محاضرات حول التحليل التمييزي

المقدمة

يعتمد أسلوب التحليل الإحصائي متعدد المتغيرات بأساليبه المختلفة على وصف وتحليل الظواهر ذات الأبعاد والمتغيرات المتعددة. فإذا كانت المشاهدات تشترك فيما بينها بمجموعة من الخصائص والصفات بدرجات متفاوتة، فإن التحليل الإحصائي متعدد المتغيرات يتناول دراسة بيانات تلك المشاهدات والتعبير عنها من خلال أكثر المتغيرات تأثيرا في الظاهرة محل الدراسة. وسيتم من خلال هذا البحث التركيز على أسلوب التحليل التمييزي الذي يستخدم لتصنيف الأفراد في جماعات على مقياس أو أكثر، أو التمييز بين الجماعات على أساس التجميع الخطي لعدة مقاييس بعد الحصول على قيمة فيشر دالة في اختبار تحليل التباين المتعدد

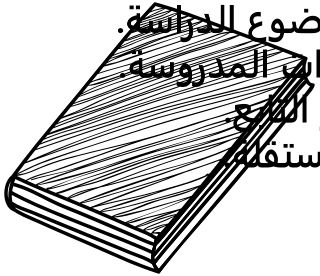
كما يستخدم التحليل التمييزي في الدراسات التي تهدف إلى تصنيف الأفراد في مجموعات على أساس متغيرات كمية منبئة

مفهوم التحليل التمييزي

يُعتبر التحليل التمييزي أحد الأساليب المهمة في التحليل متعدد المتغيرات يهدف هذا التحليل إلى دراسة المتغيرات الداخلة في النموذج بشكل مترابط، مع مراعاة العلاقات المتداخلة بينها. يتميز التحليل التمييزي بقدرته على بناء نموذج إحصائي يعكس العلاقة المتبادلة بين المتغيرات المختلفة.
وتكمن أهميته الرئيسية في قدرته على التمييز بين المجموعات أو المشاهدات باستخدام مجموعة من المتغيرات، حيث يقوم بإنشاء تركيبات خطية من المتغيرات، تُعرف بمتغيرات التمايز
يعتمد نموذج التحليل التمييزي على التوصل إلى دالة التمايز والتي تهدف إلى تعظيم الفروق بين متوسطات المجموعات وتقليل أخطاء التصنيف، وذلك عبر إيجاد تركيبات خطية مناسبة للمتغيرات

أهداف التحليل التمييزي

يهدف التحليل التمييزي إلى تحقيق مجموعة من الأهداف، أهمها:
تصميم التوليفات الخطية للمتغيرات التي تكون الأكثر فاعلية في تفسير موضوع الدراسة.
التحقق من وجود فروق ذات دلالة إحصائية بين المجموعات بالنسبة للمتغيرات المدروسة.
تحديد المتغيرات التي تساهم بأكبر قدر في التمييز بين فئات المتغير التابع.
تصنيف الحالات إلى فئات المتغير التابع بناءً على قيم المتغيرات المستقلة.
تقييم دقة التصنيف كنسبة مئوية

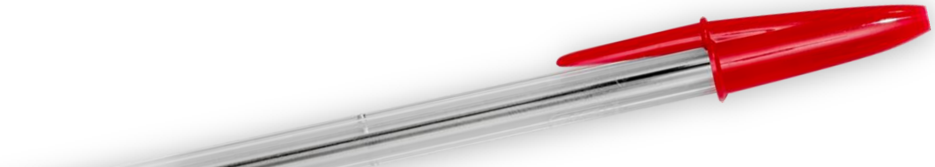


أنواع التحليل التمييزي

يمكن تصنيف التحليل التمييزي إلى ثلاثة أنواع رئيسية:
تحليل التمايز المباشر: يتم إدخال جميع المتغيرات إلى المعادلات دفعة واحدة.
تحليل التمايز الهرمي: يتم إدخال المتغيرات إلى النموذج بناءً على ترتيب أو جداول يحددها الباحث.
تحليل التمايز المتدرج: يتم اختيار المتغيرات بناءً على معايير إحصائية تُحدد أولوية إدخالها إلى النموذج

