محاضرات في مادة تحليل قواعد المعطيات I التسويقية I

GHICHI ALI

السنة الدراسية 2017 - 2018

أوامر وصف البيانات عبر برنامج (SPSS)

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)

لأجل إختبار التوزيع الطبيعي للبيانات، لدينا العديد من الوسائل يمكن أن نلخصها في الشكل البياني التالي:

إختبار التوزيع الطبيعي للبيانات
ينظر في الجانبين معا ولا تغني إحداهما على الأخرى على العموم

بالنظر لشكل التوزيع

- histograms
- box plots
- normal Q-Q plot
- detrended normal Q-Q plot

بالنظر للإختبار الإحصائي

- Kolmogorov-Smirnov (K-S) and Shapiro-Wilk

أوامر وصف البيانات عبر برنامج (SPSS)

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)

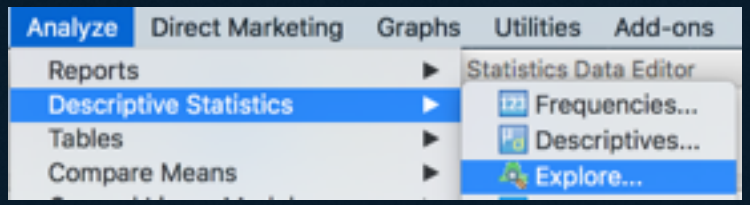
بعد تعرفنا على الإتجاهات الواجب سلوكها في تقييم مدى إتباع البيانات للتوزيع الطبيعي، نقدم المثال التالي لأجل تطبيق ما سبق تعريفه نظريا:

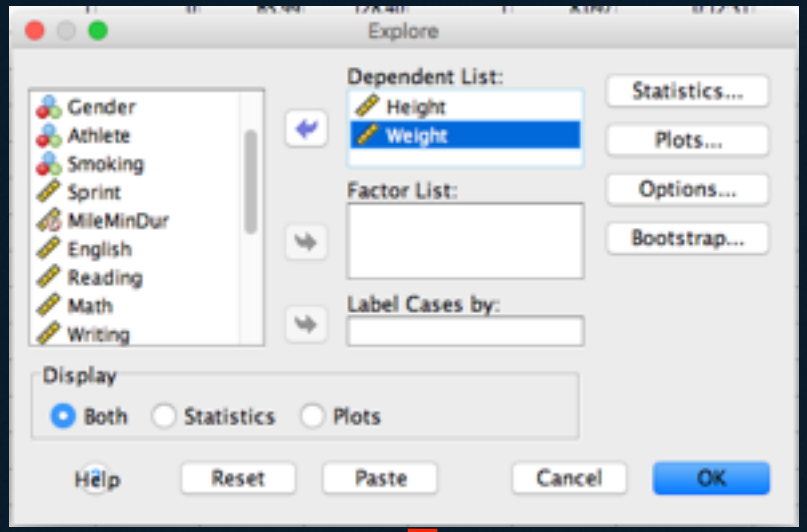
مثال: في العموم تخضع أطوال البشر إلى التوزيع الطبيعي، بينما تميل أوزانهم إلى الإلتواء، سنحاول من خلال بيانات الملف (Sample_Dataset_2014.sav) التأكد من اتباع هذه العينة لمنحنى التوزيع الطبيعي، وذلك من خلال القيام بالاختبار الإحصائي والرسومات البيانية لمتغيري الطول (heights) والوزن (weights) كما سبق وقدمنا في الأعلى.

سنستخدم الأمر (Explore) للقيام بهذا الإختبار الإحصائي.

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)

1. الأمر: (Explore) عند اختيار هذا الأمر يظهر مربع الحوار التالي:





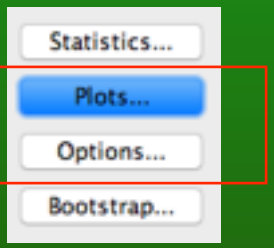
ثم ندخل المتغير المراد تحليله إلى(Dependent List)

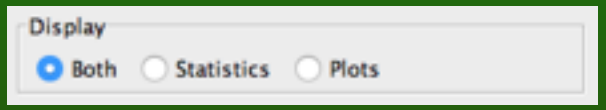
فمتغير تقسيم عرض المعلومات في (Factor List)

عبر الضغط دوما على السهم التالي:

مثلا في هذه الحالة تم إدخال المتغير (heights) والمتغير (weights)

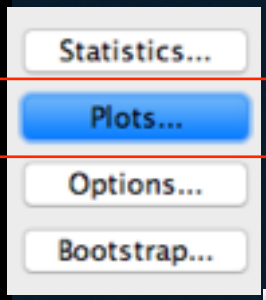
في هذه الحالة نترك (statistics) كما هي ونتعامل فقط مع (Plots) و (Options) لما لهم من علاقة مع رسوم التخطيط الطبيعي

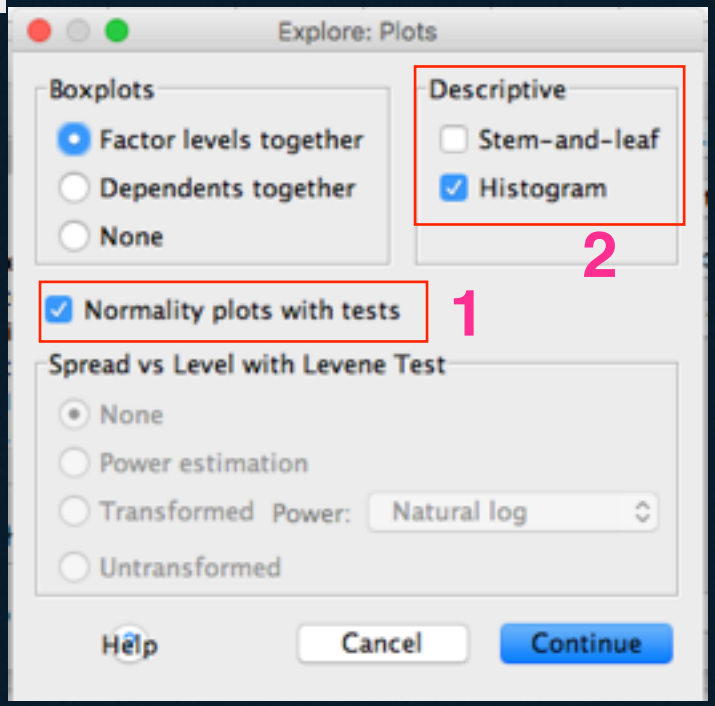


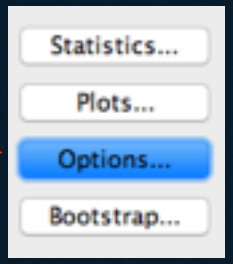


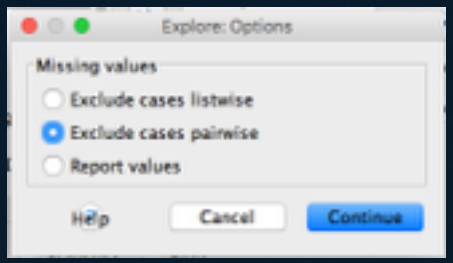
- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)

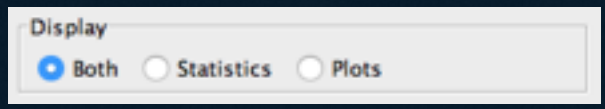
من خلال الضغط على إيقونة (Plots) نقوم بتحديد إختبار طبيعية البيانات من المربع (1)، كما نحدد المدرج التكراري في المربع (2)، ويتم ذلك بتأشير الباحث في الخانات الملائمة كما يظهر في مربع الحوار أسفله:











يتبع نترك التعيين في مكانه لعرض النتائج والرسوم البيانية معا

OK

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)

بالتطبيق نتحصل على النتائج التالي:

مختصر بيانات العينة (للمتغيرات المدروسة):

Explore

[DataSet1] /Users/arbah/Desktop/desk work univ/data base analysis/Sample_Dataset_2014.sav

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Height	408	93.8%	27	6.2%	435	100.0%
Weight	376	86.4%	59	13.6%	435	100.0%

يقدم لنا الجدول معلومات حول عدد مفردات عينة الإختبار، وكذا البيانات الناقصة في كل من متغير (heights) ومتغير (weights) حيث تتمثل هذه المعلومات في النسب المئوية وعدد مفردات العينة.

