروف-سميرنوف و شابيرو-ويلك



توجد نقط منعزلة outliers في رسم البواقي (الرواسب) لذلك لا نعتمد نتائج كولموغوروف-سميرنوف.

هل نعتمد نتائج شابيرو-ويلك؟

Value	
-2.0	
-1.0	
1.0	
2.2	(tie)
2.2	
2.2	
5.3	(tie)
5.3	
7.0	
8.1	
8.2	

في هذا الجدول لدينا ربطتين و هما المشاهدة 2.2 (مكررة ثلاث مرات)، و المشاهدة 5.3 (مكررة مرتان).

لاحظ أنه إذا كانت كل المشاهدات لها نفس التكرار فليس لدينا ربطات.

Unstanda

		Frequency
Valid	-8.87715	1
	-8.74328	1
	-8.27138	1
	-7.59304	1
	-7.54932	1
	-7.52333	1
	-7.40243	1
	-7.39217	1
	-7.35718	1
	-6.91229	1
	-6.91015	1
	-6.67113	1
	-6.43384	1
	-6.24586	1
	-6.19789	1
	-5.96361	1
	-5.73198	1
	-5.67309	1
	-5.63776	1
	-5.52083	1
	-5.47012	1

لنرجع لمتحول البواقي في مثالنا:

نلاحظ أن كل المشاهدات مكررة نفس التكرار (وهو تكرار واحد فقط لكل مشاهدة) بالتالي ليس هناك ربطات و إن اختبار شابيرو-ويلك هو ما سنعتمده في هذا المثال.

$$sig_{SW} = .000 < \alpha = .05$$

إذاً إن البواقي (و بالتالي إن الأخطاء العشوائية) لا تتوزع طبيعياً.

البحث السادس: تحليل الانحدار الخطي

Correlations

		Unstandardized Predicted Value	Unstandardized Residual
Unstandardized Predicted Value	Pearson Correlation	1	.000
	Sig. (2-tailed)		1.000
	N	1078	1078
Unstandardized Residual	Pearson Correlation	.000	1
	Sig. (2-tailed)	1.000	
	N	1078	1078

$$r_{e\hat{y}} = .000 \quad , \quad sig_{\varepsilon\hat{Y}} = 1 > \alpha = .05$$

$$H_0 : \rho_{\varepsilon\hat{Y}} = 0 \;\; v.s. \;\; H_1 : \rho_{\varepsilon\hat{Y}} \neq 0$$

إذا كان طول الأب هو $x = 80$ ، نتوقع أن يكون طول ابنه:

$$\hat{y} = 33.887 + .514x = 33.887 + .514(80) = 75.01$$

الإحصاء الحيوي

لطلاب السنة الثانية والثالثة - كلية العلوم الصحية

جامعة دمشق

" الجلسة الثانية "

مدرس المقرر: أ. سلام الجرّاح

إعداد: د. ياسر الزعيم

المدرس في قسم الإحصاء الرياضي

كلية العلوم - جامعة دمشق

نيسان– 2023

نتكلم اليوم عن:

- تذكرة بما تحدثنا عنه
- بعض الرموز الهامة في الإحصاء
- **البحث الثاني: التوزيع التكراري**
- التوزيعات التكرارية في SPSS

إن المتغير وفق نوع قيمه هو أحد نوعين:

مستمر: أي إن قيم المتغير غير قابلة للعد

(**مثل:** الزمن، الوزن، الطول، درجة الحرارة و ...)

منفصل: أي إن قيم المتغير قابلة للعد (منتهية أوغير منتهية العدد)

(**مثل:** إلقاء حجر نرد، المكالمات الهاتفية الواردة إلى المقسم، عدد السيارات المارة من طريق دولي و ...)

بينما وفق طريقة قياسه فهو اسمي أوترتيبي أوقياسي.

إنّ الاسمي والترتيبي هو منفصل،

بينما القياسي هو مستمر.

بعض الرموز الهامة في الإحصاء

$\sum$ حرف يوناني يدل على عملية الجمع.

عينة حجمها N من المتغير X: $X_1, X_2, \ldots, X_N$

إن $\sum_{i=1}^{N} X_i$ يعني $\sum_{i=1}^{N} X_i = X_1 + X_2 + \cdots + X_N$

وبالتالي إن المتوسط الحسابي بالتعريف هو $\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$

مثال: X هو نتيجة رمي حجر نرد مرة واحدة

عندئذ قيم X: $X_1 = 1, X_2 = 2, \ldots, X_6 = 6$

وأنّ $\bar{X} = \frac{\sum_{i=1}^{6} X_i}{6} = \frac{21}{6} = 3.5$

بعض الرموز الهامة في الإحصاء

مثال: U المتغير الدال على نشاط مجموعة فئران (خمسة فئران) في متاهة. بفرض قيم U هي $10, 12, 19, 21, 32$
عندئذ ما قيمة المتوسط الحسابي $\bar{U}$ ؟

$$U_1 = 10, U_2 = 12, U_3 = 19, U_4 = 21, U_5 = 32$$

$$\bar{U} = \frac{\sum_{i=1}^{5} U_i}{5} = \frac{10+12+19+21+32}{5} = 18.8$$

تمرين:

بفرض c ثابت برهن أنّ:

$$\sum_{i=1}^{N} cX_i = c\sum_{i=1}^{N} X_i$$

$$\sum_{i=1}^{N} (X_i - c)^2 = \sum_{i=1}^{N} X_i{}^2 - 2c\sum_{i=1}^{N} X_i + Nc^2$$

بعض الرموز الهامة في الإحصاء

	NAME	POS	AGE	HT	WT
	Kostas Antetokounmpo	F/C	23	6'10"	200 lbs
	Jordan Bell	F/C	25	6'8"	224 lbs
	Devontae Cacok	PF	24	6'7"	240 lbs
	Kentavious Caldwell-Pope	G	27	6'5"	205 lbs
	Alex Caruso	G	26	6'5"	186 lbs
	Quinn Cook	PG	27	6'1"	179 lbs
	Anthony Davis	F/C	27	6'10"	253 lbs
	Jared Dudley	F	35	6'6"	237 lbs

إن الـ **minimum** هو أصغر مقادير المتغير و يرمز له **min**

أما الـ **maximum** فهو أكبر مقادير المتغير و اختصاراً **max**

مثال:

بالنسبة لمتغير وزن عينة من لاعبي فريق السلة Los Angeles Lakers

إنّ

$$\max_i(X_i) = 253$$

$$\min_i(X_i) = 179$$

البحث 2: التوزيع التكراري

التوزيع التكراري Frequency Distribution

هو تصنيف و إعادة كتابة مشاهدات المتغير وفق تكرار كل مشاهدة في العينة.

حيث أن f_i يدل على تكرار المشاهدة X_i في العينة.

X	f
X_1	f_1
X_2	f_2
⋮	⋮
X_K	f_K

دوماً إن $\sum_{i=1}^{K} f_i = N$

الهدف من الجدول هو تيسير فهم المشاهدات الحاصلة في العينة.

البحث 2: التوزيع التكراري

مثال: إنّ X هو المتغير الدال على شدة الإصابة العضلية لدي فريق كرة القدم الانجليزي ليفربول، حيث أنّ عدد اللاعبين الإجمالي هو 28 لاعب. لنفرض أنّ الإصابات العضلية كانت كالتالي:

$$X_1=1, X_2=1, X_3=1, X_4=2,$$
$$X_5=1, X_6=2, X_7=5, X_8=4,$$
$$X_9=2, X_{10}=3, X_{11}=5, X_{12}=4,$$
$$X_{13}=2, X_{14}=1, X_{15}=3, X_{16}=2,$$
$$X_{17}=1, X_{18}=2, X_{19}=4, X_{20}=5,$$
$$X_{21}=2, X_{22}=4, X_{23}=2, X_{24}=4,$$
$$X_{25}=1, X_{26}=3, X_{27}=4, X_{28}=5$$

شدة الإصابة	label
ضعيفة جداً	1
ضعيفة	2
متوسطة	3
قوية	4
قوية جداً	5

الجدول (1)

البحث 2: التوزيع التكراري

تتمة المثال:

$$X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 2, X_5 = 1, X_6 = 2,$$
$$X_7 = 5, X_8 = 4, X_9 = 2, X_{10} = 3, X_{11} = 5, X_{12} = 4,$$
$$X_{13} = 2, X_{14} = 1, X_{15} = 3, X_{16} = 2, X_{17} = 1,$$
$$X_{18} = 2, X_{19} = 4, X_{20} = 5, X_{21} = 2, X_{22} = 4,$$
$$X_{23} = 2, X_{24} = 4, X_{25} = 1, X_{26} = 3, X_{27} = 4, X_{28} = 5$$

X	f
1	7
2	8
3	3
4	6
5	4

لا حظ أن $\sum_{i=1}^{5} f_i = 28$ وإن حجم العينة هو $N = 28$.

فلذا يجب الانتباه إلى مقادير المتغير وحجم العينة المسحوبة من هذا المتغير.

تمرين: رمينا قطعة معدنية عشر مرات فكانت النتائج كالتالي:

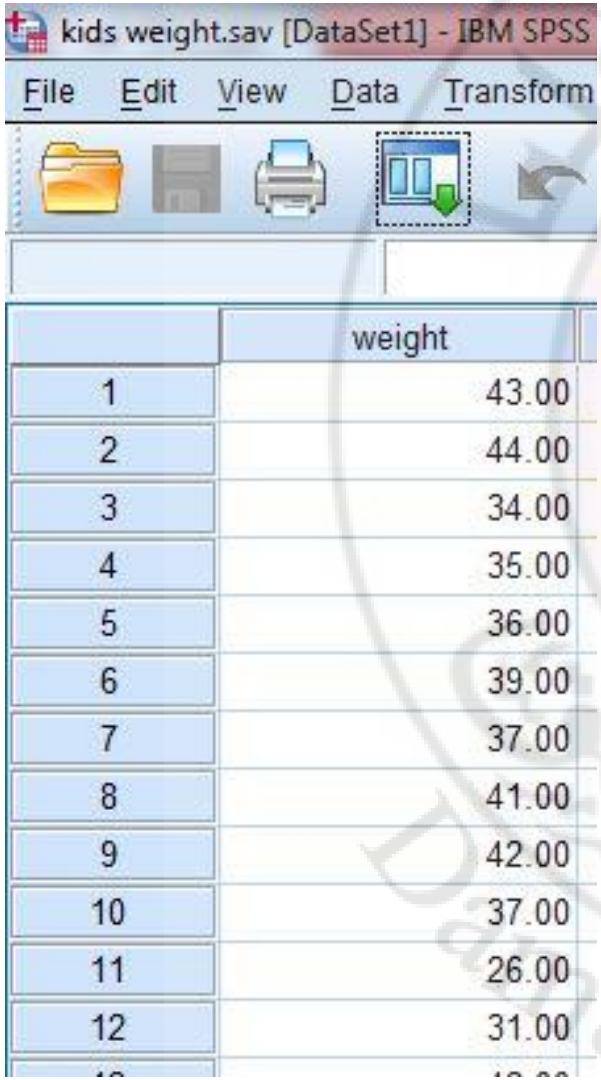
H,H,H,T,H,T,H,T,H,H

اكتب جدول التوزيع التكراري.

في هذين المثال والتمرين إن المتغير المدروس منفصل.

البحث 2: التوزيع التكراري

لنأخذ الآن **مثالاً يكون فيه المتغير المدروس مستمراً.**
Kids weight.sav فيه $N = 209$.



	weight
1	43.00
2	44.00
3	34.00
4	35.00
5	36.00
6	39.00
7	37.00
8	41.00
9	42.00
10	37.00
11	26.00
12	31.00

الأوزان		Frequency
Valid	24.00	1
	26.00	1
	27.00	2
	29.00	6
	30.00	4
	31.00	3
	32.00	4
	33.00	5
	34.00	17
	35.00	8
	36.00	9
	37.00	19
	38.00	8
	39.00	17
	40.00	13
	41.00	9
	42.00	17
	43.00	12
	44.00	14
	45.00	5
	46.00	9
	47.00	7
	48.00	2
	49.00	9
	50.00	3
	51.00	1
	52.00	3
	53.00	1
	Total	209

البحث 2: التوزيع التكراري

التوزيع التكراري المصنّف Grouped frequency distribution

هو تنظيم وترتيب المشاهدات بحسب تكرارها في طبقات (تسمى أيضاً فئات categories) معينة.

