الفصل الثاني عشر

تحليل التباين ANOVA

1.3 مقدمة:

ذكرنا عند الحديث عن اختيار الفرق بين وسطين حسابيين أن الباحث قد يرغب في اختبار ما إذا كان متوسط الدخل في إحدى الدول يساوي متوسط الدخل في دولة أخرى. أو إجراء اختبار عما إذا كان متوسط عمر المواطن في إحدى المناطق يساوي متوسط عمر الناخب في منطقة أخرى. أي أن الباحث قد يرغب في إجراء اختبار ما إذا كان متوسط مجتمع يساوي متوسط مجتمع آخر. وتكون الفرضية العدمية في هذه الحالات:

$$Ho: \mu_1 = \mu_2$$

ولكن في كثير من الحالات قد يرغب الباحث في إجراء اختبار عن تساوي ثلاث متوسطات أو أكثر. فقد يرغب في اختبار ما إذا كانت متوسطات الدخول في أربع دول متساوية أم لا، أو ما إذا كانت متوسطات أعمار المواطنين في ست مناطق متساوية أم أن هناك اختلافات بينها. ففي المثال الأول تكون الفرضية العدمية:

$$Ho: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

مقابل الفرضية البديلة بعدم تساوي بعض هذه المتوسطات (على الأقل متوسط غير متساوي) وفي المثال الثاني تكون الفرضية العدمية :

$$Ho: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

في مثل هذه الحالات لا نأخذ كل اثنين من المتوسطات على حدة ونجري اختبار الفرق بين وسطين لهما لأن هذا سوف يستغرق وقتاً أطول ومجهوداً أكبر، والأهم من ذلك أن احتمال أخذ قرار صحيح سوف يقل أو ينخفض، ويزيد بالتالي – كثيراً – احتمال الخطأ أو احتمال اتخاذ قرار غير سليم، ولتوضيح ذلك نأخذ المثال الأول الخاص بمقارنة المتوسط لأربع دول :فإذا أخذنا كل دولتين على حده فإنه يلزم إجراء هذا الاختبار ست مرات، فإذا كان مستوى المعنوية المستخدم في كل اختبار هو 0.05 (أو أن درجة الثقة هي 0.95) فإن احتمال الوصول إلى القرار الصحيح للاختبارات الستة معاً يساوي $(0.95)^6$

أي 0.95 مضروبة في نفسها ست مرات أي يساوي 0.745 ومعنى ذلك أن احتمال اتخاذ القرار الصحيح سوف ينخفض من 0.95 إلى 0.735 .فقط وبمعنى آخر فإن احتمال الخطأ في اتخاذ القرار الصحيح سوف يرتفع من مجرد 0.05 فقط إلى 0.265 والذي يساوي (0.735 - 1) وهو احتمال كبير للخطأ عند اتخاذ القرار.

وكلما زاد عدد المتوسطات كلما زاد احتمال الخطأ وقل احتمال اتخاذ قرار صحيح ففي المثال الثاني الخاص بمقارنة متوسطات أعمار الناخبين في ست مناطق فإنه يلزم إجراء الاختبار 15 مرة وبالتالي سوف ينخفض احتمال اتخاذ قرار صحيح في الخمسة عشر اختبار معاً من 0.95 إلى $(0.95)^{15}$ أي إلى 0.46 فقط وبالتالي يرتفع احتمال الخطأ في اتخاذ القرار الصحيح من مجرد 0.05 إلى 0.54 والذي يساوي 1-0.46) وهو احتمال كبير جداً للخطأ في اتخاذ القرار.

لذلك كان لابد من التفكير في أسلوب آخر بديل يوفر الوقت والمجهود وفي الوقت نفسه لا يقلل احتمال اتخاذ القرار الصحيح أو يكبر احتمال الخطأ في اتخاذ القرار، هذا الأسلوب هو الذي يسمى " تحليل التباين " والذي يختبر ما إذا كانت المتوسطات كلها متساوية مرة واحدة دون أخذهم اثنين اثنين ودون أن ينخفض احتمال اتخاذ قرار صحيح أو يزيد احتمال الخطأ عند اتخاذه .والذي يسمى اختصاراً ANOVA وهو اختصار للمصطلح الإنجليزي **Analysis of Variance**.

