## المحور5: التحليل التمييزي (Discriminant Analysis)

### 1.تعريف التحليل التمييزي:

التحليل التمييزي هو اسلوب إحصائي لتحليل البيانات حيث يهتم بمسالة التمييز بين مجموعتين او اكثر والتي تكون متشابهة في كثير من الصفات على أساس عدة متغيرات من خلال استخدام دالة التمييز وهي عبارة عن تركيب خطي للمتغيرات المستقلة، لذلك التحليل التمييزي هو أحد أساليب التحليل متعدد المتغيرات كالتحليل العاملي والتحليل العنقودي، تحليل التباين المتعدد، الارتباط القانوني، يقوم هذا التحليل على تقنية تصنيف المشاهدات إلى مجموعتين أو أكثر وفقاً لمتغيرات معينة. يعتمد نموذج التحليل التمييزي على إيجاد العلاقة بين المتغيرات بتحديد المتغير التابع والمتغيرات المستقلة، والتوصل إلى بناء نموذج يمثل العلاقة بين هذه المتغيرات بشكل علاقات خطية تمثل الظاهرة

المدروسة ويساعدنا على فهم واقع الظاهرة، وإجراء التقديرات والتنبؤ بالمستقبل بما يسمى دالة التمايز التي تعمل على تعظيم الفروق بين متوسط المجموعات وتقليل التشابه في أخطاء التصنيف.

### 2. أهداف التحليل التمييزي:

- انشاء دوال المتغير التابع والمعروفة بدوال التمايز للمتغيرات التوضيحية للفصل او التمييز بين فئات المتغير التابع.
- تعمل هذه الدوال على تعظيم التباين بين المجموعات وتصغير التباين داخل المجموعة.
- التعرف على اكثر المتغيرات التوضيحية تاثيرا في تصنيف مجموعات المتغير التابع اي ترتيب المتغيرات التي تسهم بقدر كبير في التمييز او توضيح الاختلافات بين المجموعات (فئات المتغير التابع).
- تصنيف مشاهدات جديدة وتوزيعها على المجموعات (فئات المتغير التابع).
- الوصول الى اقل نسبة خطأ للتوصيف.
- تقييم دقة التصنيف كنسبة مئوية اي تقييم دوال التصنيف (معنوية ام غير معنوية).

### 3. أنواع التحليل التمييزي:

هناك ثلاثة أنواع من التحليل التمييزي، وذلك حسب طريقة إدخال المتغيرات إلى التحليل، تتمثل في:

- التحليل المباشر (Direct Discriminant Analysis): ويتم فيه إدخال جميع المتغيرات المستقلة إلى التحليل مرة واحدة ودون استثناء و دون إعطاء أي أهمية لترتيب دخولها.
- التحليل الهرمي (Herarchical Discriminant Analysis): وهنا يتم إدخال المتغيرات إلى التحليل تبعا لما يراه الباحث من أهمية للمتغيرات المستقلة و بالترتيب الذي يعتقد أنه مناسبا.
- التحليل التدريجي (Stepwise Discriminant Analysis): في هذه الحالة يكون ترتيب إضافة المتغيرات المستقلة إلى التحليل و استبعادها منه تبعا لمعايير إحصائية فقط[27].

## 4. شروط التحليل التمييزي:

-اختبار العينة يجب أن يكون بشكل عشوائي؛

-عدم تساوي متوسطات المجموعات؛

- تساوي مصفوفة التباين والتباين المشترك بين المجموعتين؛

- أن تكون المجموعات منفصلة و قابلة للتحديد؛

- أن تتوزع المتغيرات توزيع طبيعي؛

- عدم وجود قيم متطرفة، حيث أن التحليل التمييزي أكثر حساسية وتأثرا بالقيم الشاذة ووجودها يبعد توزيع البيانات عن التوزيع الطبيعي؛

- استقلالية المشاهدات أي عدم وجود ارتباط بين المتغيرات المدروسة.