تتمة مثال أوزان الأطفال: إن الجدول التكراري لمتغير الأوزان هو

		Frequency
Valid	23.00 - 27.00	4
	28.00 - 32.00	17
	33.00 - 37.00	58
	38.00 - 42.00	64
	43.00 - 47.00	47
	48.00+	19
	Total	209

إن الحدود اليسارية (أو الحدود السفلية) للفئات تسمى بنقاط القطع cut points.
أما عرض الفئة (width) فهو الحد العلوي ناقص السفلي وهو يبلغ 4 في هذا المثال.
الآن لنفرض وزن طفل ما هو 27.5 كيلوغراماً، إلى أية فئة ينتمي وزن هذا الطفل؟

الحدود الفعلية للفئات هي $[22.5, 27.5[$, $[27.5, 32.5[$, ... , $[47.5, \infty[$

البحث 2: التوزيع التكراري

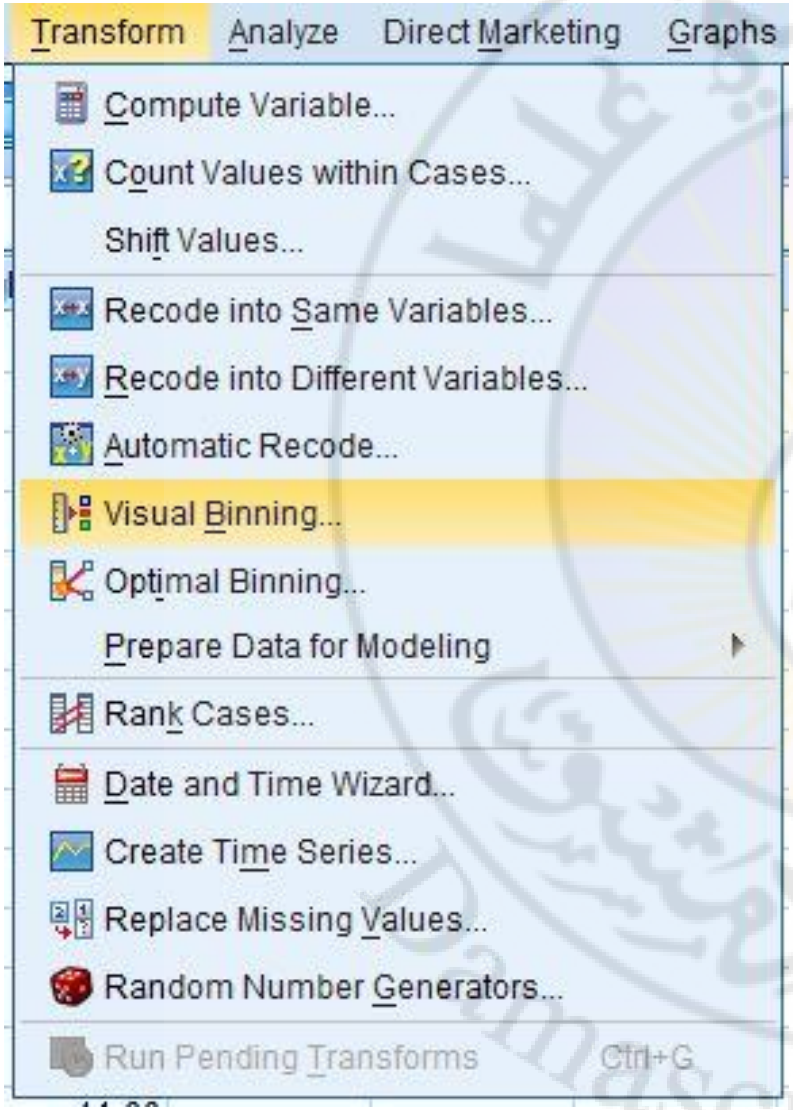
مثال: بفرض X المتغير الدال على أوزان $N = 209$ أطفال بالكيلوغرام.
و لنفرض بأنّ الجدول التكراري كان كالتالي

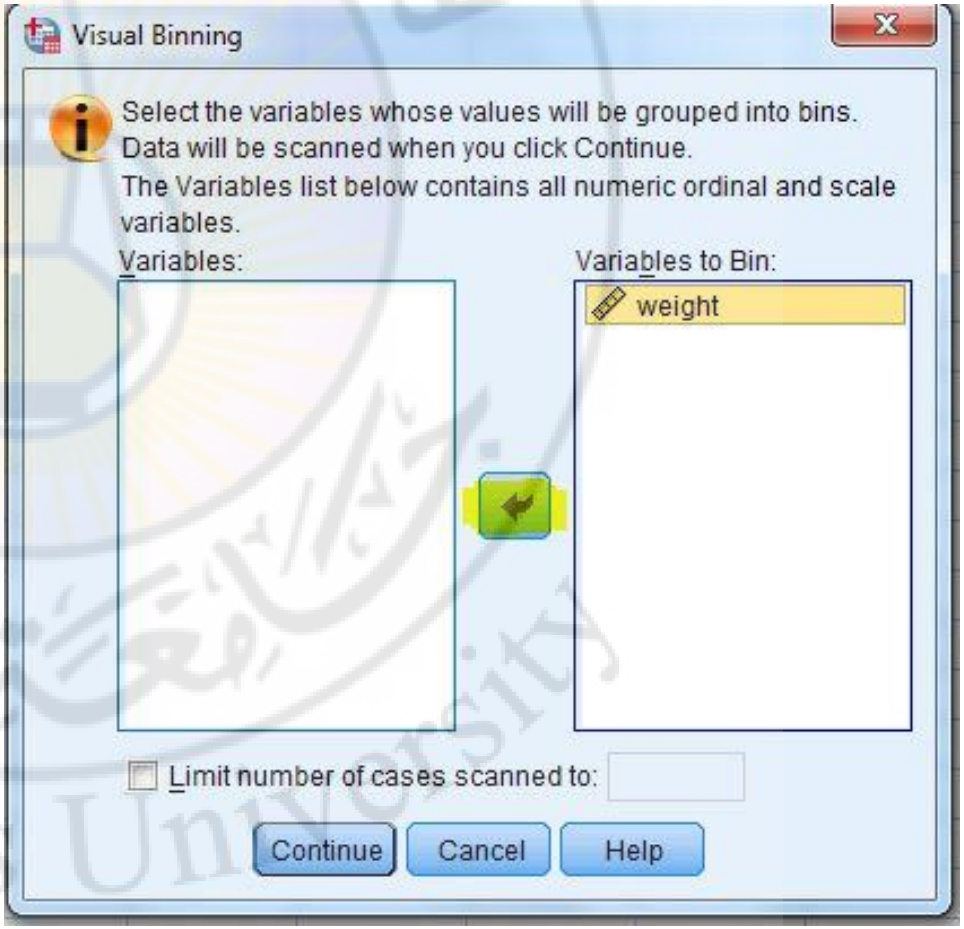
		Frequency
Valid	23.00 - 27.00	4
	28.00 - 32.00	17
	33.00 - 37.00	58
	38.00 - 42.00	64
	43.00 - 47.00	47
	48.00+	19
	Total	209

بالتالي هنا فقط المعلومات التي لدينا تفيد بأن قيم المتغير تتوضع في هذه الفئات وفق تكرارات معينة و لكننا قد فقدنا المعلومات المتعلقة بالقيم الفعلية لهذا المتغير.

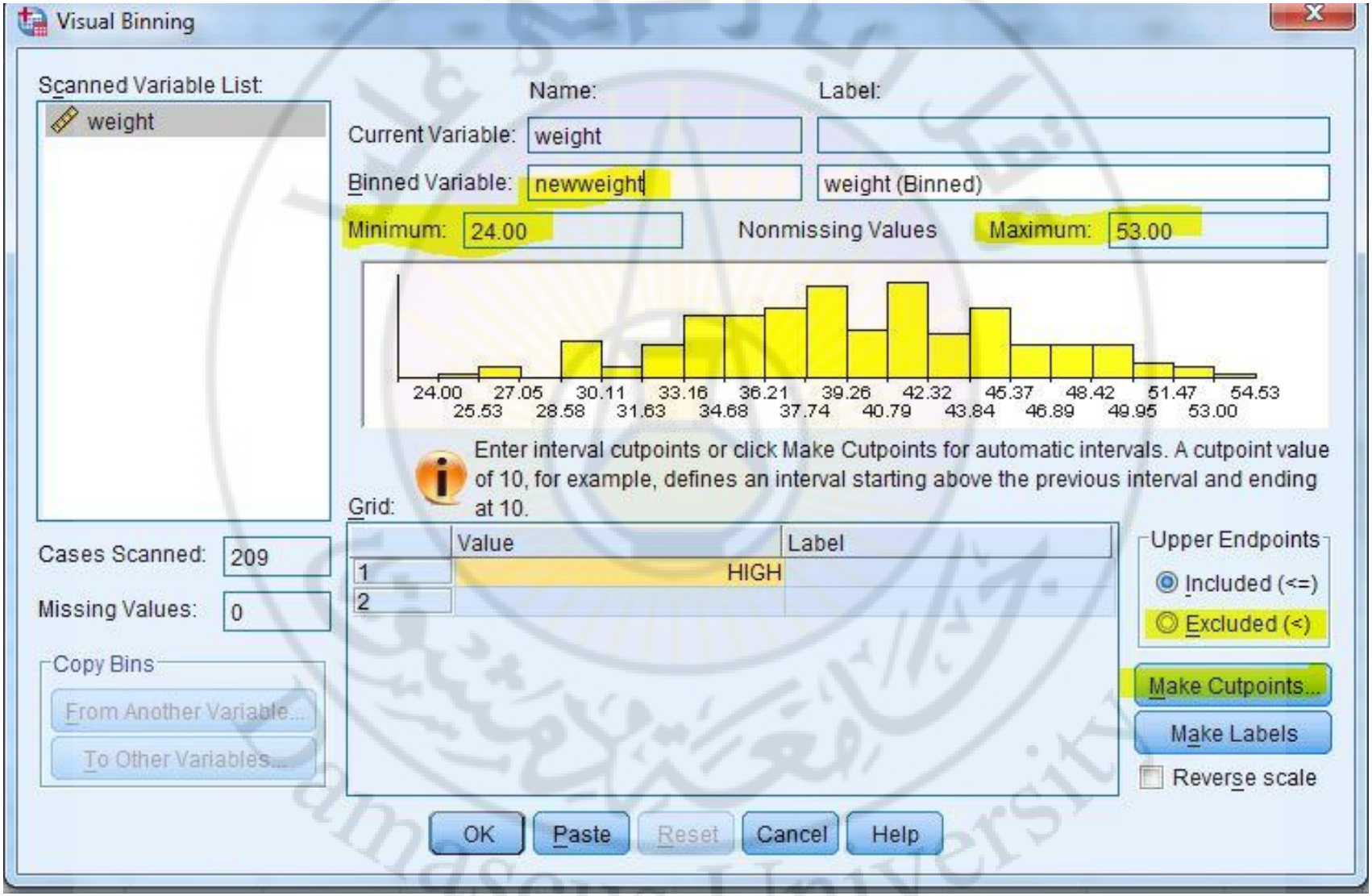
الآن نتكلم عن كيفية بناء جدول توزيع تكراري للمتغير المستمر في برنامج SPSS. بالعودة إلى مثال kids weight.sav سنتكلم عن طريقة واحدة فقط للاستفادة من هذا البرنامج في بناء جدول التوزيع التكراري.