ويعتمد هذا الأسلوب من أساليب التحليل الإحصائي على ما يعرف باختبار F والذي يعتمد أساساً على تحليل التباين. ونعلم من الفصول الأولى من الكتاب أن التباين ما هو إلا متوسط مربعات انحرافات القيم عن وسطها الحسابي. أي أن التباين يعتمد أساساً على مجموع مربعات ثم القسمة على عدد المشاهدات. ويعتمد أسلوب تحليل التباين على تقسيم مجموع المربعات الكلي إلى أقسام فيمثل كل منهما أو يقيس أحد مصادر التغير أو الاختلاف **Source of Variation** يمثل أحدها – مثلاً – التغير بسبب المعاملات (أو المجتمعات) المختلفة، ويمثل آخر التغير بسبب الأخطاء ثم تعرف الإحصائية (أو الاختبار F) بأنها خارج قسمة التباين بسبب المعاملات على التباين بسبب الأخطاء. وهكذا. أي أنه يتم حساب التباين بسبب المعاملات، والتباين بسبب الأخطاء فيحصل على

قيمة F المحسوبة وبمقارنة هذه القيمة بالقيمة الجدولية F نصل إلى قرار إما بقبول الفرض العدمي أو عدم قبوله عند مستوى المعنوية المطلوب. ولتحليل التباين تطبيقات كثيرة في مختلف المجالات: وسوف نتناول هنا أبسط حالة لتحليل التباين وهي التي تسمى التصنيف الأحادي **One -Way Classification** مع العلم بأن هناك حالات أخرى كثيرة لتحليل التباين منها على سبيل المثال التصنيف الأحادي في حالة اختلاف أحجام العينات وتحليل التباين الثنائي **Two-way Analysis** ولكننا سوف نتطرق لأبسط حالة **وهي حالة التصنيف الأحادي بافتراض تساوي وعدم تساوي أحجام العينات.**

2.3 التصنيف الأحادي (في حالة تساوي أحجام العينات):

يعد هذا التصنيف هو أبسط أنواع تحليل التباين، حيث تصنف المشاهدات إلى عدة مجموعات على أساس متغير واحد أو خاصية واحدة. وان الافتراضات الأساسية لهذا التحليل هي ما يلي:

1- نفترض أن عدد المجتمعات(المجموعات) المكونة للمتغير **a** وأنها جميعاً مستقلة.

2- أنها جميعاً تخضع للتوزيع الطبيعي بمتوسطات تساوي $\mu_k, ..., \mu_2, \mu_1$

3- أن لها جميعاً التباين نفسه σ^2 أي أن التباين لكل المجتمعات ثابت ويساوي σ^2

4- يتم سحب عينة عشوائية من كل من هذه المجتمعات وأن أحجام هذه العينات (n) كلها متساوية والتي تمثل عدد المشاهدات وتصنف البيانات عادة في هذا التحليل على النحو التالي:

5- ان y_{ij} تمثل المشاهدة (j) في العينة المسحوبة او المجموعة أو الفئة (i)

المجموعات او الفئات	المشاهدات		
	1	2	n
1	y_{11}	y_{12}	y_{1n}
2	y_{21}	y_{22}	y_{2n}
:	:	:	:
a	y_{a1}	y_{a2}	y_{an}

حيث يمثل الصف الأول مشاهدات العينة الأولى أي المسحوبة من المجتمع الأول، ويمثل الصف الثاني مشاهدات العينة المسحوبة من المجتمع الثاني، وهكذا يمثل الصف الأخير مشاهدات العينة المسحوبة من المجتمع الأخير رقم **a**. **وكما تكون خطوات الاختبار كما يلي:**