مسلمات التحليل التمييزي

يعتمد التحليل التمييزي على عدد من المسلمات التي يجب تحققها لضمان صحة النتائج، وأهمها:
التوزيع الطبيعي: يجب أن تكون المتغيرات التابعة الكمية موزعة توزيعًا طبيعيًا، لذا يُفضل استخدام عينات متوسطة أو كبيرة الحجم لتحقيق نتائج موثوقة.
تماثل التباين: يجب أن تكون تباينات المتغيرات التابعة متساوية في جميع مستويات العامل
اختيار عشوائي للعينة: يجب أن تكون العينة مختارة بشكل عشوائي، وأن تكون درجات الأفراد في أي متغير مستقلة عن درجات الآخرين.
الاستقلالية: عدم الثقة في نتائج التحليل التمييزي إذا تم انتهاك شرط استقلالية العينات



شروط التحليل التمييزي

لضمان فعالية التحليل التمييزي، يجب تحقق الشروط التالية:

أن تكون المجتمعات موضوع الدراسة منفصلة وقابلة للتحديد، حتى وإن كانت متداخلة بدرجات معينة.

أن تكون كل مفردة في المجتمع قابلة للوصف باستخدام مجموعة من المقاييس أو المتغيرات المستقلة، على أن تكون هذه المتغيرات مقاسة بقيم محدودة

أن تختلف المجتمعات محل الدراسة من حيث أوساط المتغيرات التابعة، أي أن تكون متجهات أوساط المتغيرات غير متساوية.

أن تحتوي البيانات المستخدمة في التحليل على عينة عشوائية من أعضاء كل مجتمع، بحيث تكون هذه العينات ممثلة للمجتمعات المدروسة.

أن تكون المجتمعات موضوع الدراسة ذات توزيع طبيعي، بحيث تتوزع البيانات بطريقة معتدلة.

انواع الدوال التمييزية

1. الدالة المميزة الخطية
2. االدالة المميزة التربيعية
3. الدالة المميزة اللوجستية

1. الدالة المميزة الخطية Linear Discriminant Function

تستخدم هذه الدالة عندما تكون المجتمعات المدروسة ذات توزيع طبيعي متعدد المتغيرات بمتجهات متوسط مختلفة و مصفوفة تباين مشترك متساوية وهناك حالتان:

1- حالة مجموعتين (مجتمعين)

2- حالة عدة مجاميع (مجتمعات)

حالة المجموعتين

نفرض لدينا عينة مسحوبة من مجتمعين يتوزعان توزيعاً طبيعياً بمتوسطين $\underline{\mu_1}$, $\underline{\mu_2}$ ومصفوفة تباين مشترك Σ لكل مجتمع. أي أن:

$$\underline{X_1} \sim N_p(\underline{\mu_1}\ ,\ \Sigma\ )$$
$$\underline{X_2} \sim N_p(\underline{\mu_2}\ ,\ \Sigma\ )$$

يمكن صياغة دالة بالإعتماد على مقاييس من هذه القيم وأن هذه الدالة تمكننا من إختيار أي مشاهدة و تحديد المجتمع الذي تعود إليه. إن المتغير العشوائي X له عادةً دالة كثافة إحتمالية إما f_1

$(x\ ,\theta_1)$ أو $f_2\,(x\ ,\theta_2)$ و θ_i هنا ترمز لأي معلمة يختلف التوزيع بموجبها. وبذلك، فإن هذا يعني أن:

$$f_i(x,\mu_i) = \frac{1}{(2\pi)^{\frac{p}{2}}(\Sigma)^{\frac{1}{2}}}\, e^{-\frac{1}{2}(\underline{x}-\underline{\mu_i})'\Sigma^{-1}(\underline{x}-\underline{\mu_i})} \quad , \quad i=1,2$$