## 5. خطوات التحليل التمييزي:

1. تحديد نوع التحليل التمييزي المناسب: وصفي أم تنبؤي، بسيط أم متعدد، خطي أم غير خطي، كمي أو نوعي أم لوجستي .
2. تحديد المتغيرات المستقلة المناسبة واللازمة لتحقيق أهداف البحث، وجمع البيانات عنها .
3. تحديد عدد المجموعات في المجتمع التي ستستخدم في التحليل( g≥2) ، وسحب العينات الطبقية منها، وجمع البيانات اللازمة منها .
4. اختبار البيانات والتأكد من أنها تحقق الافتراضات والشروط المفروضة عليها .
5. إجراء التحليل التمييزي حسب خطواته العملية، والحصول على النتائج المطلوبة .
6. تفسير النتائج والعمل على الاستفادة منها ، من خلال التنبؤ بانتماء وتصنيف مفردات جديدة في المجتمع

## 6. بناء الدالة التمييزية:

تقوم الدالة التمييزية على فكرة أساسية وهي تقسيم الأشخاص إلى مجموعتين هما (مصاب أو غير مُصاب) وذلك بالاعتماد على مجموعة من المُتغيرات أو العوامل وتعمل الدالة على زيادة درجة التجانس بين مفردات المجموعة الواحدة وتقليل درجة التجانس بين المجموعتين وبالتالي تسهيل إمكانية تصنيف أي مُشاهدة جديدة إلى إحدى المجموعتين بأقل خطأ للتصنيف كما تعمل الدالة على استبعاد المُتغيرات التي ليس لها تأثير معنوي في التمييز والفصل بين المجموعتين.

**ويتم حساب الدالة التمييزية كالتالي: –**

في حالة تعدد المجموعات تتعدد الدوال التمييزية ولكننا سنقتصر على الدالة التمييزية بين مجموعتين فقط.

**أولاً: – حساب متوسطات المُتغيرات في كل مجموعة وإيجاد الفرق بين متوسط: –**

$$\bar{x}_{i\,(1)} = \begin{bmatrix} \bar{x}_{1\,(1)} \\ \bar{x}_{2\,(1)} \\ \vdots \\ \vdots \\ \bar{x}_{k\,(1)} \end{bmatrix}$$

**متوسطات المُتغيرات في المجموعة الثانية: -**

$$\bar{x}_{i\,(2)} = \begin{bmatrix} \bar{x}_{1\,(2)} \\ \bar{x}_{2\,(2)} \\ \vdots \\ \vdots \\ \bar{x}_{k\,(2)} \end{bmatrix}$$

k عدد المتغيرات المستقلة

**الفرق بين متوسط المتغير في المجموعتين:**

$$d_i = \bar{x}_{i\,(1)} - \bar{x}_{i\,(2)} = \begin{bmatrix} \bar{x}_{11} - \bar{x}_{12} \\ \bar{x}_{21} - \bar{x}_{22} \\ \vdots \quad - \quad \vdots \\ \bar{x}_{k\,(1)} \quad \bar{x}_{k\,(2)} \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ \vdots \\ d_k \end{bmatrix} \quad \text{(المسافة)}$$

**ثانيًا: إيجاد التباين والتغاير المشترك بين المجموعتين: -**

$$S_{ii} = \sum x_i^2 - \frac{(\sum xi)^2}{n}$$

$$S_{ij} = \sum x_i x_j - \frac{\sum x_i \sum x_i}{n}$$

**∴ التباين المشترك**

$$V_{ii} = \frac{S_{ii} + S_{ii^{(2)}}}{n_1 + n_2 - 2}$$

**∴ التغاير المشترك**

$$V_{ij} = \frac{S_{ij\,(1)} + S_{ij^{(2)}}}{n_1 + n_2 - 2}$$

**مصفوفة التباين والتغاير المشترك بين المجموعتين.**

$$\mathrm{v} = \begin{bmatrix} \mathrm{v}_{11}\mathrm{v}_{12}\mathrm{v}_{13}\dots\dots\dots\mathrm{v}_{1k} \\ \mathrm{v}_{21}\mathrm{v}_{22}\mathrm{v}_{23} \qquad\qquad \mathrm{v}_{2k} \\ \vdots \quad \vdots \quad \vdots \qquad\qquad \vdots \\ \mathrm{v}_{k1}\mathrm{v}_{k2}\mathrm{v}_{k3}\dots\dots\dots\mathrm{v}_{kk} \end{bmatrix}$$

وهي عبارة عن مصفوفة مربعة ومتماثلة والقطر الرئيسي لها يُمثل التباين المُشترك وباقي العناصر التغاير المشترك.