1-الفرضية العدمية: أن متوسطات هذه المجتمعات متساوية. وبالرموز

$$Ho: \mu_1 = \mu_2 = ... = \mu_k$$

2-الفرضية البديلة: أن بعض هذه المتوسطات غير متساوية (أو يوجد متوسط على الأقل غير متساوي).

$$H1: \mu_1 \neq \mu_2 \neq .... \neq \mu_k$$

3-إحصائية الاختبار: في هذه الحالة يرمز لها بالرمز F وتأخذ الشكل التالي:

$$F = \frac{MSR}{MSE}$$

والتي لها توزيع F بدرجات حرية للبسط a-1 وللمقام a(n-1)

ويمكن الحصول على الإحصائية F بتنظيم الحسابات في جدول يسمى " جدول تحليل التباين " AVOVA TABLE كما يلي:

جدول تحليل التباين (ANOVA)*

(الإحصائية) المحسوبة	متوسط المربعات	درجات الحرية	مجموع المربعات	مصدر التغير
$F = \frac{MSR}{MSE}$	$MSR = \frac{SSR}{a-1}$	a-1	SSR=[A]-[C]	بسبب المعاملات
	$MSE = \frac{SSE}{a(n-1)}$	a(n-1)	SSE=[B]-[A]	بسبب الخطأ
		an-1	SST=[B]-[C]	الكلي

م.سرمد علوان صالح
أ.م.د. وقاص سعد خلف

وسوف نوضح من المثال كيفية حساب المقادير الثلاثة SSE , SSR , SST

4-حدود منطقتي القبول والرفض: ويتم الحصول عليها من جدول توزيع F بدرجات حرية للبسط a-1 وللمقام (a(n- 1)). (اختبار الطرف الأيمن).

5-المقارنة والقرار: إذا كانت قيمة F المحسوبة من تحليل التباين أقل من قيمة F الجدولية نقبل الفرض العدمي بتساوي المتوسطات والعكس صحيح.

مثال(1): البيانات التالية تمثل أعمار أربع عينات عشوائية من المواطنين سحبت من أربع مدن مستقلة (نفترض أن لها توزيعات طبيعية بمتوسطات $\mu_1, \mu_2, \mu_3, \mu_4$ وتباين مشترك يساوي σ^2) :

المدن a	المشاهدات						المجموع
	1	2	3	4	5	6	
الأولى 1	20	21	25	28	30	26	150
الثانية 2	23	22	27	20	26	20	138
الثالثة 3	19	20	21	28	20	18	126
الرابعة 4	24	29	30	28	27	24	162

والمطلوب اختبار الفرض العدمي بأن متوسطات أعمار المواطنين من المدن الأربع متساوية وذلك بمستوى معنوية 5%، أي أن المطلوب بالرموز:

$$Ho: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

الحل: تكون خطوات الحل كما يلي:

1-الفرضية العدمية:

$$Ho: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

2-الفرضية البديلة: أن بعض هذه المتوسطات غير متساوٍ (على ال أقل متوسط غير متساوي).

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

3-الإحصائية: وهي في هذه الحالة $F = \dfrac{MSR}{MSE}$

وتكون الحسابات التفصيلية لتحليل التباين كما يلي: إذ أن **4** =a **وأن 6 =n**

$$[A] = \sum_{i=1}^{i=4} \frac{y_i}{n_i} = \frac{(150)^2}{6} + \frac{(138)^2}{6} + \frac{(1126)^2}{6} + \frac{(162)^2}{6} = 13944$$

$$[B] = \sum_{j=1}^{j=6}\sum_{i=1}^{i=4} y^2 ij = (20^2 + 21^2 + \dots\dots\dots\dots + 24^2) = 14160$$