وبذلك فإن نسبة الإمكان الأعظم هنا تكون:

$$\frac{f_1(x,\mu_1)}{f_2(x,\mu_2)} = \frac{\dfrac{1}{(2\pi)^{\frac{p}{2}}(\Sigma)^{\frac{1}{2}}}\, e^{-\frac{1}{2}(\underline{x}-\underline{\mu_1})'\Sigma^{-1}(\underline{x}-\underline{\mu_1})}}{\dfrac{1}{(2\pi)^{\frac{p}{2}}(\Sigma)^{\frac{1}{2}}}\, e^{-\frac{1}{2}(\underline{x}-\underline{\mu_2})'\Sigma^{-1}(\underline{x}-\underline{\mu_2})}} \geq \lambda$$

أو أن:

$$\exp\left[-\frac{1}{2}\left\{(\underline{x}-\underline{\mu_1})'\Sigma^{-1}(\underline{x}-\underline{\mu_1})-(\underline{x}-\underline{\mu_2})'\Sigma^{-1}(\underline{x}-\underline{\mu_2})\right\}\right] \geq \lambda$$

و عند تفكيك هذه المعادلة وإضافة وطرح المقدار $\underline{\mu_2'}\Sigma^{-1}\underline{\mu_1}$ يصبح لدينا:

$$\exp\left[\underline{x}'\Sigma^{-1}(\underline{\mu_1}-\underline{\mu_2})-\frac{1}{2}\left\{(\underline{\mu_1}+\underline{\mu_2})'\Sigma^{-1}(\underline{\mu_1}-\underline{\mu_2})\right\}\right] \geq \lambda$$

وبأخذ اللوغاريتم الطبيعي للطرفين:

$$\left[\underline{x}'\Sigma^{-1}(\underline{\mu_1}-\underline{\mu_2})-\frac{1}{2}\left\{(\underline{\mu_1}+\underline{\mu_2})'\Sigma^{-1}(\underline{\mu_1}-\underline{\mu_2})\right\}\right] \geq \log\lambda$$

$$J_2(x,\mu_2) \quad \frac{1}{(2\pi)^{\frac{p}{2}}(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu_2})'\Sigma^{-1}(\underline{x}-\underline{\mu_2})}$$

أو أن:

$$\exp\left[-\frac{1}{2}\left\{(\underline{x}-\underline{\mu_1})'\Sigma^{-1}(\underline{x}-\underline{\mu_1})-(\underline{x}-\underline{\mu_2})'\Sigma^{-1}(\underline{x}-\underline{\mu_2})\right\}\right] \geq \lambda$$

وعند تفكيك هذه المعادلة وإضافة وطرح المقدار $\underline{\mu_2}'\Sigma^{-1}\underline{\mu_1}$ يصبح لدينا:

$$\exp\left[\underline{x}'\Sigma^{-1}(\underline{\mu_1}-\underline{\mu_2})-\frac{1}{2}\left\{(\underline{\mu_1}+\underline{\mu_2})'\Sigma^{-1}(\underline{\mu_1}-\underline{\mu_2})\right\}\right] \geq \lambda$$

وبأخذ اللوغاريتم الطبيعي للطرفين:

$$\left[\underline{x}'\Sigma^{-1}(\underline{\mu_1}-\underline{\mu_2})-\frac{1}{2}\left\{(\underline{\mu_1}+\underline{\mu_2})'\Sigma^{-1}(\underline{\mu_1}-\underline{\mu_2})\right\}\right] \geq \log\lambda$$

وعندما تكون $\lambda = 1$ فإن:

$$\left[\underline{x}'\Sigma^{-1}(\underline{\mu_1}-\underline{\mu_2})-\frac{1}{2}\left\{(\underline{\mu_1}+\underline{\mu_2})'\Sigma^{-1}(\underline{\mu_1}-\underline{\mu_2})\right\}\right] \geq 0$$