البحث 2: التوزيع التكراري

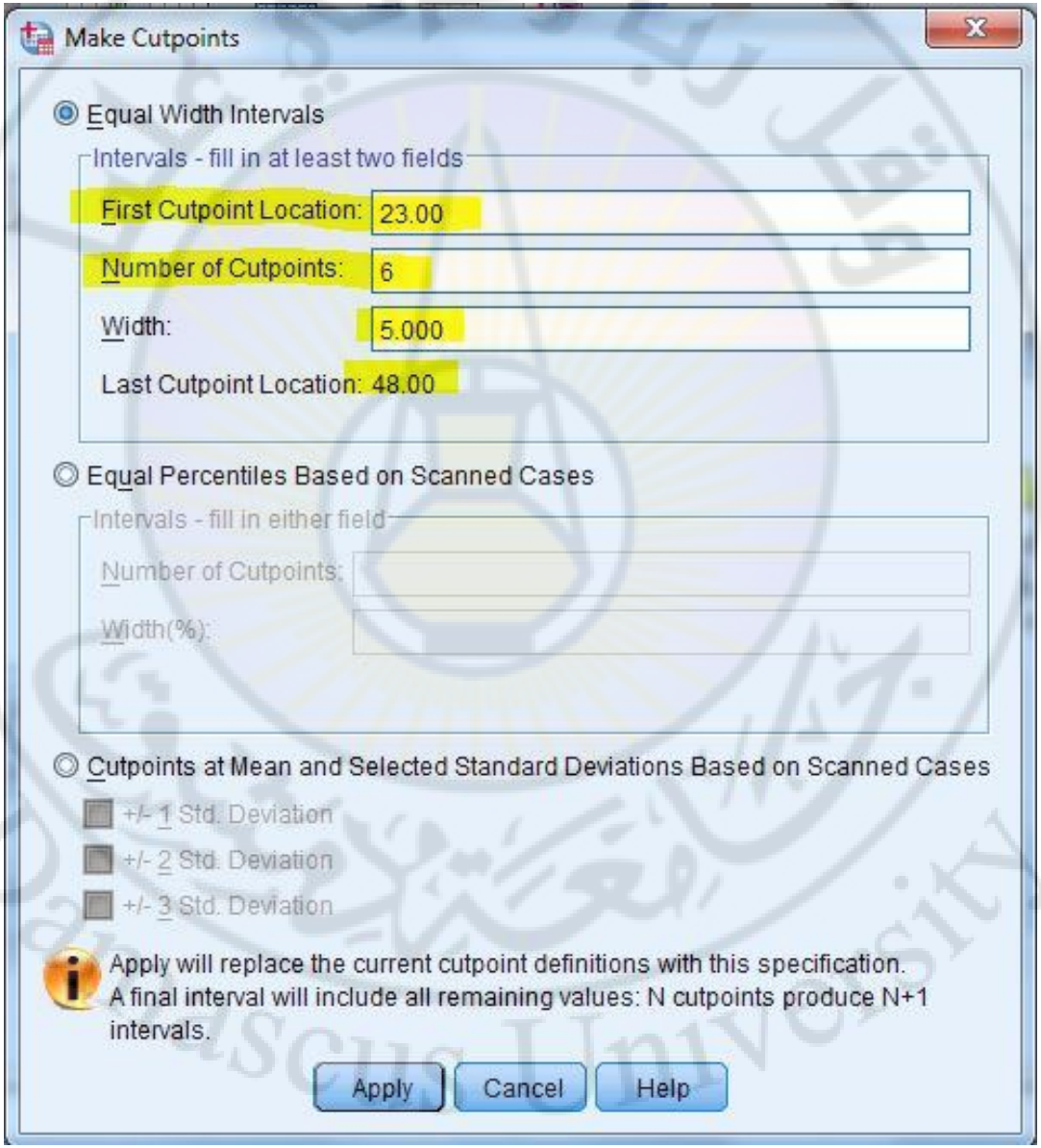




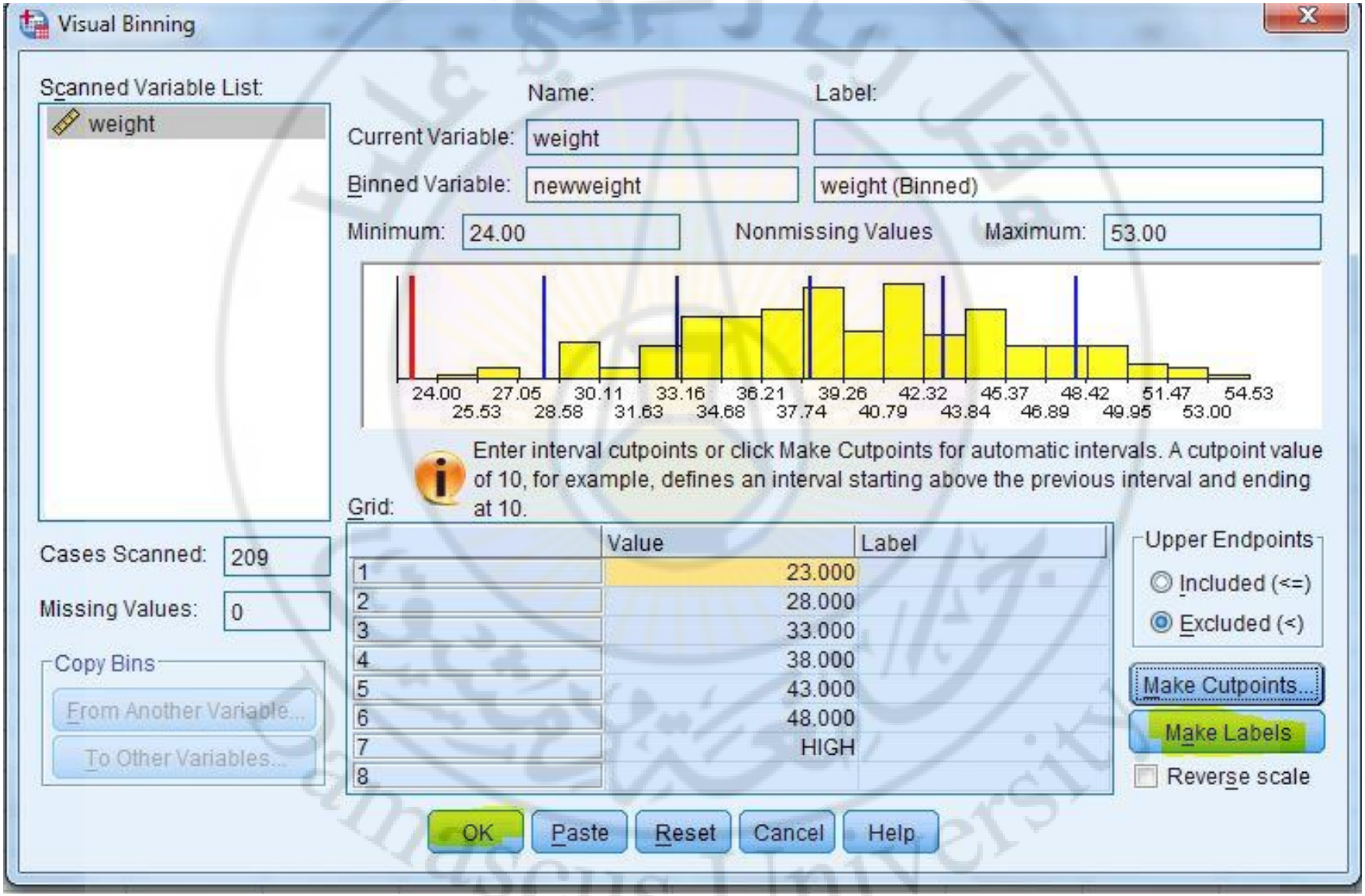
البحث 2: التوزيع التكراري

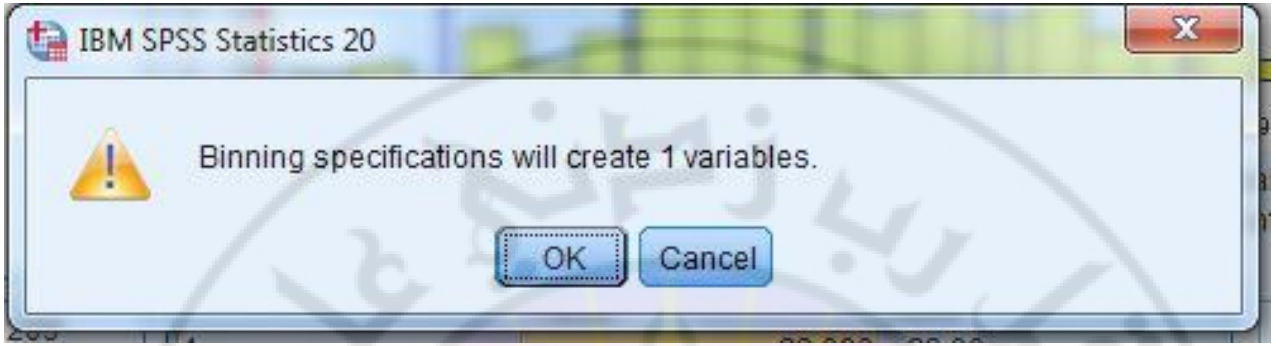


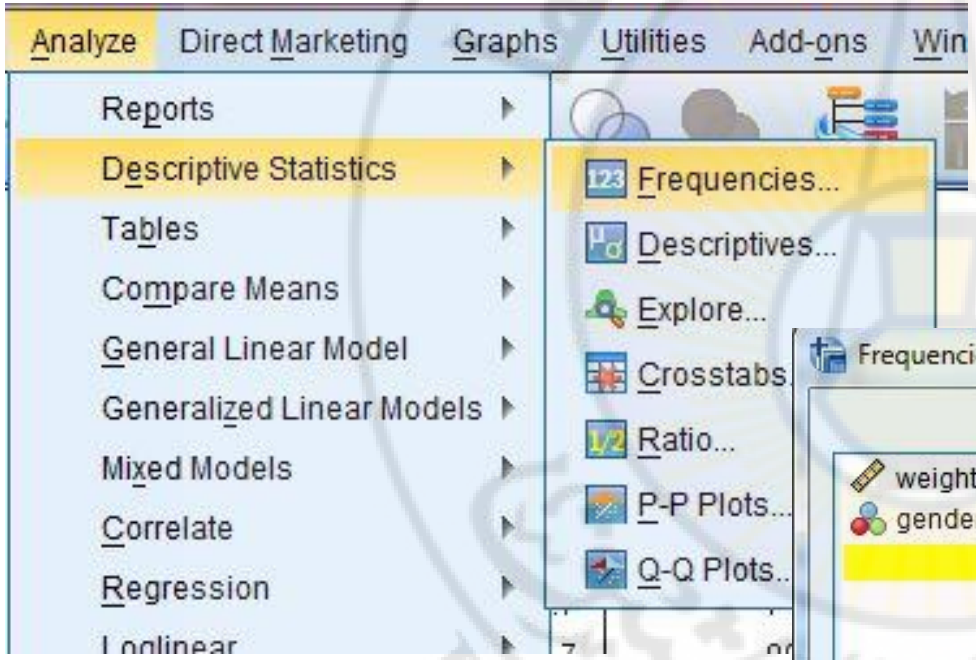
البحث 2: التوزيع التكراري



البحث 2: التوزيع التكراري




IBM SPSS Statistics 20
Binning specifications will create 1 variables.
OK
Cancel


Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Win
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...


Frequencies
weight
gender
Variable(s):
weight (Binned) [ne...
Statistics...
Charts...
Format...
Bootstrap...
Display frequency tables
OK
Paste
Reset
Cancel
Help

البحث 2: التوزيع التكراري

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	23.00 - 27.00	4	1.9	1.9	1.9
	28.00 - 32.00	17	8.1	8.1	10.0
	33.00 - 37.00	58	27.8	27.8	37.8
	38.00 - 42.00	64	30.6	30.6	68.4
	43.00 - 47.00	47	22.5	22.5	90.9
	48.00+	19	9.1	9.1	100.0
	Total	209	100.0	100.0	

لاحظ أن العرض النظري للفئات هو 4 كغ أما العرض الفعلي فهو 5 كغ و ذلك لأن 4=23-27 (الحدود النظرية) بينما 5=22.5-27.5 (الحدود الفعلية).

للتعرف على أساسيات العمل مع برنامج SPSS، كتاب الدكتور أسامة ربيع أمين: التحليل الإحصائى باستخدام برنامج SPSS.

الإحصاء الحيوي

لطلاب السنة الثانية والثالثة - كلية العلوم الصحية

جامعة دمشق

''الجلسة الثامنة''

مدرس المقرر: أ. سلام الجرّاح

إعداد: د. ياسر الزعيم

المدرس في قسم الإحصاء الرياضي

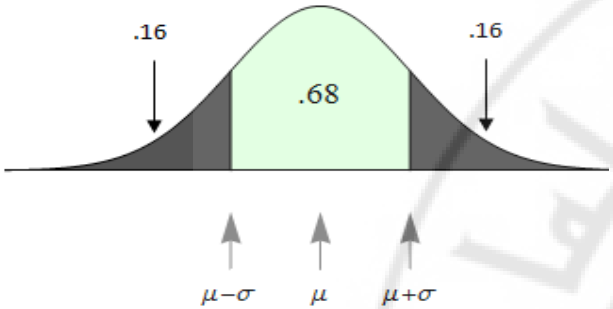
كلية العلوم - جامعة دمشق

آيار– 2023

نتكلم اليوم عن

- حساب المساحات تحت المنحني الطبيعي
- **البحث السادس: تحليل الارتباط الخطي و الانحدار الخطي**

حساب المساحات تحت المنحني الطبيعي

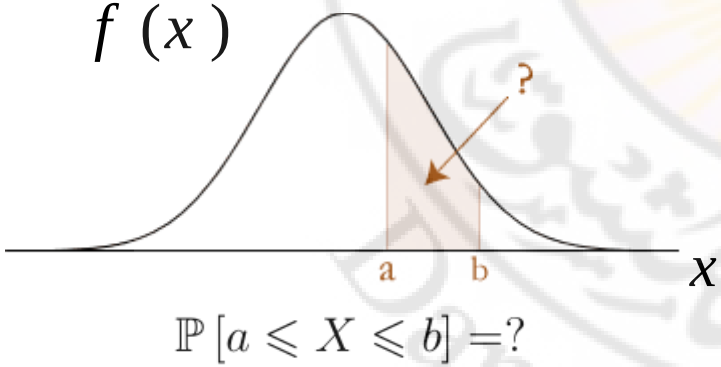


فإنّ: $X \sim N(\mu, \sigma^2)$

$$\mathbb{P}[\mu - \sigma \leqslant X \leqslant \mu + \sigma] = .68$$

$$\mathbb{P}[\mu - 2\sigma \leqslant X \leqslant \mu + 2\sigma] = .95$$

$$\mathbb{P}[\mu - 3\sigma \leqslant X \leqslant \mu + 3\sigma] = .997$$



ماذا عن حساب $P[a \leq X \leq b]$؟

نتعلم طريقتين:

SPSS

استخدام جدول التوزيع الطبيعي المعياري.

حساب المساحات تحت المنحني الطبيعي

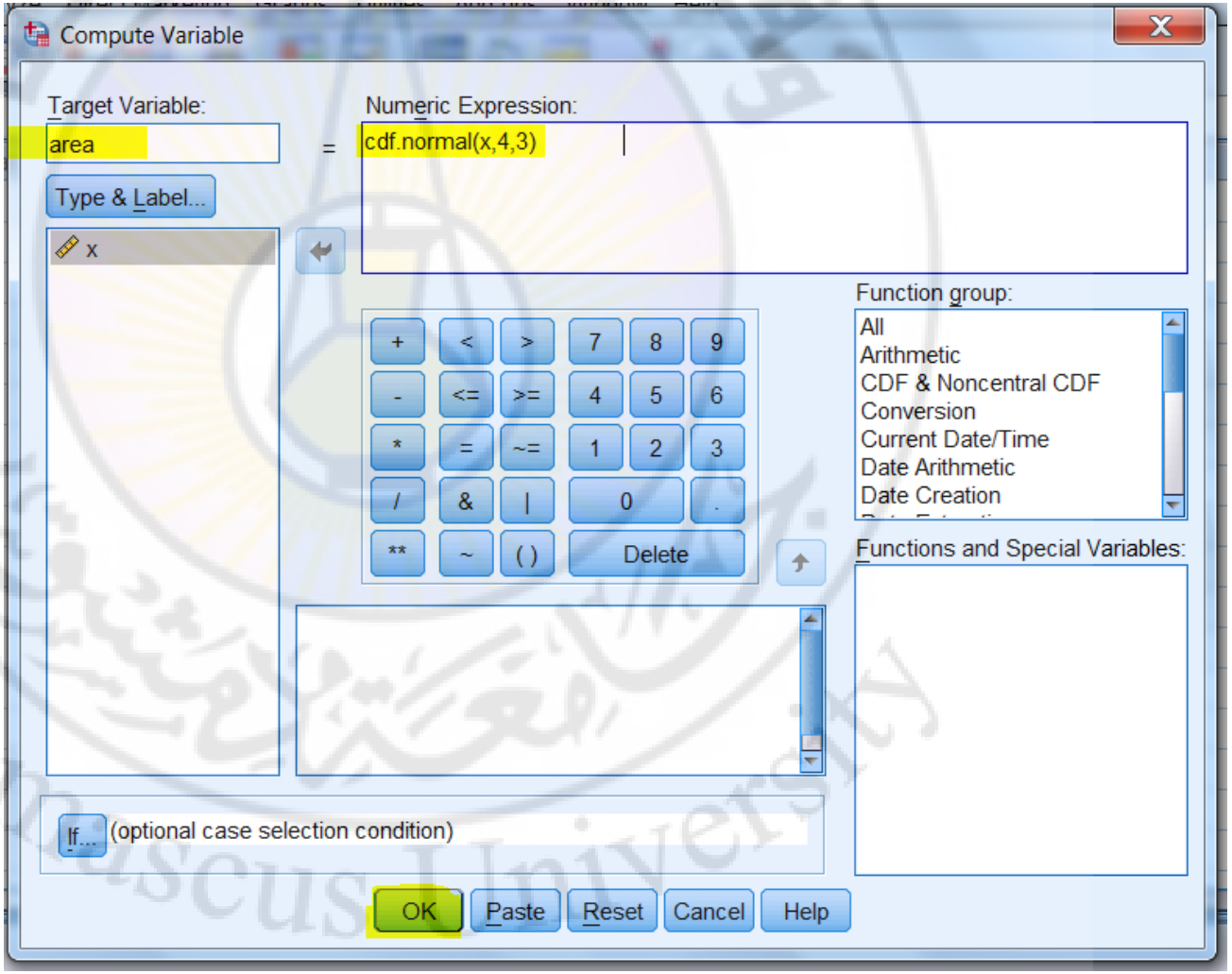
$$\mathbb{P}[-2 \leqslant X \leqslant 3]$$

$$X \sim N(\mu = 4, \sigma^2 = 9)$$

	x
1	-2.00
2	3.00

cdf.normal(x,4,3)

cdf:
Cumulative distribution function



حساب المساحات تحت المنحني الطبيعي

x	area
-2.00	.02
3.00	.37

$$X \sim N(\mu = 4, \sigma^2 = 9)$$

$$\mathbb{P}[-2 \leqslant X \leqslant 3] = \mathbb{P}[X \leqslant 3] - \mathbb{P}[X \leqslant -2] = .37 - .02 = .35$$

35% من المشاهدات تتوضع بين 2- والقيمة 3.