يتبع

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)

Descriptives

			Statistic	Std. Error
Height	Mean		68.0318	.26366
	95% Confidence Interval for Mean	Lower Bound	67.5135	
		Upper Bound	68.5501	
	5% Trimmed Mean		67.9687	
	Median		67.5700	
	Variance		28.363	
	Std. Deviation		5.32566	
	Minimum		55.00	
	Maximum		84.41	
	Range		29.41	
	Interquartile Range		6.78	
	Skewness		.230	.121
	Kurtosis		.113	.241
Weight	Mean		181.0316	2.20465
	95% Confidence Interval for Mean	Lower Bound	176.6966	
		Upper Bound	185.3666	
	5% Trimmed Mean		178.4763	
	Median		172.9600	
	Variance		1827.535	
	Std. Deviation		42.74968	
	Minimum		101.71	
	Maximum		350.07	
	Range		248.36	
	Interquartile Range		50.62	
	Skewness		1.005	.126
	Kurtosis		1.502	.251

النتائج الوصفية لتحليل البيانات:

في هذا الجدول نجد معلومات حول كل من المتوسط الحسابي، الحدود الدنيا والعليا لموقع المتوسط الحسابي حسب مجال الثقة المختار، في حالتنا هذه (95 بالمئة)، كما نجد أيضا معلومات حول التباين، الانحراف المعياري، المدى، الربيعيات وبيانات الإلتواء والتحدب

ملاحظة: القيم المعيارية لما يكون التوزيع طبيعي بالنسبة لقيم إحصائيات الإلتواء والتحدب هي (0). في حالتنا هذه قيمة الالتواء لمتغير الطول تساوي (0.23) وقيمة التحدب تساوي (0.113) وهي قيم تبتعد قليلا عن القيمة المعيارية، وعلى العكس من ذلك في متغير الوزن أين تبتعد كثيرا عن الصفر، هذه مؤشرات ابتدائية فقط .

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)

نتائج إختبار (Kolmogorov-Smirnov):

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Height	.045	408	.049	.993	408	.070
Weight	.084	376	.000	.944	376	.000

a. Lilliefors Significance Correction

ملاحظة: من أجل إختبار مدى طبيعية توزيع بيانات المتغيرات المعروضة على اختبار (K-S)، لابد أولا من إختيار التوزيع الطبيعي لأن هذا الاختبار صالح للعديد من أنواع التوزيعات كما سنبين فيما بعد، كما يجب طرح فرضية التوزيع بالطريقة الملائمة، حيث دوما ننطلق من أن البيانات تتوزع طبيعيا أي أن فرضية العدم تقترح أن البيانات تتوزع طبيعيا، على عكس باقي أنواع الفرضيات المنفية الأخرى.

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)

إختبار (Kolmogorov-Smirnov):

يعد أختبار (K-S) من الإختبارات اللامعلمية، تقنيا يمكنه اختبار مرجعية البيانات إلى العديد من التوزيعات المعروفة كتوزيع (Student, Pareto,...)، وليس فقط التوزيع الطبيعي الجرسي. إن الأمر (Explore) يتعامل مباشرة مع التوزيع الطبيعي، ولذا فمن غير المطلوب منك تحديد التوزيع المراد إختباره.

يدمج مع أختبار (K-S)، إختبار (Shapiro-Wilk) وهو إختبار معلمي، يعتمد نفس طريقة إختبار (K-S) من حيث طرح فرضية الاختبار على أن التوزيع هنا المراد إختباره يكون طبيعيا، وليكن في المعلوم أن إختبار (S-W) جد حساس لأبسط التغيرات بخصوص توزيع البيانات طبيعيا خصوصا في العينات ذات الأحجام الكبيرة. في كثير من الأحيان تتعارض نتائج الاختبارين وهما غير كافيتان للحكم على نوع التوزيع دون الإستعانة بالنظر إلى الأشكال البيانية المكملة، فعلى الباحث التعامل مع الاثنين بقدم المساوات.