وبإستخدام مقدرات الإمكان الأعظم إلى كل من Σ, μ_2, μ_1 وتعويضها هنا، يصبح لدينا:

$$W = \left[\underline{x}'S^{-1}(\underline{\bar{x}_1}-\underline{\bar{x}_2})-\frac{1}{2}(\underline{\bar{x}_1}+\underline{\bar{x}_2})'S^{-1}(\underline{\bar{x}_1}-\underline{\bar{x}_2})\right] \geq 0$$

$$\underline{\bar{x}_1} = \sum_{j=1}^{n_1}\underline{x_{1j}}/n_1$$

حيث أن:

$$\underline{\bar{x}_2} = \sum_{j=1}^{n_2}\underline{x_{2j}}/n_2$$

$$S = \left[\sum_{j=1}^{n_1}(\underline{x_{1j}}-\underline{\bar{x}_1})(\underline{x_{1j}}-\underline{\bar{x}_1})' + \sum_{j=1}^{n_2}(\underline{x_{2j}}-\underline{\bar{x}_2})(\underline{x_{2j}}-\underline{\bar{x}_2})'\right]/(n_1+n_2-2)$$

وأن جزئي المعادلة w أعلاه يتكون من دالة التمييز الخطية:

$$y = \underline{x}'S^{-1}(\underline{\bar{x}_1}-\underline{\bar{x}_2})$$

ونقطة الفصل أو التمييز z:

$$z = \frac{1}{2}D^2 = \frac{1}{2}(\underline{\bar{x}_1}+\underline{\bar{x}_2})'S^{-1}(\underline{\bar{x}_1}-\underline{\bar{x}_2})$$

حيث أن D^2 هي إحصاءة مهلونوبيس Mahalanobis.

وأن دالة التمييز الخطية ممكن أن تكتب على شكل الدالة الخطية التالية:

$$y = \underline{x}'\underline{C} \quad \text{or} \quad y = \underline{C}'\underline{x}$$

حيث:

$$C = S^{-1}(\underline{\bar{x}_1}-\underline{\bar{x}_2})$$

والأن يمكن أن نستخلص الآتي من دالة التمييز الخطية:

وأن دالة التمييز الخطية ممكن أن تكتب على شكل الدالة الخطية التالية:

$$y = \underline{x}'\underline{C} \quad \text{or} \quad y = \underline{C}'\underline{x}$$

حيث:

$$C = S^{-1}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)$$

والآن يمكن أن نستخلص الآتي من دالة التمييز الخطية:

$$\bar{y}_1 = \underline{\bar{x}}_1' S^{-1}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)$$
$$\bar{y}_2 = \underline{\bar{x}}_2' S^{-1}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)$$

وبالتالي يمكننا التعبير عن نقطة الفصل بالآتي:

$$z = \frac{\bar{y}_1 + \bar{y}_2}{2}$$

ولو إفترضنا أن $\bar{y}_1 < \bar{y}_2$ فإن خطة التصنيف للمشاهدة الجديدة y هو أنها:

تعود للمجموعة الأولى في حالة كون $y \le z$

تعود للمجموعة الثانية في حالة كون $y > z$

أو يمكن إستخدام w أعلاه بالكامل لغرض التصنيف للمشاهدة x بكونها:

تعود للمجموعة الأولى في حالة كون $w \le 0$

تعود للمجموعة الثانية في حالة كون $w > 0$

وجدير بالذكر أن المصفوفة S تستخرج بالشكل التالي (مفترضين n = 3):

$$S = \begin{bmatrix} V_{11} & V_{12} & V_{13} \\ & V_{22} & V_{23} \\ & & V_{33} \end{bmatrix}$$

حيث أن:

$$V_{ii} = \frac{S_{ii}(1) + S_{ii}(2)}{n_1 + n_2 - 2} \quad , \quad V_{ij} = \frac{S_{ij}(1) + S_{ij}(2)}{n_1 + n_2 - 2}$$
$$S_{ii} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \quad , \quad S_{ij} = \sum x_i x_j - \frac{\left(\sum x_i\right)\left(\sum x_j\right)}{n}$$

الإختبارات المستخدمة في التحليل المميز

مع أن كفاءة الدالة المميزة الخطية تقاس وفقاً لنسبة التصنيف الصحيح للمشاهدات حسب مجاميعها الأصلية، إلا أنه من الممكن التنبؤ مسبقاً بشئ عن هذه الكفاءة لأنها تزداد وفقاً لزيادة الفرق ما بين أوساط المجاميع من جهة، وتقارب قيم مصفوفات التباين – التباين المشترك لهذه المجاميع من جهة أخرى. وفي لغة الإحصاء ومفهومه، فإن ذلك يعني تطبيق إختبارات إحصائية ونستعين بنتائجها لتقييم كفاءة الدالة المميزة الخطية. وفيما يلي الإختبارات المناسبة في هذا الجانب.