مثال آخر لنحسب للمتحول السابق الاحتمالات التالية:

$\mathbb{P}[X \leqslant 5]$

$\mathbb{P}[2 \leqslant X]$

$$\mathbb{P}[X \leqslant 5] = \mathbb{P}[-\infty \leqslant X \leqslant 5]$$

$$\mathbb{P}[2 \leqslant X] = \mathbb{P}[2 \leqslant X \leqslant \infty]$$

حساب المساحات تحت المنحني الطبيعي

x	area
2.00	.25
1000000.00	1.00
-1000000.00	.00
5.00	.63

$$\mathbb{P}[X \leqslant 5] = \mathbb{P}[-\infty \leqslant X \leqslant 5] = \mathbb{P}[X \leqslant 5] - \mathbb{P}[X \leqslant -\infty] = .63 - 0 = .63$$

$$\mathbb{P}[2 \leqslant X] = \mathbb{P}[2 \leqslant X \leqslant \infty] = \mathbb{P}[X \leqslant \infty] - \mathbb{P}[X \leqslant 2] = 1 - .25 = .75$$

تمرين:

$$Z \sim N(\mu = 0, \sigma^2 = 1)$$

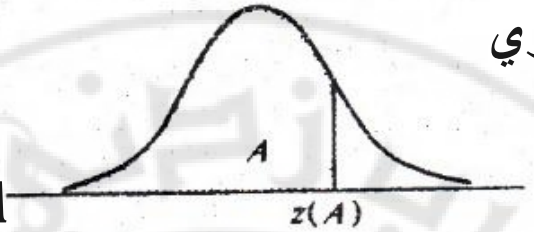
احسب هذه الاحتمالات:

$$\mathbb{P}[-1 \leqslant Z \leqslant 1] \qquad \mathbb{P}[-1 \leqslant Z] \qquad \mathbb{P}[Z \leqslant 0]$$

جدول التوزيع الطبيعي المعياري

$$Z \sim N(0,1)$$

$$\mathrm{P}\left(\mathrm{Z} \leq z(A) = \frac{X-\mu}{\sigma}\right) = A$$



$\mathrm{P}(\mathrm{Z} \leq 0)$

$\mathrm{P}(\mathrm{Z} \leq 1.54)$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936

البحث السادس: تحليل الارتباط الخطي

متغيرين X و Y كميين، نقيس شدة العلاقة الخطية بينهما من خلال:

معامل الارتباط الخطي بيرسون Pearson correlation coefficient.

Karl Pearson (1857–1936)

بفرض $x_1, x_2, \ldots, x_N$ و $y_1, y_2, \ldots, y_N$ عينتين من المتحولين X و Y، إنّ r_{xy} (أو r_{yx}) هو مقياس لشدة وجهة العلاقة الخطية بين المتحولين X و Y

$$r_{xy} = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}}$$

إنّ r_{xy} محسوب لأجل عينات من المتحولين، أمّا فيما يتعلق بمجتمعي المتحولين فإنّ المطلوب هو حساب:

(rho) $$\rho_{XY} = \frac{\sum_{i=1}^{Total}(X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{Total}(X_i - \mu_X)^2}\sqrt{\sum_{i=1}^{Total}(Y_i - \mu_Y)^2}}$$

وهذا غير ممكن، لذا سنتعلم في SPSS كيف نحسب r_{xy}

ومن ثم نصل للمجتمعات من خلال الاختبار الإحصائي التالي:

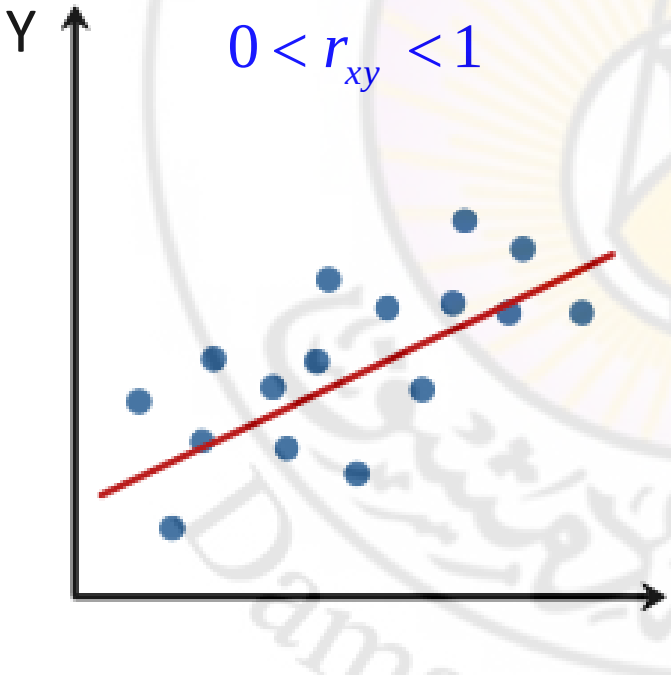
$H_0 : \rho_{XY} = 0 \; versus \; H_1 : \rho_{XY} \neq 0$ عند مستوى أهمية α

البحث السادس: تحليل الارتباط الخطي

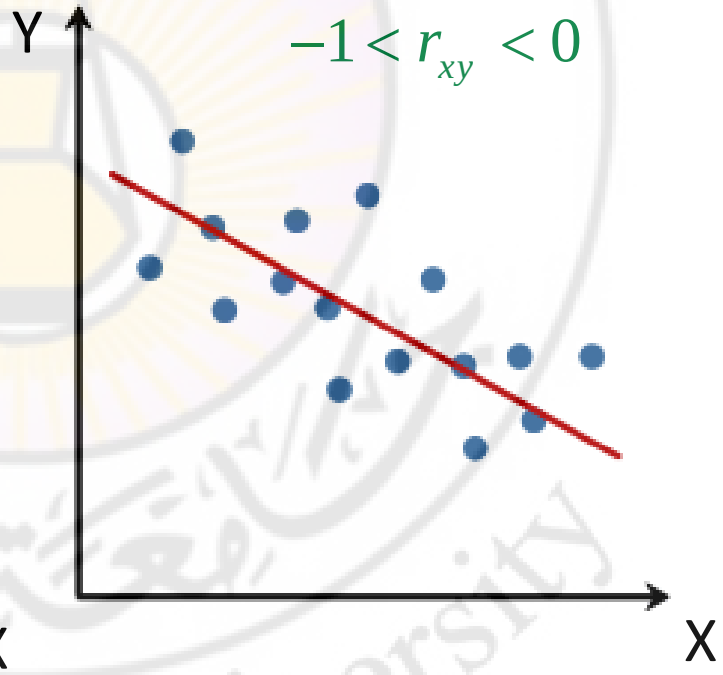
ارتباط خطي موجب ضعيف القوة	$0 \le r \le 0.3$
ارتباط خطي سالب ضعيف القوة	$-0.3 \le r \le 0$
ارتباط خطي موجب متوسط القوة	$0.3 < r \le 0.6$
ارتباط خطي سالب متوسط القوة	$-0.6 \le r < -0.3$
ارتباط خطي موجب قوي	$0.6 < r \le 0.9$
ارتباط خطي سالب قوي	$-0.9 \le r < -0.6$
ارتباط خطي موجب تام	$0.9 < r \le 1$
ارتباط خطي سالب تام	$-1 \le r < -0.9$

البحث السادس: تحليل الارتباط الخطي

علاقة خطية إيجابية (طردية) بين المتحولين



علاقة خطية سلبية (عكسية) بين المتحولين



شروط استخدام معامل بيرسون:

لحساب ارتباط بيرسون الخطي بين متحولين X و Y كميين، يجب توفر ما يلي:

inc_aft	inc_bef
12.00	8.00
10.00	8.00
11.00	8.00
18.00	9.00
12.00	7.00
15.00	8.00
13.00	8.00
22.00	9.00

training.sav

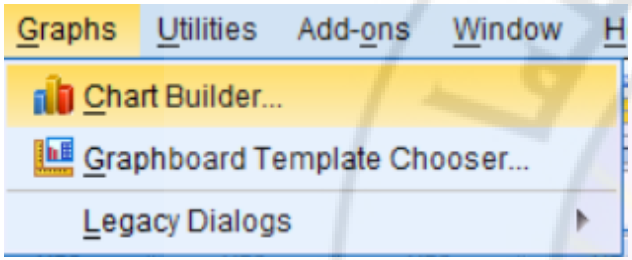
1) أن تكون البيانات ازدواجية related pairs
أي أنّ كل مشاهدة من X تقابلها مشاهدة من Y.

2) خلو كل من المتحولين من المشاهدات القاصية extremes

3) (بديل عن الشرط السابق) أن يتوزع كل من المتحولين توزعاً طبيعياً.