طرح الفرضيات:

الفرضية الصفرية (H0): بيانات العينة مسحوبة من توزيع طبيعي

الفرضية البديلة (H1): بيانات العينة مسحوبة من توزيع غير طبيعي

يتبع

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)

الحكم على الفرضية: إن الحكم على الفرضية يتم من خلال قيمة (P-Value) التي يصدرها برنامج (SPSS) في جدول مخرجات الأمر (Explore):

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Height	.045	408	.049	.993	408	.070
Weight	.084	376	.000	.944	376	.000

a. Lilliefors Significance Correction

- فإذا كانت قيمة (P-Value) أقل من مستوى المعنوية غالبا (0.05) عند مجال ثقة بـ (95 بالمئة)، حينها نرفض (H0)، ونقبل الفرضية البديلة (H1)، وبالتالي فإن التوزيع الذي تم سحب بيانات العينة منه غير طبيعي.
- أما إذا كانت قيمة (P-Value) أكبر من مستوى المعنوية غالبا (0.05) عند مجال ثقة بـ (95 بالمئة)، وبالتالي لا نرفض فرضية العدم (H0)، ونقول حينها أنه لا توجد دلائل بالشكل الكافي للتوصل إلى إستنتاج أن البيانات غير طبيعية.

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Height	.045	408	.049	.993	408	.070
Weight	.084	376	.000	.944	376	.000

a. Lilliefors Significance Correction

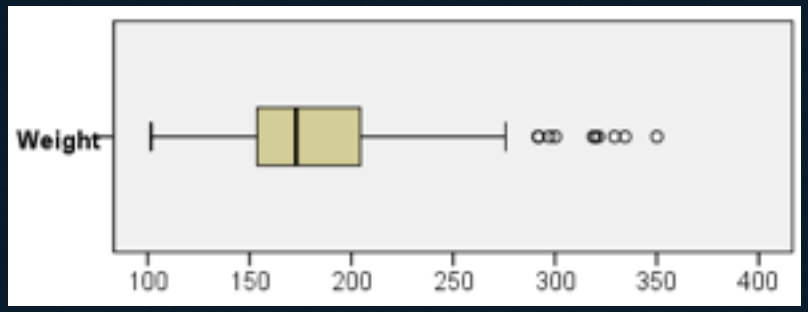
التعليق على نتائج الإختبار :

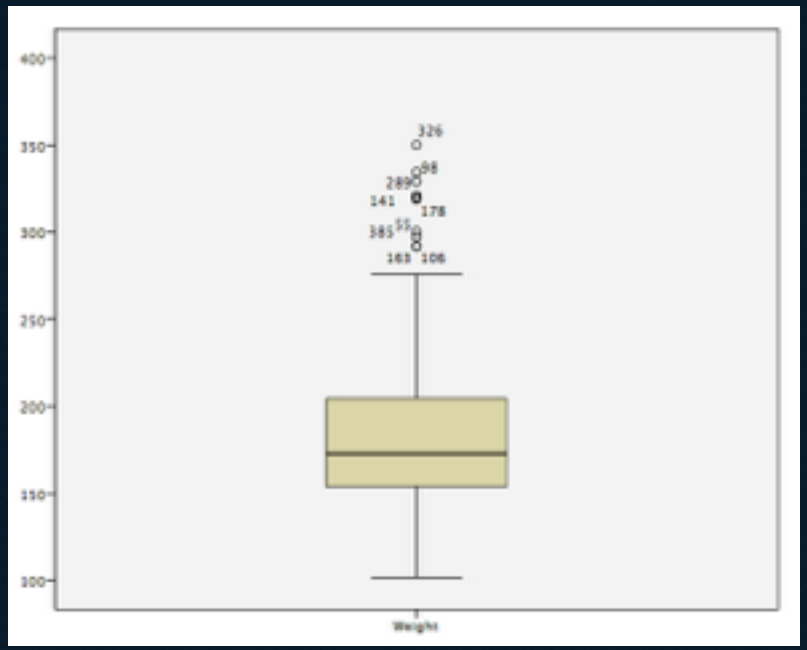
من خلال نتائج الجدول نلاحظ أن قيم (K-S and Shapiro-Wilk test p-values) بالنسبة للوزن هي أقل من (p < 0.001)، ومنه فرفض فرضية العدم أمر واضح، أما بالنسبة لمتغير الطول، فتبين قيمة (p-values) لاختبار (K-S) عن قيمة (0.049) وهي تكاد تقترب من مستوى المعنوية (0.05)، وعلى العكس من نتيجة اختبار (K-S)، يظهر إختبار (Shapiro-Wilk) عن قيمة (0.070) وهي قيمة أكبر من مستوى المعنوية (0.05)، يتوصل الأختباران إلى نتائج متضادة، أين يبين إختبار (K-S) أن البيانات غير طبيعية، بينما يرى إختبار (Shapiro-Wilk) أن البيانات طبيعية.