$$[C] = \frac{\left[\sum_{j=1}^{j=6}\sum_{i=1}^{i=4} yij\right]^2}{N = a * n = \sum_{i=1}^{i=4} n_i = n_1 + n_2 + n_3 + n_4} = \frac{(20 + 21 + \dots\dots\dots\dots + 24)^2}{6+6+6+6 = 6*4} = 13824$$

then

$$SSR = [A] - [C] = 13944 - 13824 = 120$$

$$SSE = [B] - [A] = 14160 - 13944 = 216$$

$$SST = [B] - [C] = 14160 - 13824 = 336$$

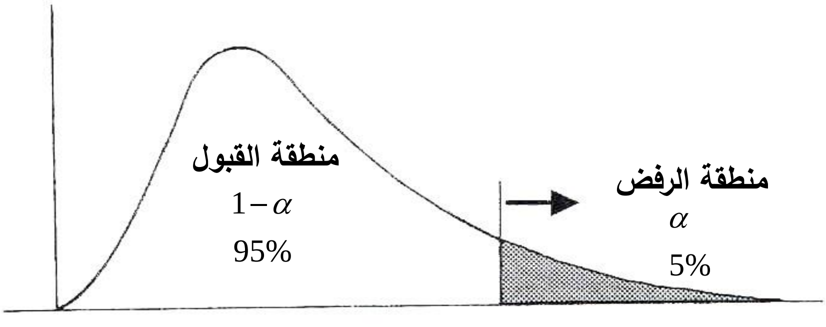
ومن ثم نكون جدول تحليل التباين كما يلي:

مصدر التغير	مجموع المربعات	درجات الحرية	متوسط المربعات	الإحصائية F
بسبب المعاملات	SSR= 120	3	MSR = 120/ 3 = 40	$F = \dfrac{40}{10.8} = 3.7$
بسبب الخطأ	SSE= 216	20	MSE= 216/20= 10.8	
الكلي	SST= 336	23		

أي أن قيمة الإحصائية (أو **F** المحسوبة) هي 3.7

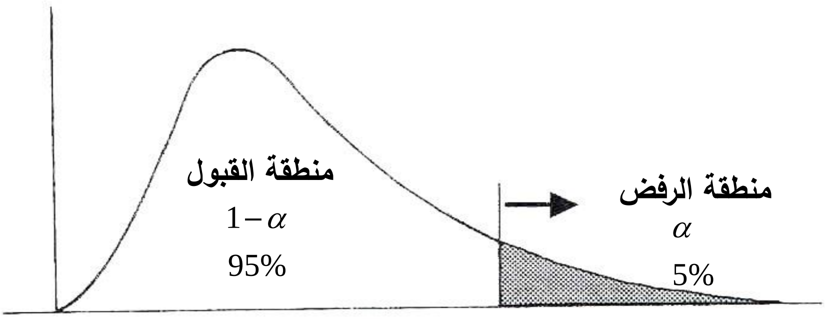
4-حدود منطقتي القبول والرفض: من جدول توزيع **F** وعند مستوى معنوية **<u>5%</u> وبدرجات حرية <u>3</u> للبسط، <u>20</u> للمقام** نجد أن **F** الجدولية تساوي **<u>3.10</u>**

ويمكن توضيح ذلك بالرسم كما يلي:



5-المقارنة والقرار: وحيث أن قيمة الإحصائية المسحوبة والتي تساوي **3.7** أكبر من القيمة الجدولية فإنها تقع في منطقة الرفض وبالتالي فإن القرار هو رفض الفرضية العدمية بتساوي متوسطات أعمار المواطنين في المدن الأربع وذلك بمستوى معنوية **5%**

أما إذا استخدمنا مستوى معنوية 1% فإن قيمة F من الجدول تصبح 4.94 أي تصبح حدود منطقتي القبول والرفض كما يلي:



في هذه الحالة فإن قيمة الإحصائية والتي تساوي 3.7 تقع في منطقة القبول، وبالتالي فإن القرار يكون قبول الفرض العدمية بتساوي متوسطات أعمار المواطنين في المدن الأربع وذلك بمستوى معنوية 1%.