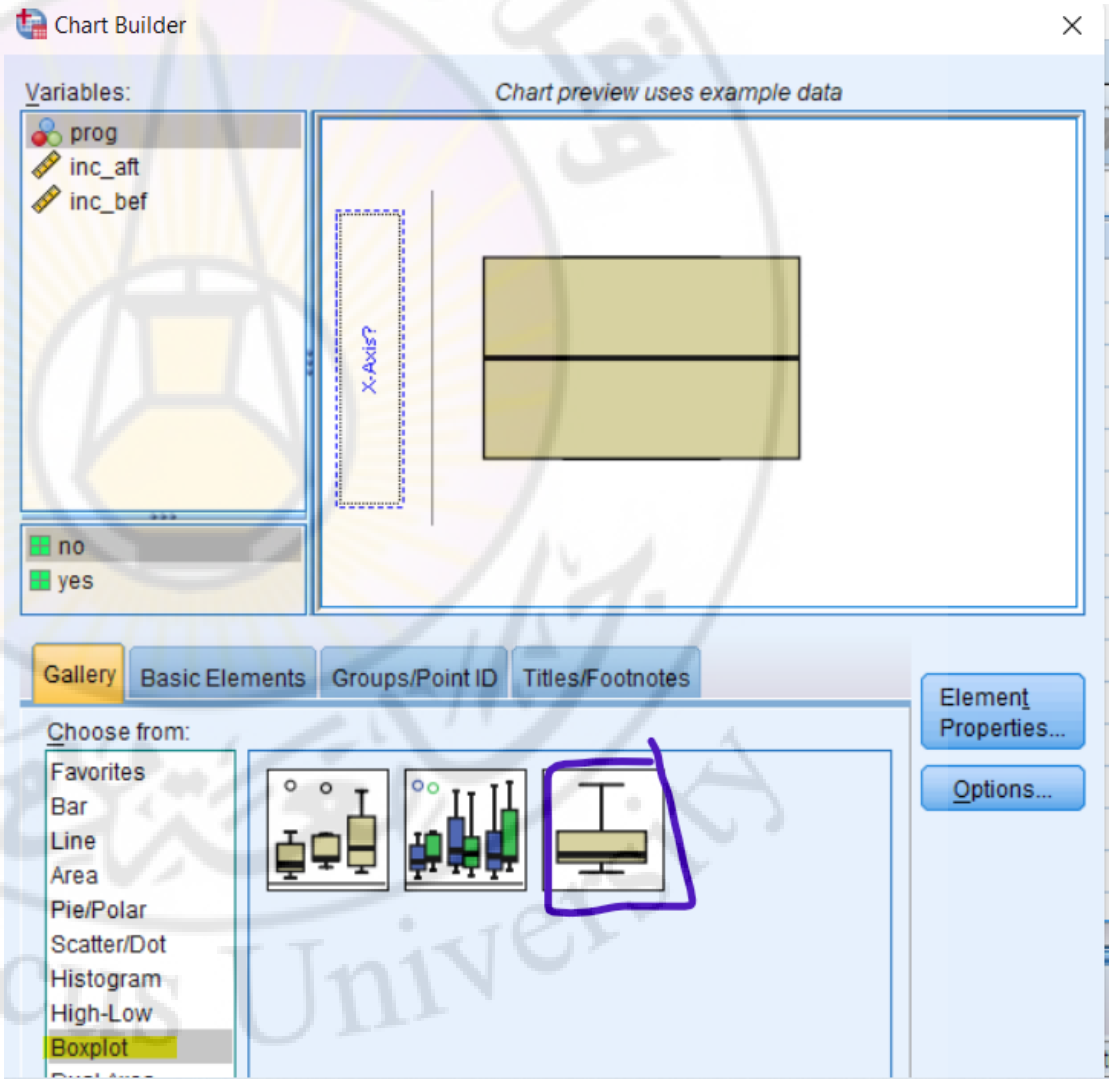
البحث السادس: تحليل الارتباط الخطي

مثال: training.sav نريد حساب ارتباط بيرسون بين inc_aft و inc_bef



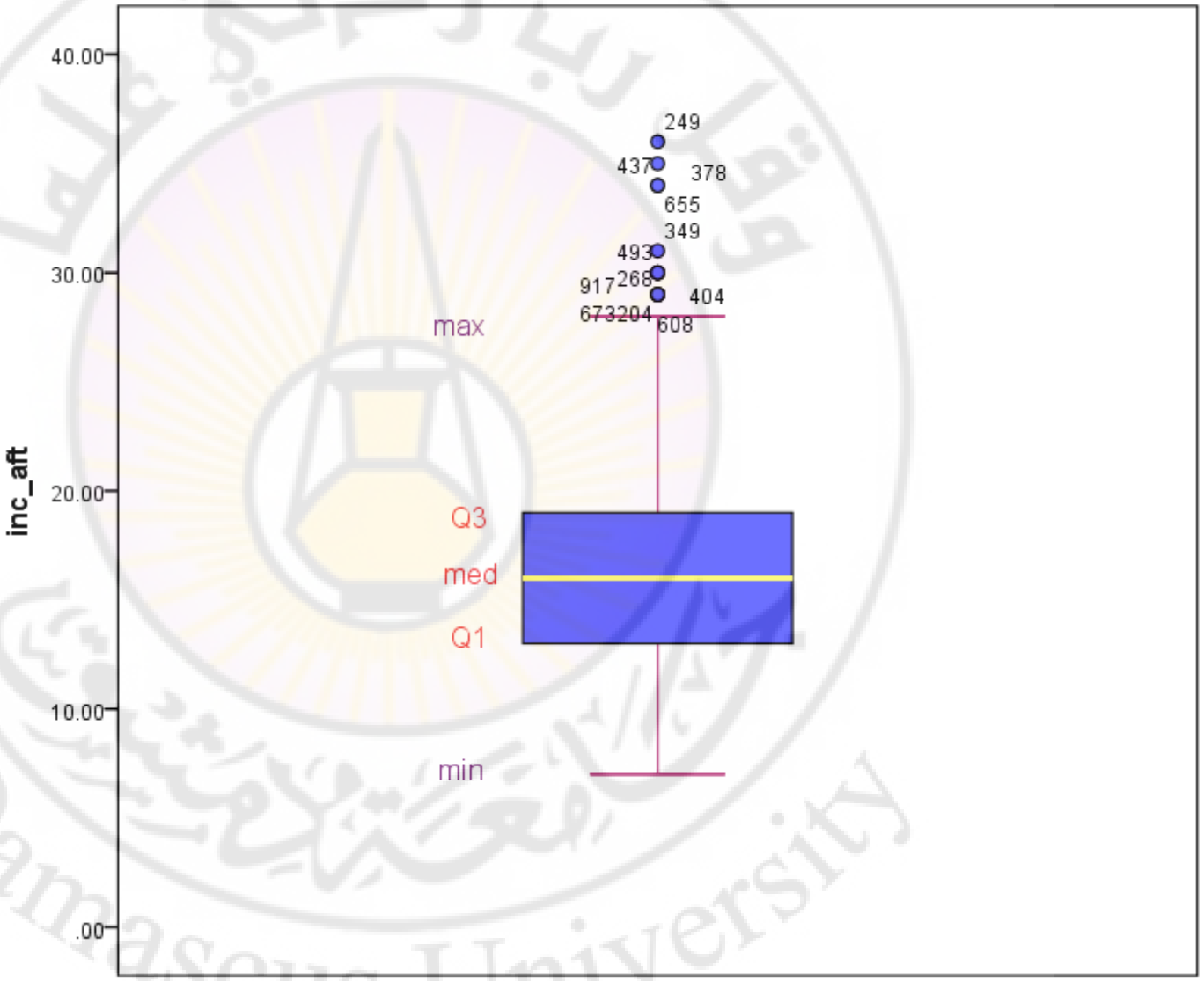
التأكد من خلو مشاهدات المتحولين من النقط القاصية، ولهذا نستخدم رسم الصندوق box plot

أي نرسم مخطط الصندوق لكل متحول على حدة



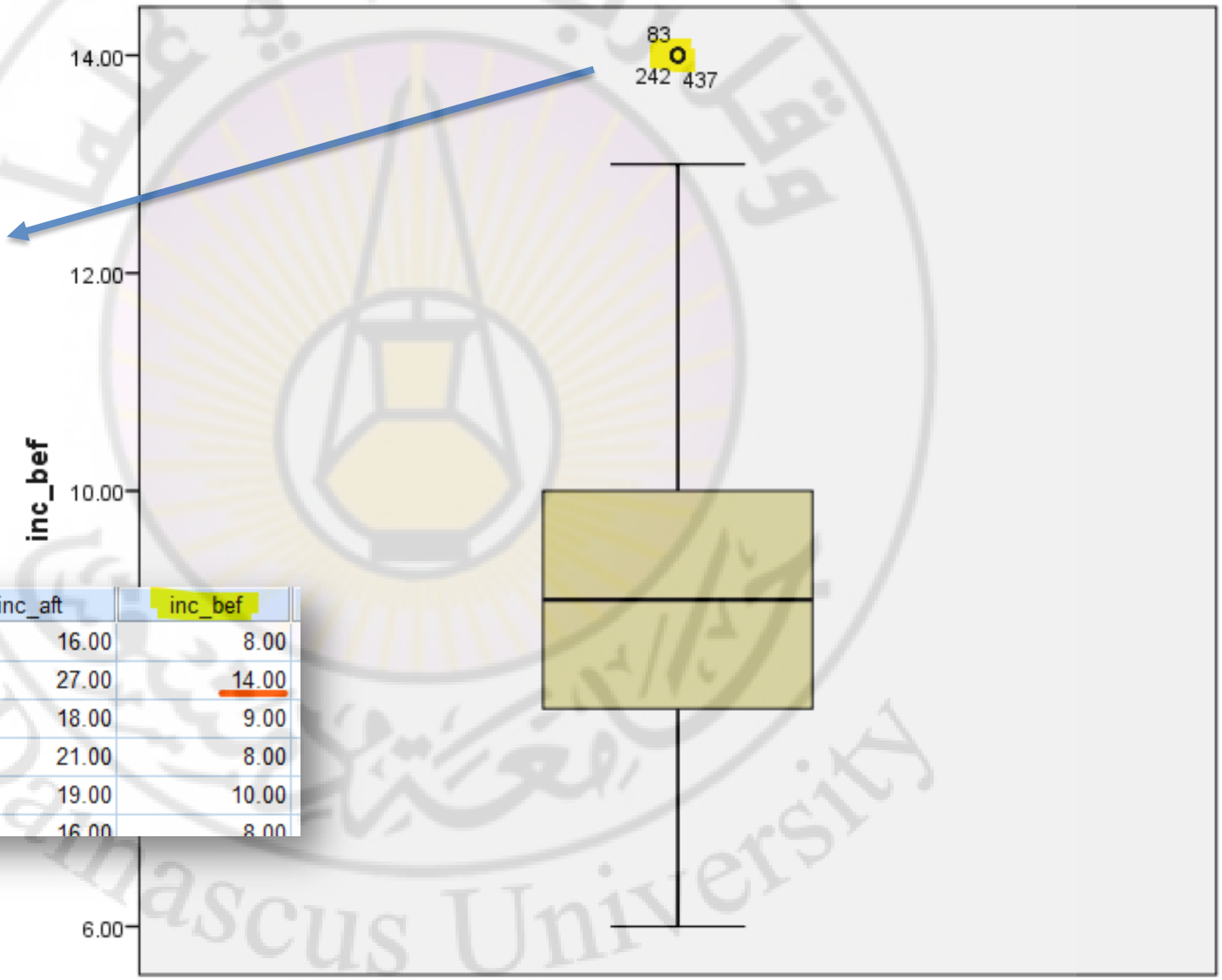
البحث السادس: تحليل الارتباط الخطي

ندعو النقط دائرية الشكل بالنقط المنعزلة outliers وهي نقط لا تؤثر بشكل كبير على قيمة معامل بيرسون.



رسم الصندوق للمتحول inc_aft

البحث السادس: تحليل الارتباط الخطي

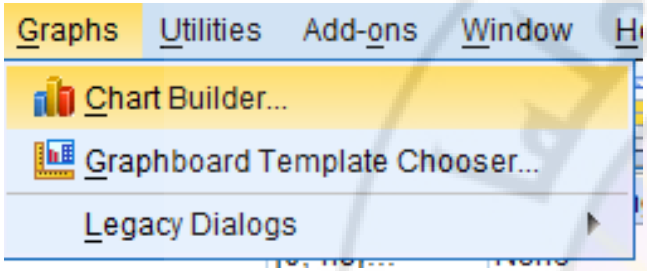


الأرقام على النقط المنعزلة هي دلالة على رقم السطر في بيانات المتحول

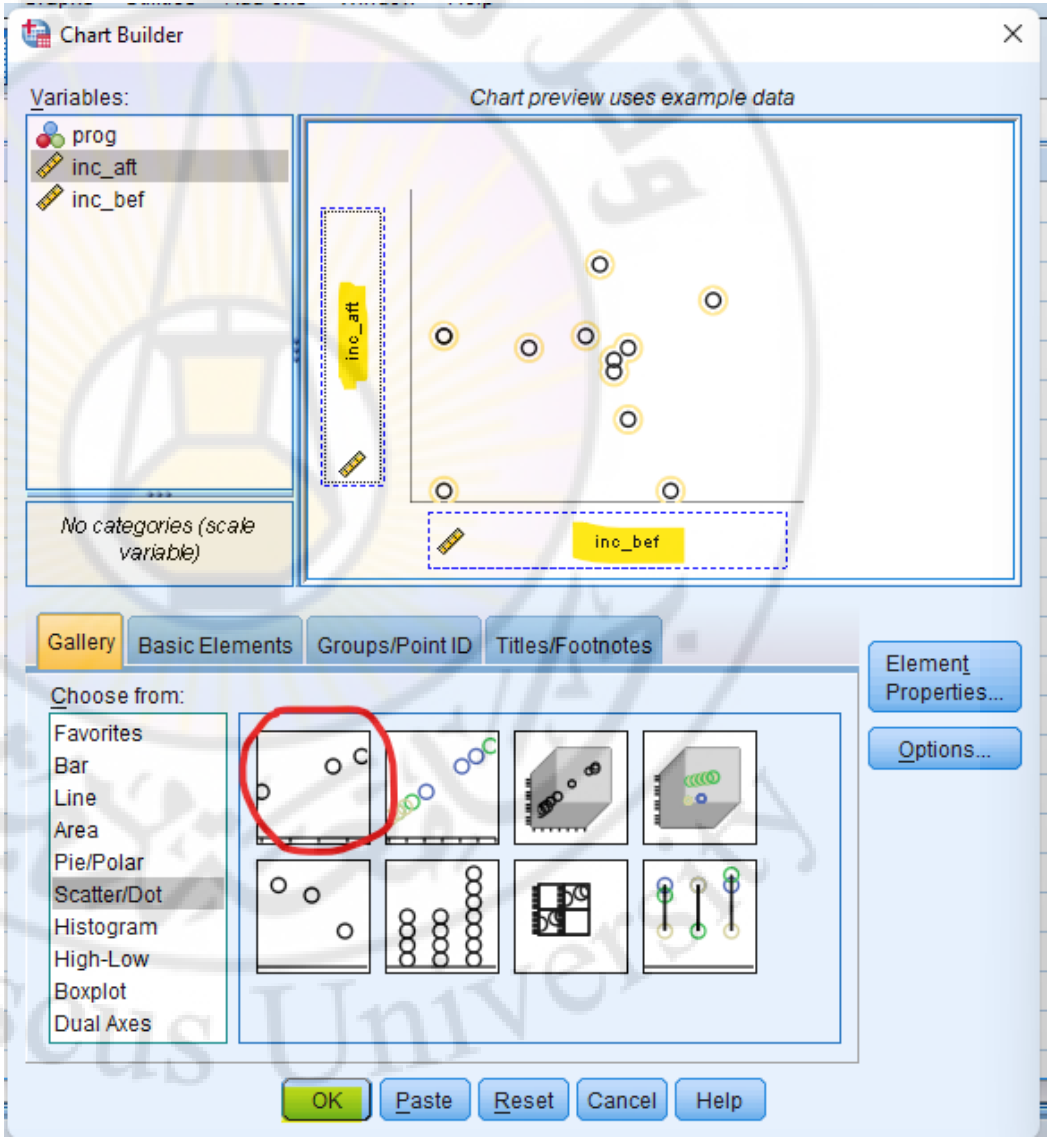
	prog	inc_aft	inc_bef
82	1	16.00	8.00
83	0	27.00	14.00
84	0	18.00	9.00
85	1	21.00	8.00
86	1	19.00	10.00
87	1	16.00	8.00

البحث السادس: تحليل الارتباط الخطي

شكل الانتشار (التبعثر) Scatter plot:

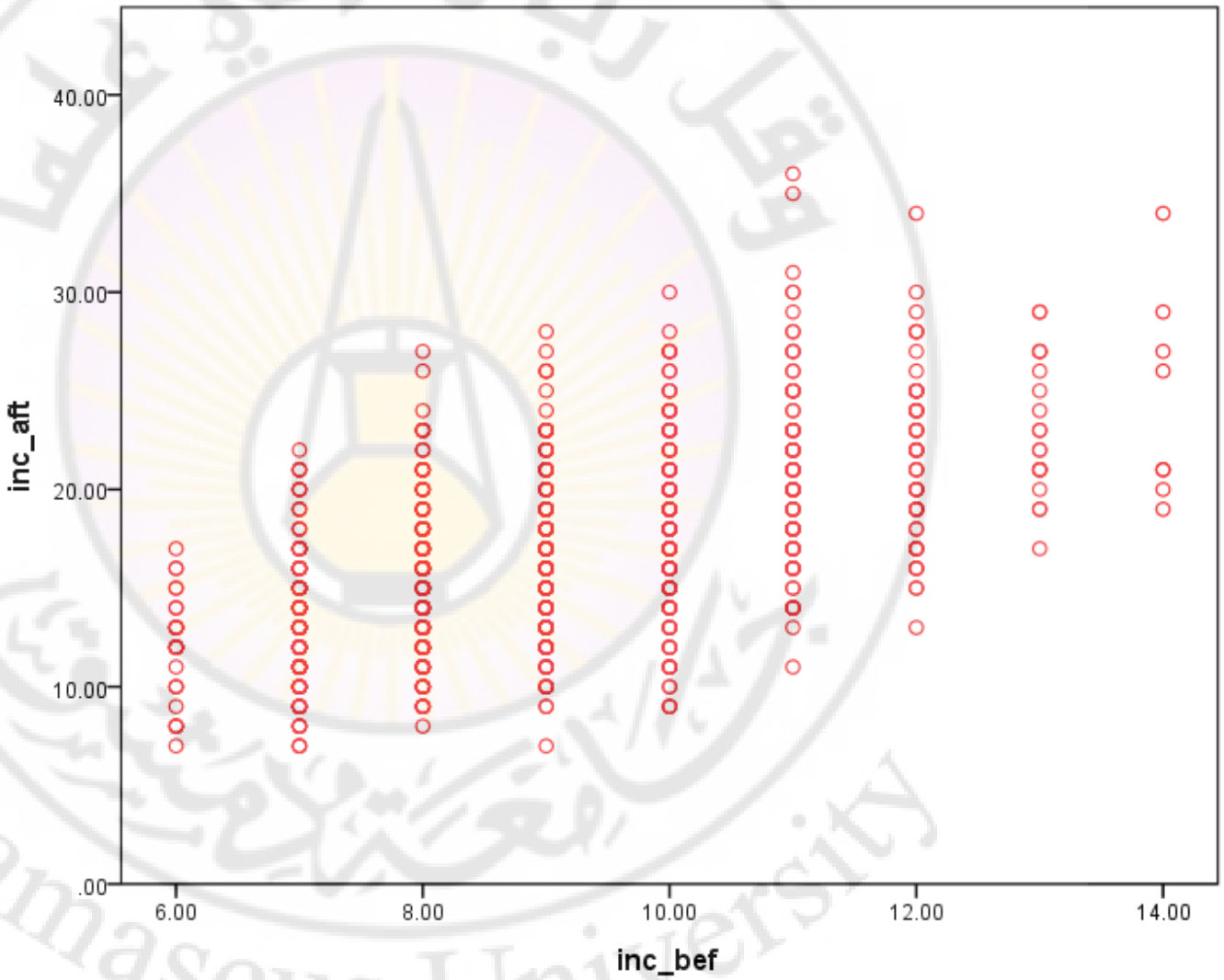


training.sav

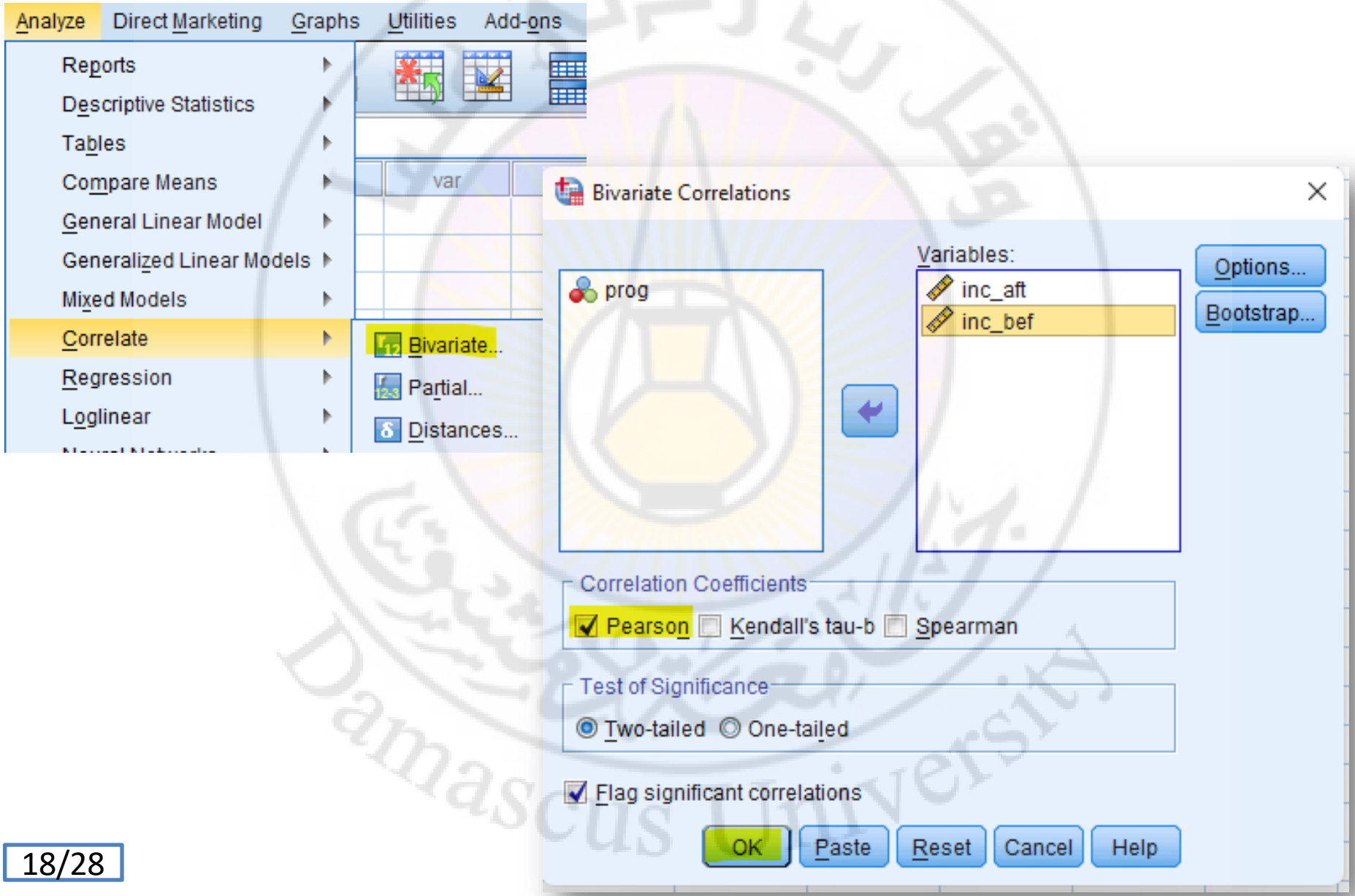


البحث السادس: تحليل الارتباط الخطي

يبدو أن العلاقة بين المتغيرين المدروسين خطية طردية (إيجابية)، حيث أن المتغيرين يزدادان معاً و ينقصان معاً.



البحث السادس: تحليل الارتباط الخطي



البحث السادس: تحليل الارتباط الخطي

Correlations

		inc_aft	inc_bef
inc_aft	Pearson Correlation	1	.589**
	Sig. (2-tailed)		.000
	N	1000	1000
inc_bef	Pearson Correlation	.589**	1
	Sig. (2-tailed)	.000	
	N	1000	1000

$r = 0.589$

$H_0 : \rho = 0 \quad v.s.$

$H_1 : \rho \neq 0$

$sig = .000 < \alpha = .05$

conclude H_1

بما أنّ القرار هو قبول الفرضية البديلة، فهذا يعني بأنّ الارتباط الخطي المشاهد بين المتغيرين المدروسين هو ارتباط خطي حقيقي (معنوي، هام، ذو دلالة إحصائية) في مجتمعي الدراسة وهو ارتباط خطي موجب متوسط الدرجة.

البحث السادس: تحليل الانحدار الخطي

مقدمة:

إنّ الانحدار الخطي البسيط simple linear regression هو عبارة عن بناء المعادلة:

$$y=\beta_0+\beta_1 x+\varepsilon \quad \Leftrightarrow \quad y_i=\beta_0+\beta_1 x_i+\varepsilon_i \;\; ; \;\; i=1,2,\ldots,N$$

beta zero beta one epsilon

في هذه المعادلة: لدينا المشاهدات $(x_i, y_i)\;;\; i=1,2,\ldots,N$ معلومة.

أمّا المعالم (الوسطاء) parameters β_0 و β_1 فهي مجاهيل يجب تقديرها.

ε_i هي أخطاء عشوائية مجهولة ويفترض أنّ $\varepsilon_i \sim N(0,\sigma^2)$

ويتم تقدير الأخطاء هذه بمقادير ندعوها الرواسب residuals

معادلة خط الانحدار الخطي البسيط $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

simple linear regression model (equation)

الهدف هو التنبؤ بالمتحول Y من خلال المتحول X.

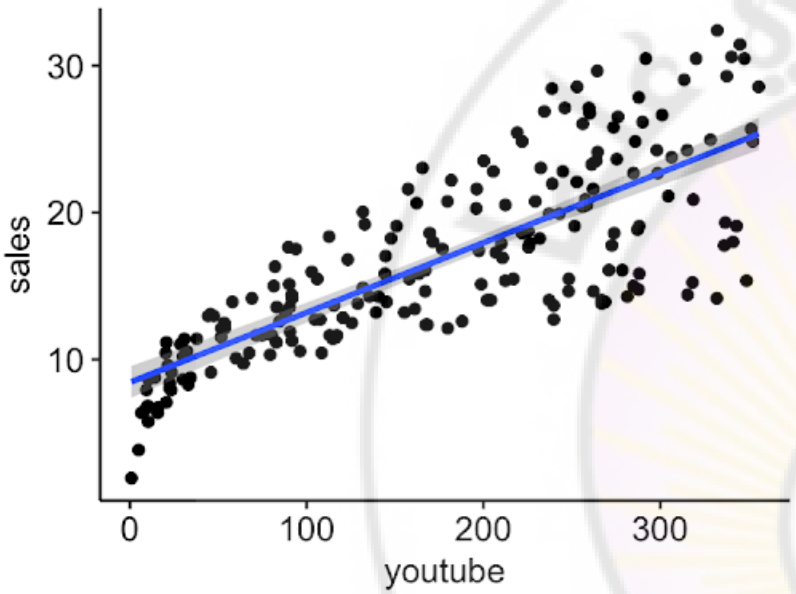
لذلك ندعو X بالمتحول المستقل (متحول التنبؤ - متحول الانحدار – متحول التفسير – المتحول المميز)

independent (predictor – regressor – explanatory - feature) variable

ندعو Y بالمتحول التابع (المتنبأ به – المتحول المفسر – المنحدر عليه – متحول النتيجة – متحول الهدف)

dependent (predicted – explained – regressand –outcome – target) variable

البحث السادس: تحليل الانحدار الخطي



مثال: التنبؤ بحجم المبيعات (ألف ليرة سورية) لمنتَج من خلال معرفة عدد المشاهدات (ألف شخص) للدعاية على موقع يوتيوب.

pair	عدد المشاهدات	حجم المبيعات
1	0	0
2	.3	2.2
3	30	325
4	21	297
5	13	128
... وهكذا	etc...	... و إلى آخره

x y

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
$$\widehat{\text{sales}}_i = \hat{\beta}_0 + \hat{\beta}_1 * (\text{youtube})_i$$

شروط بناء نموذج الانحدار الخطي البسيط (معادلة خط مستقيم):

1- كلا المتحولين X و Y مستمرين.

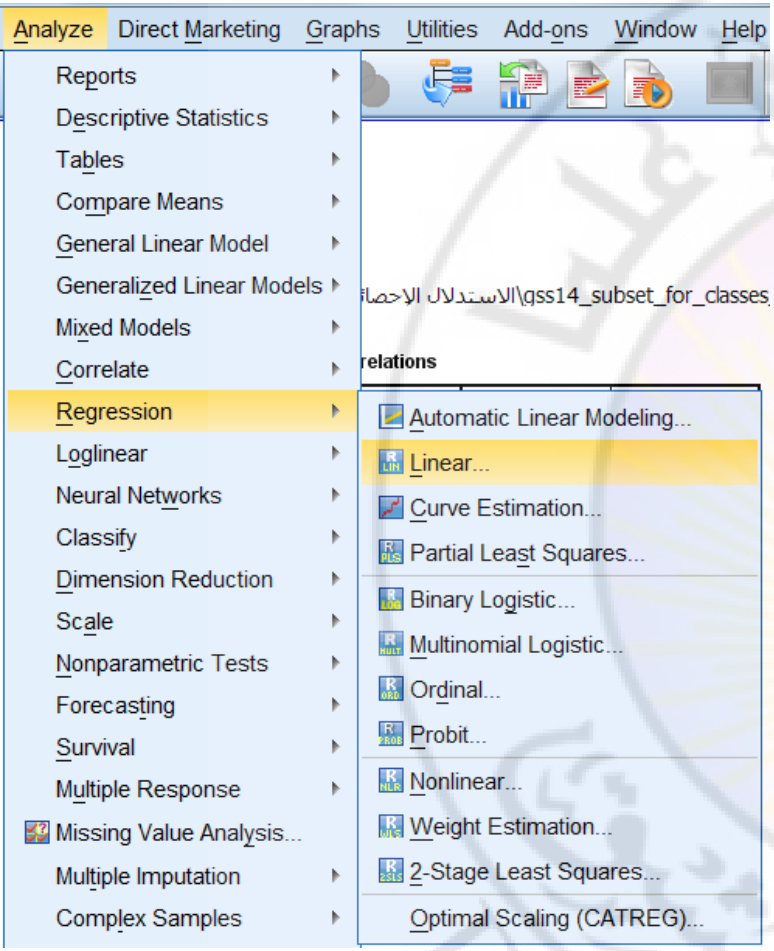
2- خلو كل من المتحولين من النقط القاصية.

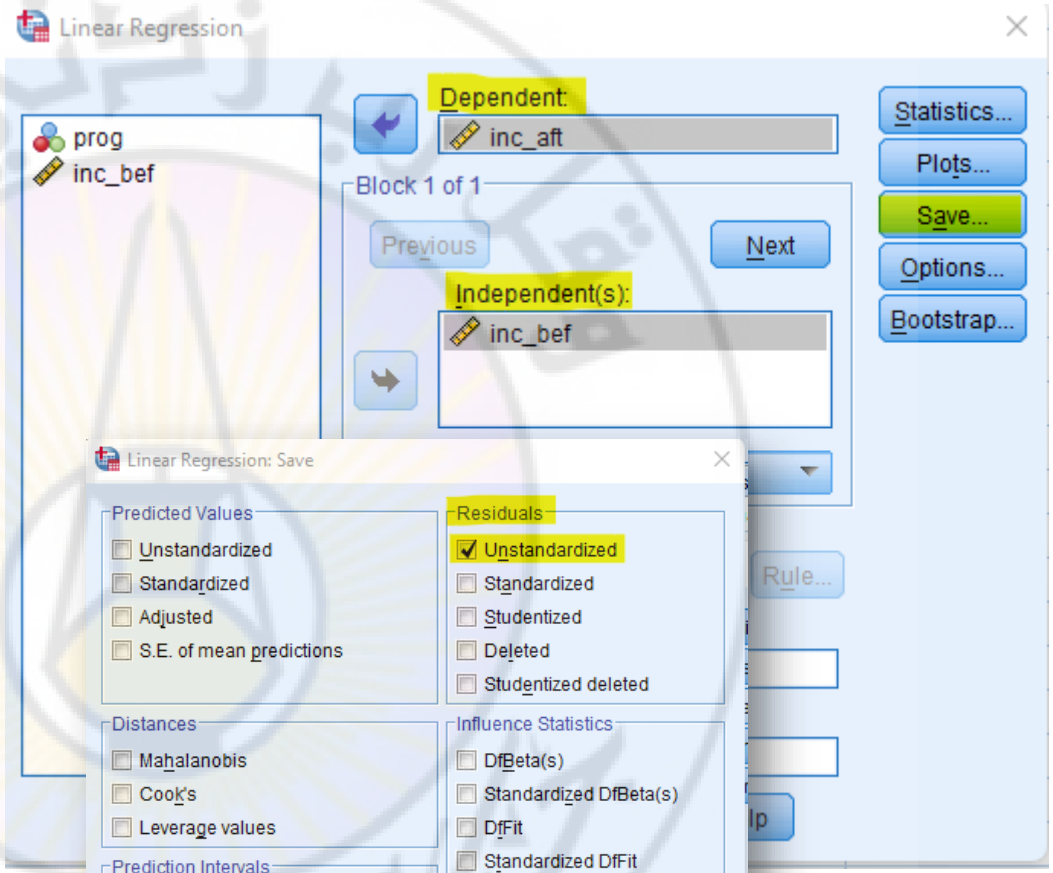
3- الارتباط الخطي الحقيقي.