1. اختبارمعنوية الفرق بين الأوساط عن طريق إختبار الفرضية:

$$H_0 : \underline{\mu}_1 = \underline{\mu}_2$$
$$H_1 : \underline{\mu}_1 \ne \underline{\mu}_2$$

ويتم الإختبار بإستخدام إختبار F الذي يعتمد على إحصاءة هوتلنك Hotteling T^2 والتي تكون:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2$$

حيث أن D^2 هي إحصاءة مهلونوبيس Mahalanobis والتي صيغتها

$$D^2 = (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)' S^{-1} (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)$$

وبالتالي، فإن إختبار F سيكون:

$$F = \frac{n_1 + n_2 - p - 1}{n_1 + n_2 - 2} T^2 \sim f_{p,\, n_1 + n_2 - p - 1}$$

والذي نريد نتيجته الرفض بمعنوية عالية لعكس زيادة نسبة التصنيف الصحيح. حيث:

n_1 : حجم العينة الأولى

n_2 : حجم العينة الثانية

P : عدد المتغيرات

2. إختبار تساوي مصفوفتي التباين- التباين المشترك للمجموعات من خلال الفرضية:

$$H_0 : \Sigma_1 = \Sigma_2$$
$$H_1 : \Sigma_1 \ne \Sigma_2$$

وأن إحصاءة الإختبار لهذه الفرضية Q وهي:

$$Q = MC^* \sim \chi^2_{(k-1)(p-1)}$$

حيث أن:

$$M = \ln \frac{|S|^{n_1 + n_2}}{|S_1|^{n_1} |S_2|^{n_2}}$$
$$\text{or} \quad M = \sum_{i=1}^{2} n_i \ln|S| - \sum_{i=1}^{2} n_i \ln|S_i|$$
$$S = \frac{A_1 + A_2}{n_1 + n_2 - 2}$$
$$A_1 = (n_1 - 1) S_1$$
$$A_2 = (n_2 - 1) S_2$$

وان إحصاءة الإختبار لهذه الفرضية Q وهي:

$$Q = MC^{*} \sim \chi^{2}_{(k-1)(p-1)}$$

حيث أن:

$$M = \ln \frac{|S|^{n_1+n_2}}{|S_1|^{n_1}|S_2|^{n_2}}$$

$$\text{or} \quad M = \sum_{i=1}^{2} n_i \ln|S| - \sum_{i=1}^{2} n_i \ln|S_i|$$

$$S = \frac{A_1 + A_2}{n_1 + n_2 - 2}$$

$$A_1 = (n_1 - 1)S_1$$

$$A_2 = (n_2 - 1)S_2$$

$$C^{*} = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\left[\sum \frac{1}{n_i} - \frac{1}{\sum n_i}\right]$$

وأن:

k = عدد المجاميع

p = عدد المتغيرات

إن الفرضية أعلاه والتي تنص على تساوي مصفوفتي التباين - التباين المشترك بين المجموعتين هي أهم فرضية هنا لأن عدم تحققها يعني أننا لا يمكن أن نستخدم الدالة التمييزية الخطية وإنما نستخدم الدالة التمييزية التربيعية.

إحتمال خطأ التصنيف: the probability of misclassification

هو إحتمال تصنيف مشاهدة معينة إلى المجموعة الأولى بينما هي تعود في الحقيقة إلى المجموعة الثانية و بالعكس. نفترض لحساب خطأ التصنيف أن حجم العينة يكون كبير لذلك فإننا نضمن كون توزيع المشاهدات يقترب من التوزيع الطبيعي (حسب نظرية الحد المركزي). حيث

أن هذا الخطأ يعتمد على أن توزيع العينة هو التوزيع الطبيعي أو يقترب من التوزيع الطبيعي. هذا الإحتمال يكون:

$$P_{12} = P(\text{classifying x to be from group(1)/ x is from group(2)})$$
$$= \phi(-D/2)$$

حيث D^2 هي إحصاءة مهالونوبيس.