**بناء الدالة التمييزية:**

تأخذ الدالة التمييزية بمعاملات معيارية الشكل التالي:

$$\hat{L} = \hat{\alpha}_1\ x_1 + \alpha_2\ x_2 + \qquad\qquad + \alpha_k\ x_k$$

**حيث**

$$\hat{\alpha} = \mathrm{v}^{-1}\ d$$

$-1$

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix} = \begin{bmatrix} \mathrm{v}_{11}\ \mathrm{v}_{12} \cdots\cdots\cdots \mathrm{v}_{1k} \\ \mathrm{v}_{21}\ \mathrm{v}_{22} \cdots\cdots\cdots \mathrm{v}_{2k} \\ \vdots \quad \vdots \qquad\qquad \vdots \\ \mathrm{v}k_1\mathrm{v}k_2 \cdots\cdots\cdots \mathrm{v}_{kk} \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{bmatrix}$$

**حيث**

$\hat{\alpha}$ معاملات الدالة التمييزية المعيارية.

$\mathrm{v}^{-1}$ : معكوس مصفوفة التباين والتغاير المشترك.

$d_i$ : مصفوفة المسافة بين متوسط المُتغيرات في كلا المجموعتين.

## 7. اختبار قدرة الدالة على التمييز:

١- اختبار F (F test)

وذلك لاختبار قدرة الدالة على التمييز وعن طريق الفرضية التي تنص على ان الدالة ليس لديها القدرة على التمييز ($H_0$) ضد الدالة لديها القدرة على التمييز ($H_1$) ويعتمد هذا الاختبار على قياس الاختلافات بين المجموعات وداخل المجموعات بين المفردات ويتم ذلك من خلال تكوين جدول تحليل التباين التالي: -

| Source | SS | Df | Ms | F |
|---|---|---|---|---|
| بين المجموعات Between x's | SSB | k-1 | $M_{SB}$ | $M_{SB}$ |
| الخطأ Within x's | SSE | n-k | $M_{SE}$ | $M_{SE}$ |
| الكلي Total | SST | n-1 | | |

**حيث ان: -**

**١- مجموع مربعات الأخطاء يحسب كالتالي: -**

$$SSE = D^2 = \hat{\alpha}_1 d_1 + \hat{\alpha}_2 d_2 + \cdots\cdots\cdots\cdots + \hat{\alpha}_k d_k$$

**٢- مجموع مربعات بين المتغيرات: -**

$$SSB = \frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 2)} \times (D^2)^2$$

**٣- مجموع مربعات الكلى: -**

$$SST = SSB + SSE$$

**ويتم الاختيار كالتالي:**

**١- صياغة الفروض:**

$H_0$ : الدالة ليس لها قدرة على التمييز

$H_1$ : الدالة لها القدرة على التمييز

**٢- القيمة المحسوبة:**

$$F = \frac{M_{SB}}{M_{SE}}$$

**٣- القيمة الجدولية:**

$$F\,(k-1, n-k)$$

٤- القرار:

إذا كانت F المحسوبة أكبر من F الجدولية نرفض الفرض العدمي ونقبل بالفرض البديل ويكون للدالة قدرة عالية على التمييز والعكس صحيح.

٢- اختبار ويلكس لمدا (A) Wil'ks Lambda:

تأخذ الفروض الشكل الآتي:

الدالة ليس لها مقدرة على التمييز

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

الدالة لها القدرة على التمييز

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

القيمة المحسوبة

$$\Lambda = \prod_{i=1}^{k} \frac{1}{1+\lambda_i}$$

$\lambda_i$ الجذر الكامن (eigenvalues) لكل المُتغيرات.

k عدد المُتغيرات

القرار: تنحصر قيمة

$$o \leq \Lambda \leq 1$$

**إذا كان**

$\Lambda = 1$ معناها تساوي متوسطات المجموعتين وبالتالي عدم مقدره الدالة على التمييز والفصل.

$\Lambda = 0$ معناها عدم تساوي متوسطات المجموعتين والدالة لها القدرة عالية على التمييز.