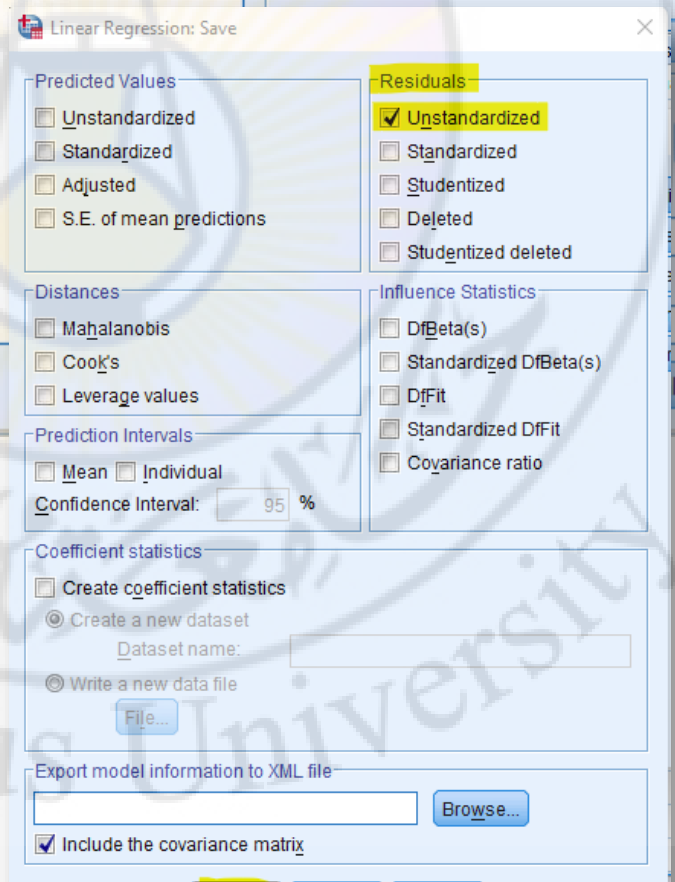
أمّا بعد بناء النموذج يجب توافر شروط لازمة لتعميم النموذج على المجتمعين الخاصين بالدراسة.

إنّ **الراسب** هو الفرق بين القيمة المشاهدة للمتحول التابع والقيمة المتوقعة له، أي أنّ $e_i = y_i - \hat{y}_i$ حيث $i = 1, 2, ..., N$

مثال training.sav







البحث السادس: تحليل الانحدار الخطي

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.589[a]	.347	.346	3.77714

a. Predictors: (Constant), inc_bef

b. Dependent Variable: inc_aft

$$R = r_{xy} = 0.589$$

$$R^2 = (\mathrm{R})^2 = 0.347$$

إنّ R^2 يسمى معامل التحديد **coefficient of determination** وهو:
النسبة المئوية التي يفسرها المتحول المستقل من التغير الكلي في قيم المتحول التابع.

$R^2 = 0.347 = 34.7\%$ أي أن 34.7% من التغير والتشتت في قيم المتحول $y = \mathrm{inc\_aft}$ مسؤول عنها المتحول $x = \mathrm{inc\_bef}$

البحث السادس: تحليل الانحدار الخطي

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7555.079	1	7555.079	529.556	.000[b]
	Residual	14238.272	998	14.267		
	Total	21793.351	999			

SSR=sum of squares due to regression

(regression sum of squares)

مجموع مربعات الانحدار

SSR=7555.079

SSE=residual sum of squares

مجموع مربعات الخطأ (البواقي)

SSE=14238.272

SST=total sum of squares

مجموع المربعات الكلي

SST=21793.351

H_0 : (النموذج الحالي المقترح غير مناسب للتنبؤ بالمتحول التابع)

H_1 : (النموذج الحالي المقترح مناسب و ملائم للتنبؤ بالمتحول التابع)

البحث السادس: تحليل الانحدار الخطي

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.548	.665		2.329	.020
	inc_bef	1.680	.073	.589	23.012	.000

a. Dependent Variable: inc_aft

$\hat{\beta}_0 = 1.548 \quad \& \quad \hat{\beta}_1 = 1.680$

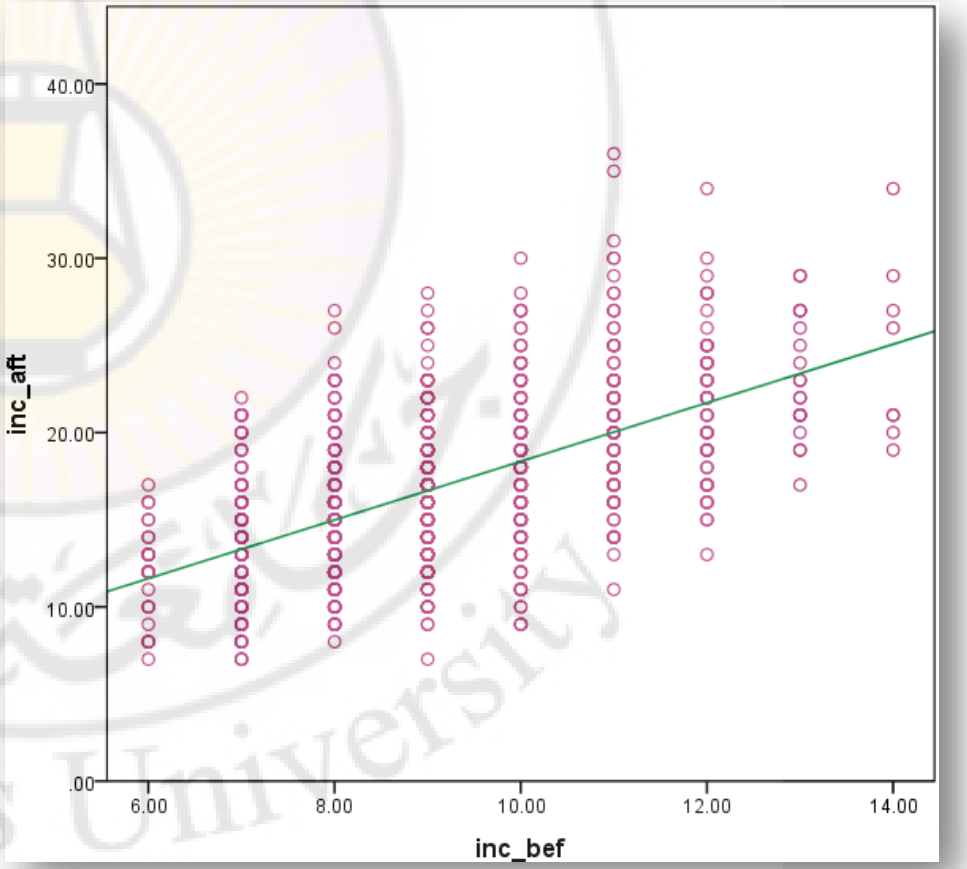
$\hat{y} = 1.548 + 1.680x$

$H_0 : \beta_1 = 0 \quad v.s. \quad H_1 : \beta_1 \neq 0$

$sig = .000 < \alpha = .05$

$t_0 = 23.012$

أي إنّ معرفة الراتب قبل الدورة التدريبية هام للتنبؤ بالراتب بعد الدورة التدريبية.



البحث السادس: تحليل الانحدار الخطي

إذاً معادلة الانحدار الخطي البسيط هي $\hat{y} = 1.548 + 1.680x$

لنفرض أنّ موظفاً راتبه قبل الدورة التدريبية $x = 13.5$ ، عندئذ نتوقع بأن يكون راتبه بعد انتهاء مدة الدورة التدريبية هو

$$\hat{y} = 1.548 + 1.680 * 13.5 = 24.228$$

التوزع الطبيعي للرواسب

	prog	inc_aft	inc_bef	RES_1
1	0	12.00	8.00	-2.99000
2	0	10.00	8.00	-4.99000
3	0	11.00	8.00	-3.99000
4	1	18.00	9.00	1.32971
5	0	12.00	7.00	-1.30970
6	1	15.00	8.00	.01000
7	0	13		
8	1	22		
9	1	18		
10	0	9		

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	.051	1000	.000	.987	1000	.000

إنّ رواسب النموذج لا تتوزع طبيعياً وهذا ليس جيداً.

الإحصاء الحيوي

لطلاب السنة الثانية والثالثة - كلية العلوم الصحية

جامعة دمشق

”الجلسة السادسة“

مدرس المقرر: أ. سلام الجرّاح

إعداد: د. ياسر الزعيم

المدرس في قسم الإحصاء الرياضي

كلية العلوم - جامعة دمشق

آيار– 2023

نتكلم اليوم عن:

- **البحث الرابع: القيم المعيارية**
- **البحث الخامس: المتغير العشوائي و التوزيعات الاحتمالية**

البحث 4: القيم المعيارية

القيم z (القيم المعيارية):

z values (z-scores or standardized values)

إنّ القيم المعيارية للمتحول X هي بالتعريف: $z = \frac{X - \bar{X}}{s}$

أو نكتب: $z_i = \frac{X_i - \bar{X}}{s} \;\; ; \;\; \forall i = 1,2,\dots,N$

و يتحقق أنّ: $\bar{z} = 0 \;\; , \;\; \mathrm{SD}(z) = 1$

إنّ القيم z ليس لها واحدة قياس أي هي أعداد scalars

البحث 4: القيم المعيارية

مثال kids weight.sav

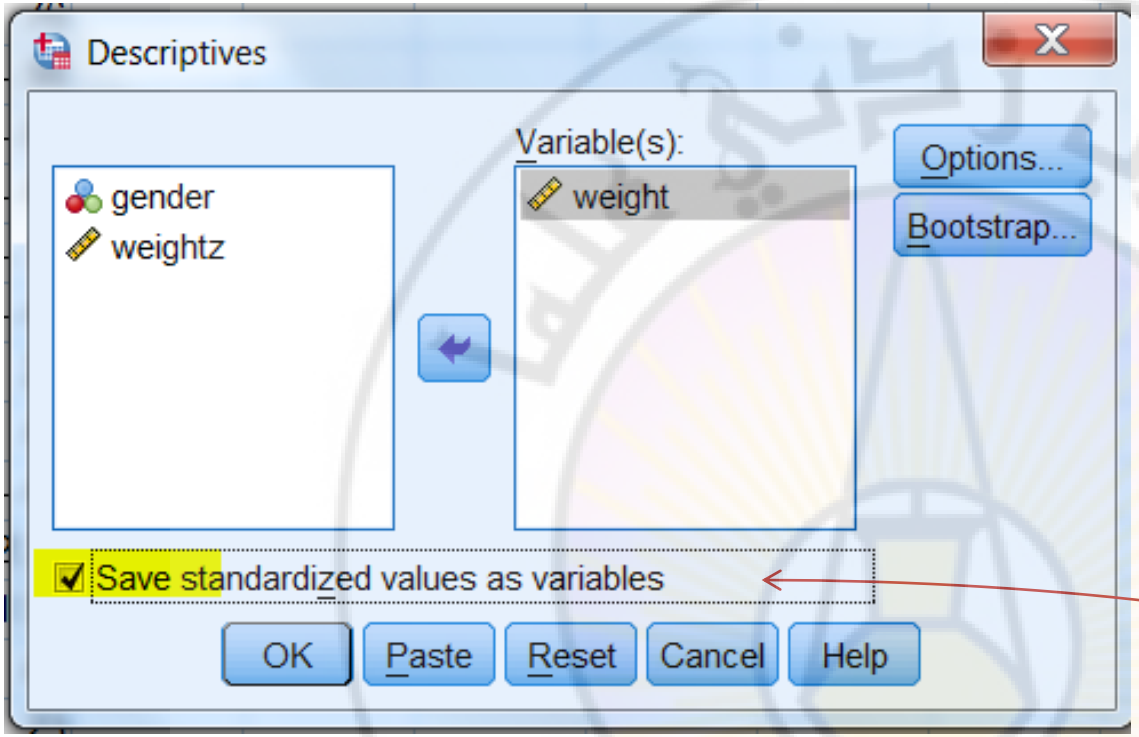
إنشاء المتحول: weightz=(weight-39.65)/5.76