ويتم ايجاد هذه القيمة من جداول التوزيع الطبيعي القياسي. إن خطأ التصنيف هو عامل مهم لإثبات كفاءة الدالة المميزة. والتي تعطي أقل خطأ تصنيف هي الدالة الأكثر كفاءة و تكون الأفضل من بين دوال التمييز.

وبالإمكان أيضاً إستخدام طريقة إعادة التعويض في هذا الجانب وكما هو في أدناه.

طريقة التعويض: Resubstitution Method

تسخدم هذه الطريقة لإيجاد إحتمال خطأ التصنيف وأن إسلوب هذه الطريقة يعتمد على أنه لو كان n_j يمثل عدد المشاهدات التي تعود للمجموعة (j) وأن n_{ij} هو عدد المشاهدات في المجموعة (j) وصنفت وفق دالة التمييز على انها تعود للمجموعة (i)، فإن تقدير إحتمال خطأ التصنيف في هذه الحالة وبشكلٍ عام سيكون:

$$P_{ij} = \frac{n_{ij}}{n_j}$$

وفي حالة المجموعتين التي نحن فيها الأن، فإن إحتمال خطأ التصنيف الكلي للدالة المميزة سيكون:

$$\hat{p} = \frac{n_{12} + n_{21}}{n_1 + n_2}$$

أما متوسط إحتمال خطأ التصنيف سيكون:

$$\hat{p} = \frac{\hat{p}_{21} + \hat{p}_{12}}{2}$$

حالة عدة مجاميع

في حالة كون مسألة التمييز بين أكثر من مجموعتين (k من المجاميع)، يتم التصنيف عن طريق المقارنة بين كل مجموعتين وتكون لذلك عدة دوال مميزة y_{ij} وعددها C_2^k وتكتب على النحو التالي:

$$y_{ij} = \underline{x}' S^{-1}(\underline{\bar{x}}_i - \underline{\bar{x}}_j)$$

ولو كانت لدينا ثلاثة مجاميع ولكل مجموعة p من المتغيرات ($p \geq 2$) فإنه من المناسب حساب:

$$\underline{C_1} = S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_2})$$
$$\underline{C_2} = S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_3})$$
$$\underline{C_3} = S^{-1}(\underline{\bar{x}_2} - \underline{\bar{x}_3})$$

وبذلك يكون لدينا دوال التمييز التالية:

$$Y_{12} = \underline{x}'\underline{C_1} \quad , \; Y_{13} = \underline{x}'\underline{C_2} \quad , \quad Y_{23} = \underline{x}'\underline{C_3}$$

ثم نجد ما يلي:

$$W_{12} = \left[\underline{x}'S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_2}) - \frac{1}{2}(\underline{\bar{x}_1} + \underline{\bar{x}_2})' S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_2})\right]$$
$$W_{13} = \left[\underline{x}'S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_3}) - \frac{1}{2}(\underline{\bar{x}_1} + \underline{\bar{x}_3})' S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_3})\right]$$
$$W_{23} = \left[\underline{x}'S^{-1}(\underline{\bar{x}_2} - \underline{\bar{x}_3}) - \frac{1}{2}(\underline{\bar{x}_2} + \underline{\bar{x}_3})' S^{-1}(\underline{\bar{x}_2} - \underline{\bar{x}_3})\right]$$

فتكون العلاقة بينها:

$$W_{23} = W_{13} - W_{12}$$

وتكون قاعدة التصنيف إذا كانت لدينا عدد المتغيرات $(p \geq 2)$ لكل مجموعة حسب الآتي:

تصنف المشاهدة x لواحدة من المجاميع التالية:

- ضمن المجموعة (1) إذا كانت $W_{12} > 0$ و $W_{13} > 0$
- ضمن المجموعة (2) إذا كانت $W_{13} > 0$ و $W_{13} > W_{12}$
- ضمن المجموعة (3) إذا كانت $W_{13} > 0$ و $W_{12} > W_{13}$