إذا اقتربت قيمة $\Lambda$ من الواحد دليل على عدم مقدرة الدالة على التمييز وإذا اقتربت من الصفر دليل على قدرة الدالة على التمييز.

وتستخدم إحصائية "ويلكس لمدا" لاختبار معنوية المتغيرات الداخلة في النموذج حيث يتم الإبقاء على المُتغيرات لها أدنى قيمة لإحصائية Wilk's Lambda وأعلى قيمة لـ F.

## 8. نقطة الفصل (القطع):

بعد تكوين الدالة التمييزية واختبار قدرتها على التمييز والفصل بين المجموعتين يبدأ الاستخدام الثاني لها وهو كيفية تصنيف المشاهدة الجديدة إلى أي المجموعتين تنتمي ويتم ذلك من خلال الخطوات الآتية: -

١- **تحديد نقطة الفصل وهي تمثل متوسط المتوسطين:**

$$\bar{\bar{L}} = \frac{\bar{L}_{(1)} + \bar{L}_{(2)}}{2}$$

**حيث ان**

$\bar{\bar{L}}$ : نقطة الفصل.

$\bar{L}_{(1)}$ متوسط القيم التمييزية للمجموعة الأولى.

$\bar{L}_{(2)}$ متوسط القيم التمييزية للمجموعة الثانية.

**قاعدة التصنيف Classification Role:**

من خلال هذه القاعدة يُمكن تصنيف أو التنبؤ بانتماء مفردة جديدة لإحدى المجموعتين بأقل خطأ تصنيف على النحو التالي:

١) إذا كان $\bar{L}_{(1)} > \bar{L}_{(2)}$

وإذا كانت القيمة التمييزية للمفردة الجديدة أكبر من نقطة الفصل تصنف ضمن المجموعة الأولى وإذا كانت القيمة التمييزية للمفردة الجديدة أقل من نقطة الفصل تصنف ضمن المجموعة الثانية وإذا ساوت نقطة الفصل تصنف عشوائيًا ضمن أي مجموعة من المجموعتين.

٢) إذا كان $\bar{L}_{(1)} < \bar{L}_{(2)}$

وإذا كانت القيمة التمييزية للمفردة الجديدة أعلى من نقطة الفصل تصنف ضمن المجموع الثانية وإذا كانت أقل تصنف ضمن المجموعة الأولى وإذا تساوت معها تصنف عشوائيًا ضمن أي مجموعة في المجموعتين.

**تمرين1:**

شركة تسويقية ترغب في تصنيف العملاء بناء على سلوكهم الشرائي إلى فئتين:

- $Y = 0$ :العملاء غير المهتمين بالعرض.
- $Y = 1$ :العملاء المهتمين بالعرض.

تم جمع البيانات حول عينة من 10 عملاء، وتشمل المتغيرات المستقلة:

- $X_1$: عدد الإعلانات التي شاهدها العميل خلال الشهر الماضي.
- $X_2$: متوسط الوقت الذي يقضيه العميل على الموقع الإلكتروني (بالدقائق).

**البيانات:**

| حالة | $X_1$ (عدد الإعلانات) | $X_2$ (متوسط الوقت) | الفئة(Y) |
|---|---|---|---|
| 1 | 3 | 10 | 0 |
| 2 | 4 | 12 | 0 |
| 3 | 5 | 15 | 0 |
| 4 | 6 | 20 | 1 |
| 5 | 7 | 25 | 1 |
| 6 | 8 | 30 | 1 |
| 7 | 9 | 35 | 1 |
| 8 | 10 | 40 | 1 |
| 9 | 2 | 8 | 0 |
| 10 | 11 | 45 | 1 |

**المطلوب:**

1. استخدم التحليل التمييزي الخطي (LDA) لبناء دالة تمييزية.
2. قم بتصنيف عميل جديد لديه:
   - $X_1 = 7$ (عدد الإعلانات).
   - $X_2 = 22$ (متوسط الوقت).