3.3 التصنيف الأحادي (في حالة عدم تساوي أحجام العينات) :

يعد هذا التصنيف هو أبسط أنواع تحليل التباين، حيث تصنف المشاهدات إلى عدة مجموعات على أساس متغير واحد أو خاصية واحدة. وان الافتراضات الأساسية لهذا التحليل هي ما يلي:

1-نفترض أن عدد المجتمعات **a** وأنها جميعاً مستقلة.

2-أنها جميعاً تخضع للتوزيع الطبيعي بمتوسطات تساوي $\mu_1, \mu_2, ..., \mu_k$

3-أن لها جميعاً التباين نفسه σ^2 أي أن التباين لكل المجتمعات ثابت ويساوي σ^2

4-يتم سحب عينة عشوائية من كل من هذه المجتمعات وأن أحجام هذه العينات كلها غير متساوية ولن يختلف أسلوب التحليل على الإطلاق إلا في أشياء بسيطة جداً.

مثال(2): الجدول الاتي يمثل عدد الوحدات المعيبة لمنتج معين من ثلاث انواع من المكائن **(M_1,M_2,M_3)** وكما مبين ادناه

المكائن **a**	عدد الوحدات المعيبة							**المجموع**
	1	2	3	4	5	6	7	
M_1	70	83	87	78	…..	…..	…..	318
M_2	64	45	56	50	71	…..	…..	286
M_3	48	94	83	84	80	87	90	566

هل هناك فرقاً بين الوحدات المعيبة المنتجة من قبل المكائن الثلاثة عند مستوى معنوية 0.05

الحل: تكون خطوات الحل كما يلي:

1-الفرضية العدمية: لا توجد فروق جوهرية بين الوحدات المعيبة المنتجة من قبل المكائن الثلاثة

$$Ho: \mu_1 = \mu_2 = \mu_3$$

2-الفرضية البديلة: توجد فروق جوهرية بين الوحدات المعيبة المنتجة من قبل المكائن الثلاثة

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3$$

3-الإحصائية: وهي في هذه الحالة $F = \frac{MSR}{MSE}$

وتكون الحسابات التفصيلية لتحليل التباين كما يلي: إذ أن **a= 4** وأن $n_1=4, n_2=5, n_3=7$

$$[A]=\sum_{i=1}^{i=3}\frac{y_i}{n_i}=\frac{(318)^2}{4}+\frac{(286)^2}{5}+\frac{(566)^2}{7}=87405.342$$

$$[B]=\sum_{j}\sum_{i=1}^{i=3}y^2 ij=(70^2+83^2+........................+90^2)=89394$$

$$[C]=\frac{\left[\sum_{j}\sum_{i=1}^{i=4}yij\right]^2}{N=\sum_{i=1}^{i=3}n_i=n_1+n_2+n_3+n_4}=\frac{(70+83+........................+90)^2}{5+5+6}=85556.25$$

then

$$\text{SSR}=[A]-[C]=87405.342-85556.25=1849.092$$

$$\text{SSE}=[B]-[A]=89394-87405.342=1988.658$$

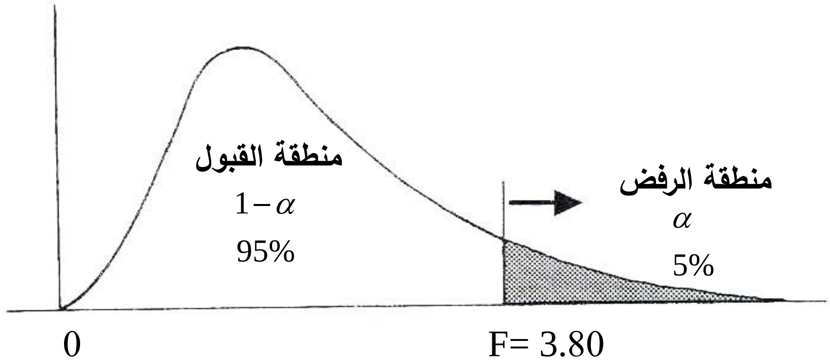
$$\text{SST}=[B]-[C]=89394-85556.25=3837.75$$