$\bar{X} = 39.65 \ kg$

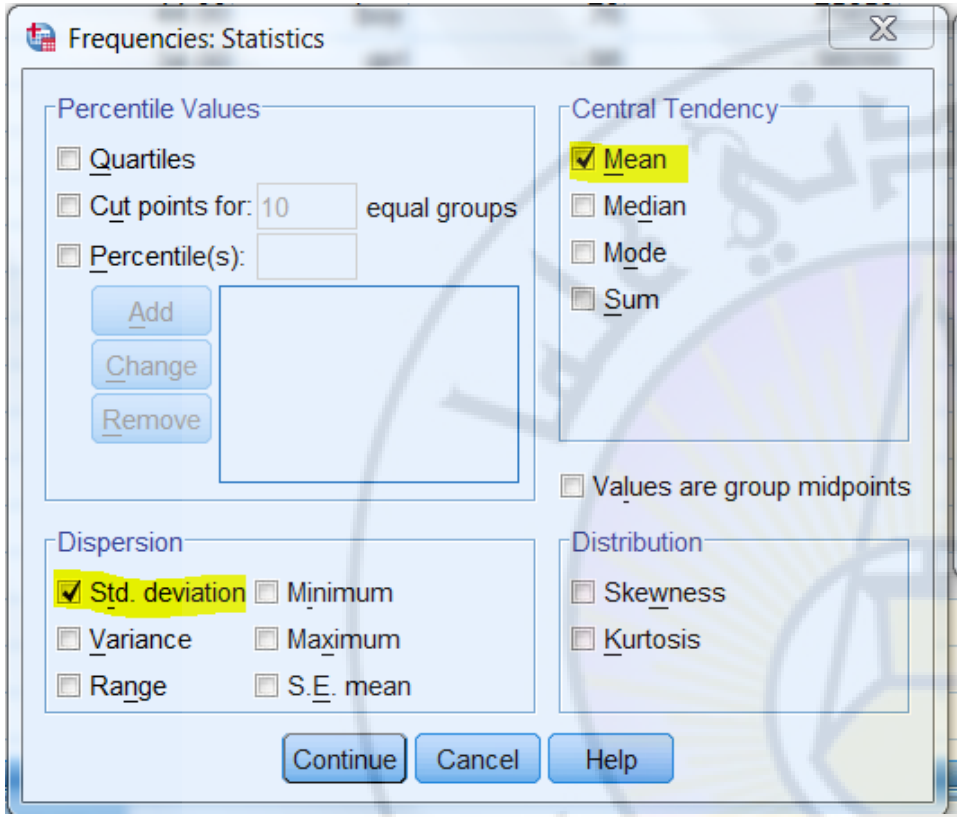
$s^2 = 33.124 \ kg^2$

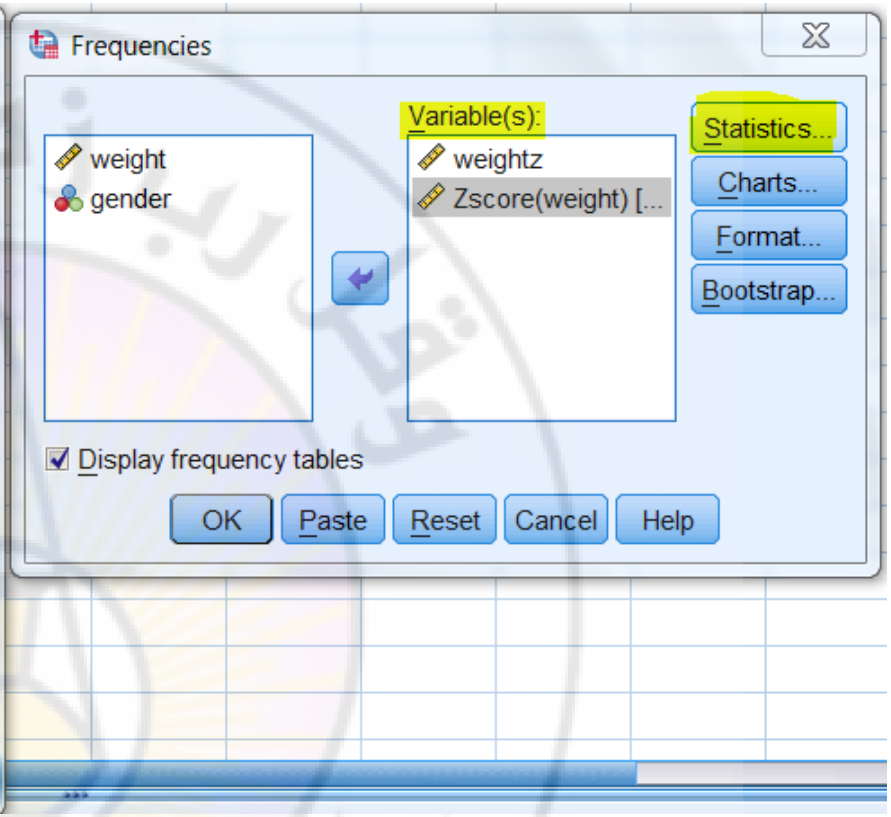
	weight	weightz
1	43.00	.58
2	44.00	.76
3	34.00	-.98
4	35.00	-.81
5	36.00	-.63
6	39.00	-.11
7	37.00	-.46
8	41.00	.23
9	42.00	.41
10	37.00	-.46
11	26.00	-2.37
12	31.00	-1.50
13	42.00	.41

في مثال أوزان الأطفال:



	weight	gender	weightz	Zweight
1	43.00	boy	.58	.58277
2	44.00	boy	.76	.75653
3	34.00	girl	-.98	-.98099
4	35.00	girl	-.81	-.80724
5	36.00	girl	-.63	-.63349
6	39.00	girl	-.11	-.11223
7	37.00	girl	-.46	-.45973
8	41.00	girl	.23	.23527
9	42.00	girl	.41	.40902
10	37.00	girl	-.46	-.45973
11	26.00	girl	-2.37	-2.37100





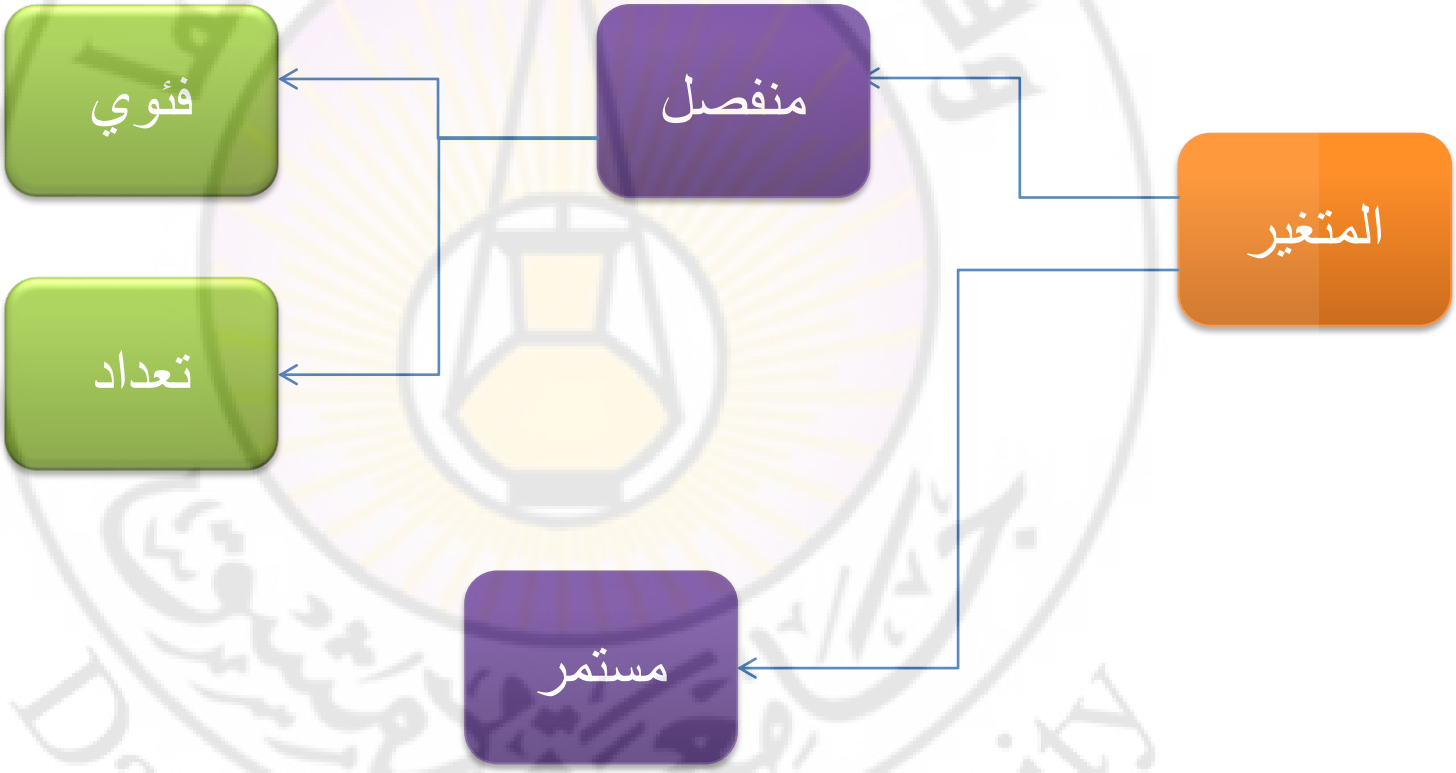
Statistics

		weightz	Zscore (weight)
N	Valid	209	209
	Missing	0	0
Mean		-.0007	0E-7
Std. Deviation		.99919	1.00000000

المتغيرات بلغة SPSS

~~Dispersion~~

~~CTM~~



البحث الخامس: المتغير العشوائي

مثال: X هو نتيجة رمي قطعة نقود مرة واحدة:

$$x_1 = H \ , \ x_2 = T$$

$$\mathbb{P}[X = X_1] = \frac{1}{2} \quad , \quad \mathbb{P}[X = X_2] = \frac{1}{2}$$

مثال: X هو نتيجة رمي حجر نرد مرة واحدة، عندئذ:

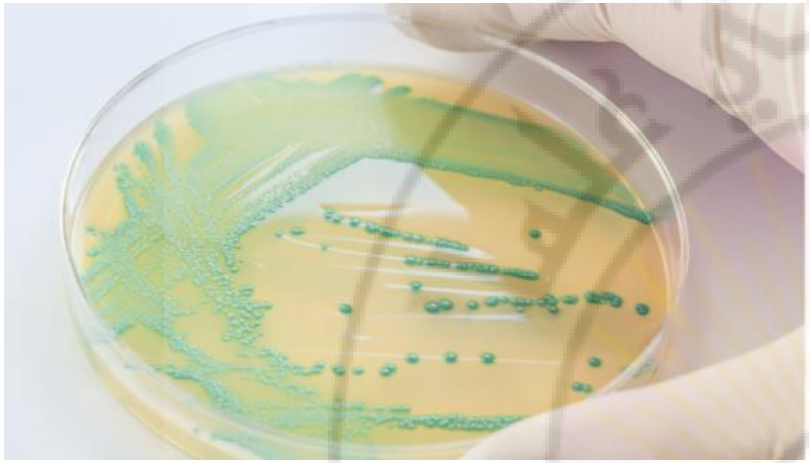
$$x_1 = 1, x_2 = 2, \ldots, x_6 = 6$$

مثال: X هو عدد السيارات المارة عبر طريق دولي خلال سنة، عندئذ:

$$x \in \{0, 1, 2, 3, \ldots, N, \ldots, +\infty\}$$

ولكل من هذه القيم فرصة معينة في الحدوث.

البحث الخامس: المتغير العشوائي



مثال: X هو عمر نوع من البكتريا (بالدقائق) بعد استخدام إحدى المعقمات، عندئذ:

$$0 < x_i < +\infty \quad ; \quad \forall i$$

ولكل من هذه القيم احتمال حدوث.

المتغير(المتحول) العشوائي:

هو خصيصة تدل على أن عناصر مجموعة تختلف فيما بينها صفةً حيث أنّ لكل عنصر من هذه العناصر فرصة للحدوث.

مثلاً المتغير الدال على الجنس: ذكر أوأنثى.

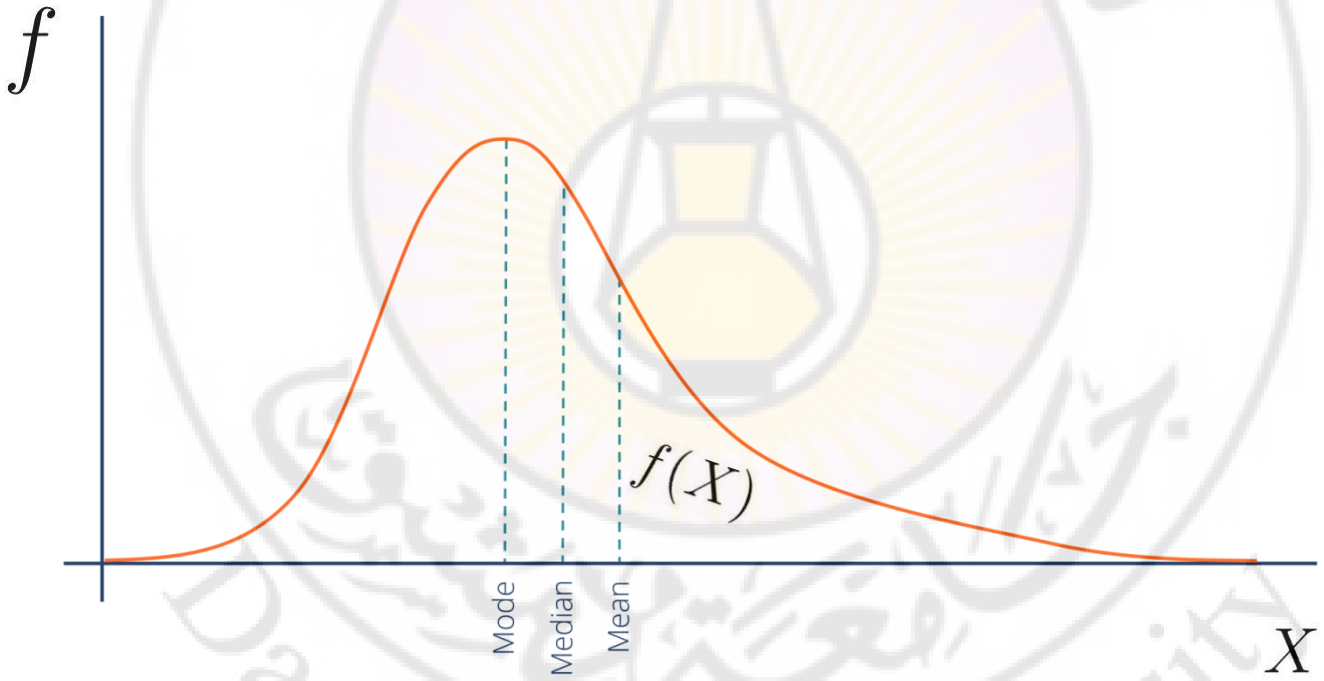
إنّ احتمال أن يكون الشخص ذكراً يساوي احتمال أن يكون الشخص أنثى (وهو يساوي النصف في مثالنا).

نستنتج من التعريف: أنّنا نعلم كل النتائج الممكنة للمتغير وفرصة وقوع كل منها، ولكن كون أنّنا لا نعلم النتيجة إلّا بعد إجراء التجربة، قلنا أنه متغير عشوائي.

البحث الخامس: المتغير العشوائي

اصطلاح:

X متغير عشوائي بالتالي ندعو معادلة المنحني التكراري بـ تابع الاحتمال Probability function ونرمز لها بـ f(x).



كل المتغيرات في SPSS تعتبر عينات عشوائية لـ متغيرات **عشوائية**.

البحث الخامس: التوزيع الطبيعي

تعريف:

نقول بأنّ X هو متغير عشوائي طبيعي بالمتوسط μ والتباين σ^2، إذا كانت معادلة منحنيه التكراري (تابعه الاحتمالي) هي بالشكل:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad ; \quad x\in(-\infty,+\infty)$$