فكيف نتعامل مع هذا التناقض؟ هنا وجب النظر في مخرجات الرسوم البيانية الأخرى.

يتبع

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)





أولا: النظر إلى رسمة الصندوق لمتغير الوزن

من رسمة الصندوق لمتغير الوزن في الشكل المقابل نفهم لما كانت نتائج اختبار طبيعة البيانات سلبية، فالملاحظ يرى أن توزيع البيانات هنا ملتوي نحو اليمين، ويظهر ذلك من خلال ما يلي:

1. طول جناح الصندوق الأيمن على الأيسر
2. الوسيط يميل إلى الجانب الأيسر منه إلى الأيمن
3. النقاط الشادة في الجهة العليا من نهاية التوزيع

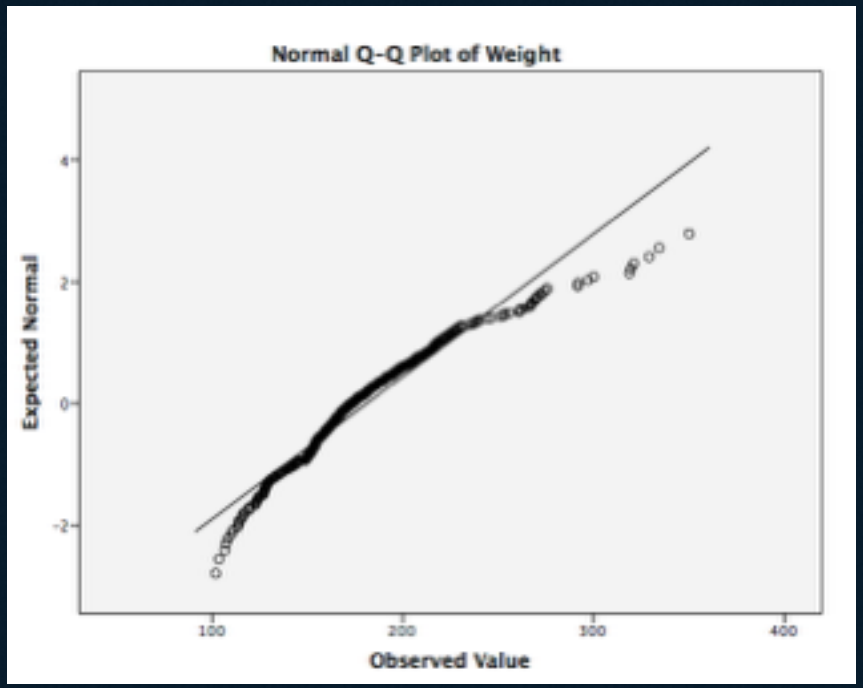
بعدها يمكننا النظر في كل من الرسم البياني لـ:

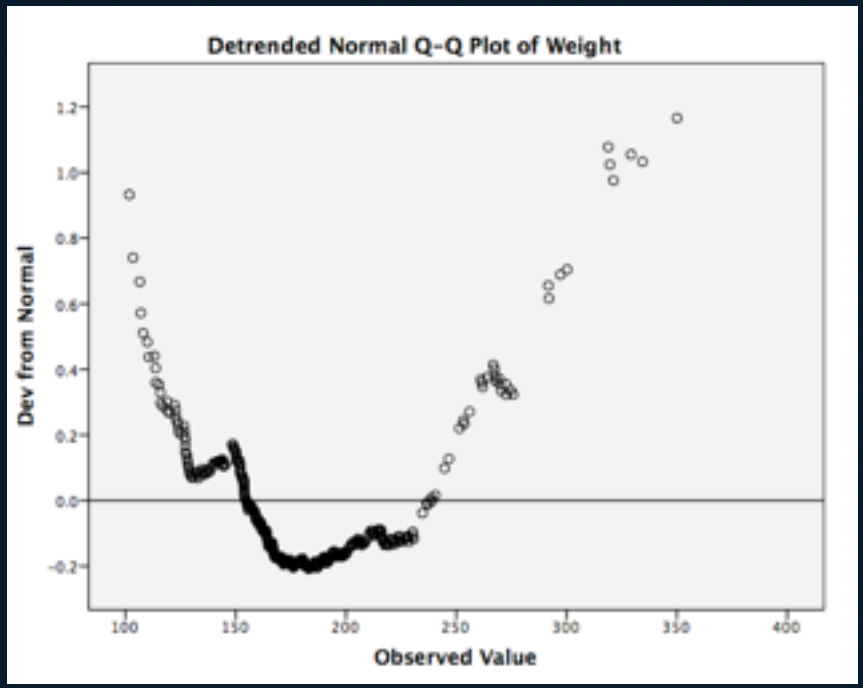
- رسم خط انتشار (Normal Q-Q plots)