ومن ثم نكون جدول تحليل التباين كما يلي

مصدر التغير	مجموع المربعات	درجات الحرية	متوسط المربعات	الإحصائية F
بسبب المعاملات	SSR= 1849.092	2	MSR = 924.546	$F=6.04$
بسبب الخطأ	SSE= 1988.658	13	MSE= 152.973	
الكلي	SST= 3838.75	15		

أي أن قيمة الإحصائية (أو **F** المحسوبة) هي 6.04

4-حدود منطقتي القبول والرفض: من جدول توزيع **F** وعند مستوى معنوية **<u>5%</u> وبدرجات حرية <u>2</u> للبسط، <u>13</u> للمقام** نجد أن **F** الجدولية تساوي **<u>3.80</u>** ويمكن توضيح ذلك بالرسم كما يلي:



المقارنة والقرار: وأذ أن قيمة الإحصائية المسحوبة والتي تساوي **6.04** أكبر من القيمة الجدولية **3.80** فإنها تقع في منطقة الرفض وبالتالي فإن القرار هو رفض الفرضية العدمية وقبول الفرضية البديلة التي تنص ان هناك فروق جوهرية بين الوحدات المعيبة المنتجة من قبل المكائن الثلاثة

مثال(3): أختبر الفرضية $Ho: \mu_1 = \mu_2 = \mu_3$ بعد اكمال المعلومات في جدول تحليل التباين لبيانات مصنع معين ينتج ثمانية انواع السلع (n=8) بتأثير ثلاثة مستويات من درجات الحرارة (a=3) عند مستوى معنوية 0.05 وكما مبين ادناه

مصدر التغير	مجموع المربعات	درجات الحرية	متوسط المربعات	الإحصائية F
بسبب المعاملات	....	....	....	....
بسبب الخطأ	....	....	2	
الكلي	62	....		

الحل:

عدد درجات الحرية الكلية 23=1-24=1- (8*3)=1- (n*a)

عدد درجات الحرية بسبب الخطأ 21=7*3= (1-8)3= (1-a(n

عدد درجات الحرية بسبب المعاملات 2=1-3=1-a

وعليه فان

مجموع المربعات بسبب الخطأ $MSE = \frac{SSE}{a(n-1)} \Rightarrow 2 = \frac{SSE}{21} \Rightarrow SSE = 42$

أ.م.د. وقاص سعد خلف
م.سرمد علوان صالح

مجموع المربعات بسبب المعاملات $SSR = SST - SSE = 62 - 42 = 20$

وعليه فان

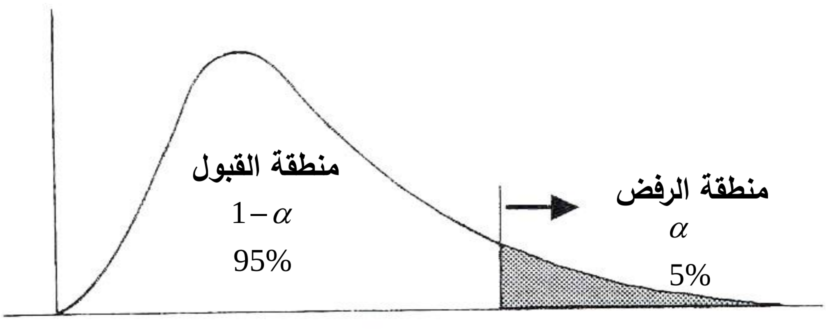
متوسط المربعات بسبب المعاملات $MSR = \frac{SSR}{a-1} = \frac{20}{2} = 10$