**1. حساب المتوسط لكل متغير داخل كل فئة:**

الفئة $Y = 0$

- المتوسط لـ: $X_1$

$$\bar{X}_{1,0} = \frac{\sum_{i=1}^{n_0} X_{1,i}}{n_0} = \frac{3+4+5+2}{4} = \frac{14}{4} = 3.5$$

- المتوسط لـ: $X_2$

$$\bar{X}_{2,0} = \frac{\sum_{i=1}^{n_0} X_{2,i}}{n_0} = \frac{10+12+15+8}{4} = \frac{45}{4} = 11.25$$

الفئة $Y = 1$

- المتوسط لـ: $X_1$

$$\bar{X}_{1,1} = \frac{\sum_{i=1}^{n_1} X_{1,i}}{n_1} = \frac{6+7+8+9+10+11}{6} = \frac{51}{6} = 8.5$$

- المتوسط لـ $X_2$:

$$\bar{X}_{2,1} = \frac{\sum_{i=1}^{n_1} X_{2,i}}{n_1} = \frac{20 + 25 + 30 + 35 + 40 + 45}{6} = \frac{195}{6} = 32.5$$

إذن:

$$\bar{X}_{1,0} = 3.5, \quad \bar{X}_{2,0} = 11.25, \quad \bar{X}_{1,1} = 8.5, \quad \bar{X}_{2,1} = 32.5$$

**2. حساب معاملات المتغيرات $w_1$ و$w_2$:**

المعادلة العامة لمعاملات المتغيرات هي:

$$w_i = \bar{X}_{i,1} - \bar{X}_{i,0}$$

- بالنسبة لـ $w_1$:

$$w_1 = \bar{X}_{1,1} - \bar{X}_{1,0} = 8.5 - 3.5 = 5$$

- بالنسبة لـ $w_2$:

$$w_2 = \bar{X}_{2,1} - \bar{X}_{2,0} = 32.5 - 11.25 = 21.25$$

إذن:

$$w_1 = 5 \quad \text{و} \quad w_2 = 21.25$$

**3. حساب الثابت $w_0$:**

المعادلة المستخدمة لحساب $w_0$ هي:

$$w_0 = -\frac{1}{2}\left(\sum_{i=1}^{p} \bar{X}_{i,1}^2 - \sum_{i=1}^{p} \bar{X}_{i,0}^2\right)$$

- حساب: $\sum_{i=1}^{p} \bar{X}_{i,1}^2$

$$\sum_{i=1}^{p} \bar{X}_{i,1}^2 = \bar{X}_{1,1}^2 + \bar{X}_{2,1}^2 = 8.5^2 + 32.5^2 = 72.25 + 1056.25 = 1128.5$$

- حساب: $\sum_{i=1}^{p} \bar{X}_{i,0}^2$

$$\sum_{i=1}^{p} \bar{X}_{i,0}^2 = \bar{X}_{1,0}^2 + \bar{X}_{2,0}^2 = 3.5^2 + 11.25^2 = 12.25 + 126.5625 = 138.8125$$

- حساب الفرق:

$$\sum_{i=1}^{p} \bar{X}_{i,1}^2 - \sum_{i=1}^{p} \bar{X}_{i,0}^2 = 1128.5 - 138.8125 = 989.6875$$

- حساب: $w_0$

$$w_0 = -\frac{1}{2} \times 989.6875 = -494.84375$$

إذن:

$$w_0 = -494.84375$$

**4. بناء دالة التمييز النهائية:**

دالة التمييز تصبح:

$$D(X) = 5X_1 + 21.25X_2 - 494.84375$$

**5. تصنيف الحالة الجديدة:**

لدينا حالة جديدة مع القيم التالية:

$$X_1 = 7 \quad \text{و} \quad X_2 = 22$$

نحسب قيمة $D(X)$ لهذه الحالة:

$$D(X) = 5(7) + 21.25(22) - 494.84375$$
$$D(X) = 35 + 467.5 - 494.84375 = 7.65625$$

**اتخاذ القرار:**

- إذا كانت: $D(X) > 0$ الحالة تنتمي إلى الفئة $Y = 1$.
- إذا كانت: $D(X) \leq 0$ الحالة تنتمي إلى الفئة $Y = 0$.

في هذه الحالة:

$$D(X) = 7.65625 > 0$$

إذن:

**الحالة تنتمي إلى الفئة $Y = 1$**