ورسم خط انتشار (Detrended Q-Q plots)

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)

- رسم خط انتشار (Q–Q and detrended Q–Q plots) لمتغير الوزن

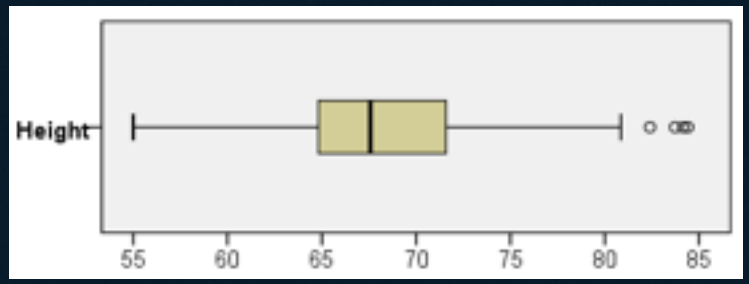


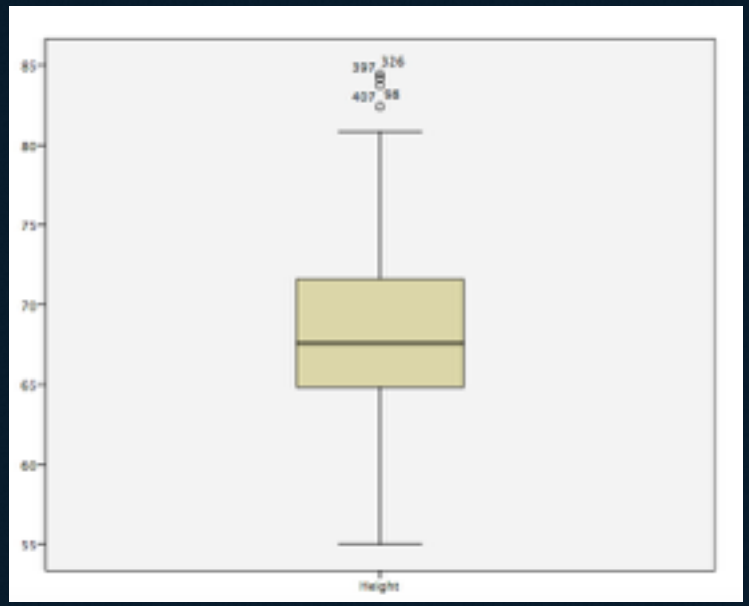


من الشكلين في الأعلى لكل من (Q–Q and detrended Q–Q plots) نلاحظ إنحرافات منتظمة عن خط التوزيع الطبيعي، ويمكن ملاحظة أن شكل توزيع بيانات (detrended Q–Q) مقعر (برابولي)، حيث تتغير إنحرافات التوزيع على محور العينات (y–axis) بمقدار يتراوح بين (0.2– حتى 1.2).

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)