وان احصائية الاختبار F $F = \frac{MSR}{MSE} = \frac{10}{2} = 5$

وان جدول تحليل التباين سيكون بالشكل الآتي

مصدر التغير	مجموع المربعات	درجات الحرية	متوسط المربعات	الإحصائية F
بسبب المعاملات	**20**	**2**	**10**	**5**
بسبب الخطأ	**42**	**21**	2	
الكلي	62	**23**		

من جدول توزيع **F** وعند مستوى معنوية **5%** **وبدرجات حرية** **2** **للبسط،** **21** **للمقام** نجد أن **F** الجدولية تساوي **3.47** ويمكن توضيح ذلك بالرسم كما يلي:



F= 3.47

وعليه يتم رفض الفرضية العدمية $Ho: \mu_1 = \mu_2 = \mu_3$

تمارين الفصل الثالث

س(1) تبيع شركة دورو لإنتاج الصابون في ثلاثة اغلفة مختلفة وبنفس السعر وكما مبين في الجدول ادناه ولمدة خمسة أشهر. علماً ان المبيعات تتوزع توزيعاً طبيعياً ولها نفس التباين

الغلاف الأول	الغلاف الثاني	الغلاف الثالث
90	78	87
91	81	83
84	79	79
82	82	81
88	80	80

اختبر فيما إذا كان متوسط المبيعات لكل غلاف متساوياً ام لا عند مستوى معنوية (0.05) علماً ان القيمة الجدولية $F_{0.05}(2,12) = 3.88$

س(2) يرغب احد الباحثين في اختبار تساوي متوسطات الزيادة في الوزن لثلاث أنواع من الاعلاف التي يمكن ان تتناولها العجول الصغيرة، ولقد قام الباحث باختيار ثلاث أنواع من العجول الصغيرة ووزع الأنواع الثلاثة من الاعلاف على العينات الثلاثة بطريقة عشوائية وبعد فترة زمنية قصيرة من تطبيق هذا البرنامج الغذائي قام بقياس الزيادة بالكيلو غرام وكانت كالاتي

النوع الأول	النوع الثاني	النوع الثالث
14	20	9
15	22	14
16	21	21
14	27	08
16	23	07
15	19	19

فهل توجد فروقات جوهرية بين متوسطات الزيادة في الوزن عند مستوى 0.05

س(3) البيانات التالية تمثل عدد الوحدات المعيبة لعينه من خمسة أنواع من المكائن تحت مستوى معنوية 0.05. هل يمكننا القول بان هناك فروق جوهرية بين الوحدات المعيبة المنتجة من قبل المكائن الخمسة

الماكنة الاولى	25	15	20	22	24	26
الماكنة الثانية	40	44	49	45	46	40
الماكنة الثالثة	30	30	32	35	……	……
الماكنة الرابعة	60	65	64	65	62	……
الماكنة الخامسة	65	66	64	……	……	……

س(4) الجدول الاتي يمثل درجات مجموعة من الطلبة يتم تدريسهم مقررات مادة الأساليب الكمية بثلاث طرائق مختلفة

الطريقة الاولى	55	65	80	72	44	36
الطريقة الثانية	70	54	49	45	67	……
الطريقة الثالثة	60	35	82	45	……	……

فهل هناك فرقاً بين طرائق التدريس الثلاثة عند مستوى معنوية 0.05

س(5) أختبر الفرضية $Ho: \mu_1 = \mu_2 = \mu_3 = \mu_4$ بعد اكمال المعلومات في جدول تحليل التباين لبيانات مصنع معين ينتج نوعاً السلع بتأثير أربع مستويات من درجات الحرارة (a=4) عند مستوى معنوية 0.05 وكما مبين ادناه

مصدر التغير	مجموع المربعات	درجات الحرية	متوسط المربعات	الإحصائية F
بسبب المعاملات	…….	…….	…….	…….
بسبب الخطأ	…….	…….	8	
الكلي	280	31		