ثانيا: لننظر إلى رسمة الصندوق الطول



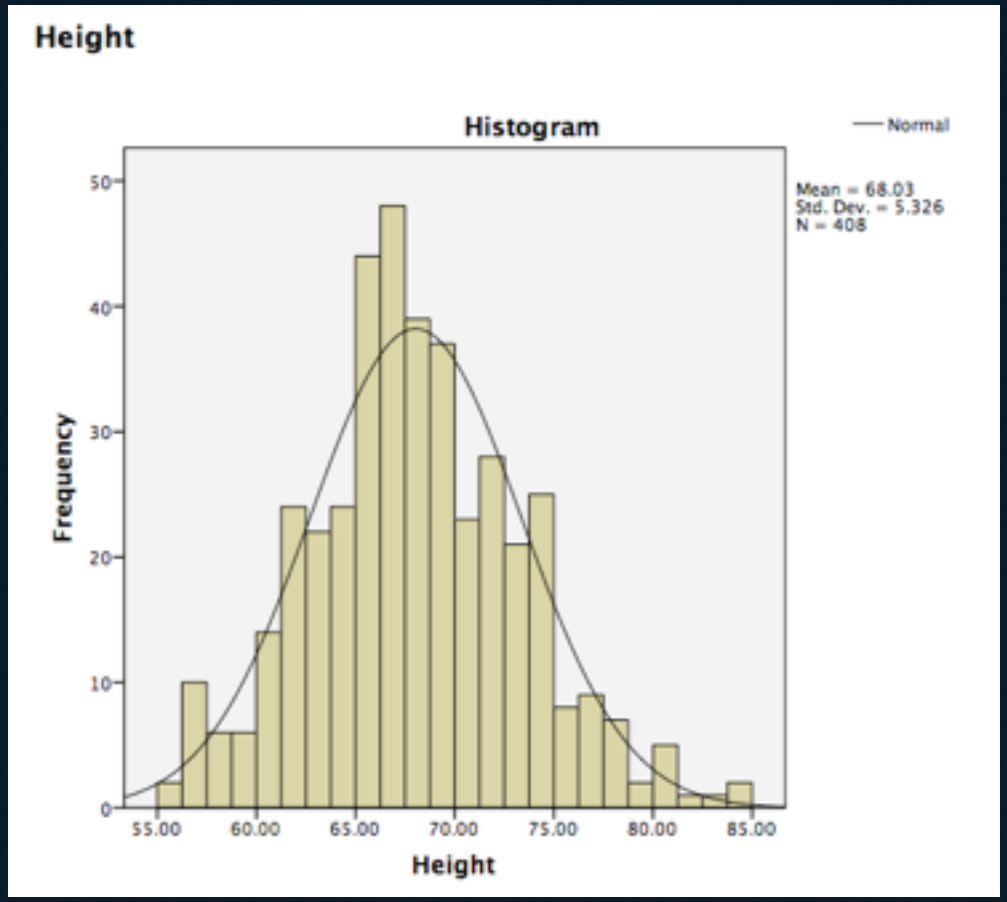


من رسمة الصندوق لمتغير الطول في الشكل المقابل نلاحظ أن توزيع البيانات هنا ملتوي نحو اليمين بشيء يسير، ويظهر ذلك من إقتراب الوسيط من متصف الصندوق، والنتيجه ليست مثلها في متغير الوزن، فالتوزيع في حالة الطول يظهر أقرب لأن يكون منتظم نحو المركز منه في حالة متغير الوزن.

ولأجل التاكد أكثر لابد من النظر في باقي الرسوم البيانية الأخرى ومنها (Histogram) وأيضا بالطبع مخرجات الأمر (Explore) المتمثلة في كل من (Normal Q-Q plots) و(detrended Q-Q plots).

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)

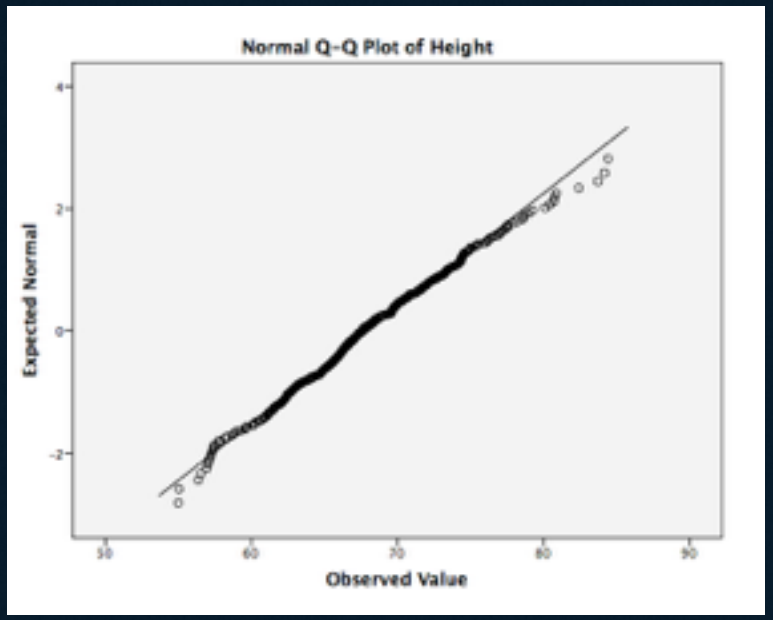
- المدرج التكراري (Histogram) لمتغير الطول

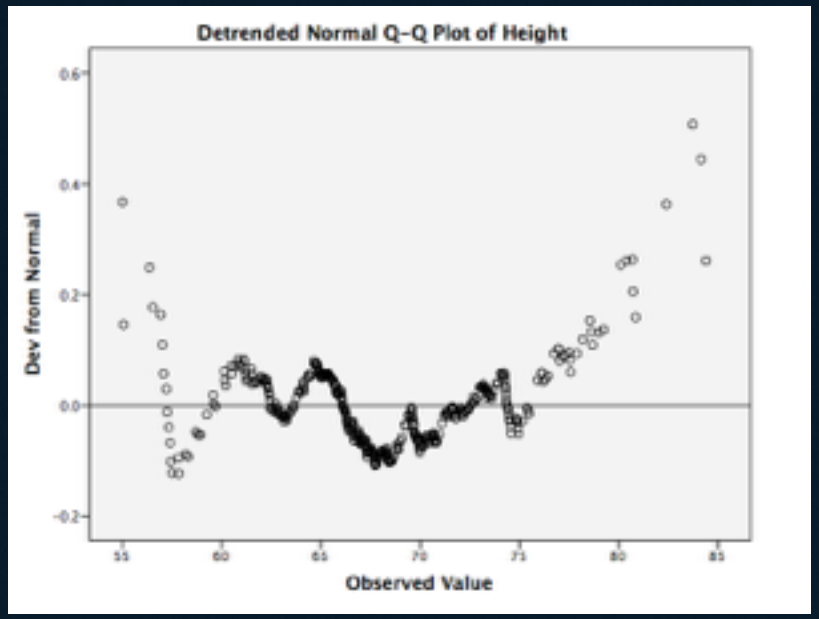


إن الهدف من النظر إلى المدرج التكراري هو معرفة مدى تماثل الشكل الجرسي، وليس بالضرورة التاكد من مصداقية التوزيع، لكنه إشارة فقط شكلية لمدى تطابق المنحنى البياني

مع مخرجات البيانات والشكل الجرسي المعروف للتوزيع الطبيعي، كما يمكن معرفة مدى تمركز مؤشرات النزعة المركزية حول وسط المنحنى الجرسي والمدرج التكراري على حد السواء، فكلما كان الشكل معتدل ومتطابق على الجانبين وتوسطه احصائيات النزعة المركزية كلما كان ميل الباحث للحكم على أن البيانات تخضع للتوزيع الطبيعي.

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)
  - رسم خط انتشار (Q-Q and detrended Q-Q plots) لمتغير الطول





من الشكلين في الأعلى لكل من (normal Q-Q plot) نلاحظ أن توزيع النقاط ينتشر بالتطابق على خط التوزيع الطبيعي، والانحرافات طفيفة ولا تظهر إلا على الديلين فقط. أما في شكل توزيع بيانات (detrended Q-Q) فيمكن ملاحظة الإنحرافات عن الخط الطبيعي بوضوح إلا أنها ليست كبيرة مثل ما تم مشاهدته في حالة متغير الوزن، وقد كانت بين (-0.2 و0.6) وهي قيم معقولة خصوصا أنها قيم شادة.

- الأوامر المختلفة لاختبار مدى خضوع البيانات للتوزيع الطبيعي (Normality Test)

نتائج وتوصيات:

من خلال الإلتواء في صندوق العرض، والنحرافات المنتظمة في (Q-Q plots)، ونتائج الإختبارين أين تم تسجيل قيمة أقل بكثير عن مستوى المعنوية ($p < 0.001$)، تبين العديد من الدلائل على أن بيانات متغير الوزن لا تخضع للتوزيع الطبيعي.

بالنسبة لمتغير الطول، تظهر الدلائل ضعيفة، حيث أن أحد الإختبارات كان أقل (K-S test p = 0.049) بينما الآخر لا (Shapiro-Wilk p = 0.070). وبعد النظر في الرسوم البيانية، وجدنا أن هناك بعض الإنحرافات عن الخط الطبيعي، ولكن هذه الانحرافات لم تكن تظهر بالشكل الكبير، ولأجل العمل التطبيقي، سيكون من غير اللائق إعتبار توزيع بيانات متغير الطول غير طبيعية.

محاضرات في مادة تحليل قواعد المعطيات I التسويقية I