أما إذا كان لدينا متغير وحيد لكل مجموعة (p=1) وأن أوساط المجاميع رتبت للسهولة بالشكل التالي $\bar{X}_1 < \bar{X}_2 < \bar{X}_3$ فإن قواعد التصنيف للمشاهدة x تكون كالاتي:

- ضمن المجموعة (1) إذا كانت $x < \frac{1}{2}(\underline{\bar{x}_1} + \underline{\bar{x}_2})$
- ضمن المجموعة (2) إذا كانت $\frac{1}{2}(\underline{\bar{x}_1} + \underline{\bar{x}_2}) \leq x \leq \frac{1}{2}(\underline{\bar{x}_2} + \underline{\bar{x}_3})$
- ضمن المجموعة (3) إذا كانت $x > \frac{1}{2}(\underline{\bar{x}_1} + \underline{\bar{x}_2})$

2. دالة التميز التربيعية Quadratic Discriminant Function

وكما ذكرنا سابقاً، يستخدم هذا النوع من الدوال في حالة عدم تساوي مصفوفة التباين - التباين المشترك بين المجموعات. و تقدر معالم هذه الدالة بطريقة MLE بافتراض أن حجم العينة كبير بحيث يصبح من الممكن أن نفترض بأن المشاهدات تقترب من التوزيع الطبيعي (النظرية المركزية).

وفي حالة المجتمعين، فإن تقدير μ_1 هو $\underline{\bar{X}_1}$ و أن تقدير μ_2 هو $\underline{\bar{X}_2}$ وتقدير Σ_1 هو S_1 وتقدير Σ_2 هو S_2. وأن مقياس التمييز هو V حيث

$$V = \frac{f_1(x_i)}{f_2(x_i)} > or < 1$$

$$G = \ln V = \ln f_1(x_i) - \ln f_2(x_i) > or < 0$$

ملاحظة:

في حالة وجود p من المتغيرات لكل مجموعة فإن G هذه بطبيعة الحال ستكون:

$$G = \ln f_1(x_1, x_2, \ldots, x_p) - \ln f_2(x_1, x_2, \ldots, x_p)$$

ووفقاً لما جاء في أعلاه، فإن دالة التمييز التربيعية التقديرية في حالة متغيرين ستكون:

$$\hat{G} = \frac{1}{2}\ln\frac{S_2}{S_1} - \frac{1}{2}\left(\underline{\bar{x}_1}' S_1^{-1}\underline{\bar{x}_1} - \underline{\bar{x}_2}' S_2^{-1}\underline{\bar{x}_2}\right) + \underline{x}'\left(S_1^{-1}\underline{\bar{x}_1} - S_2^{-1}\underline{\bar{x}_2}\right) - \frac{1}{2}\underline{x}'\left(S_1^{-1} - S_2^{-1}\right)\underline{x}$$

وبالتالي، فإن التصنيف للمشاهدة x سيكون

- تعود للمجموعة الأولى إذا كانت $\hat{G} > 0$
- تعود للمجموعة الثانية إذا كانت $\hat{G} < 0$

3. دالة الإنحدار اللوجستية للتمييز Logistic Regression Discriminant Function

تستخدم في حالة كون توزيع البيانات غير التوزيع الطبيعي. أي عندما يكون توزيع المتغيرات التوضيحية من عائلة التوزيع الأسي. وتعتمد هذه الطريقة على الاحتمالات السابقة واللاحقة لمجتمع (Y_2, Y_1) وتكون دالة التمييز التالية:

$$Z = \ln \frac{p(x \mid Y_1)}{p(x \mid Y_2)}$$

أما قاعدة التصنيف للمشاهدة x سيكون:

- تعود للمجموعة الأولى Y_1 إذا كانت $Z > 0$
- تعود للمجموعة الثانية Y_2 إذا كانت $Z < 0$

بعض الطرق اللامعلمية

تستخدم عندما يكون التوزيع غير طبيعي أو تكون هناك قيم شاذة مثل:

1. طريقة الرتب
2. الطرق اللا معلمية باستخدام التقديرات الحصينة (طريقة HUBER)
3. طريقة الدمج مابين طريقة الرتب و طريقة HUBER

طريقة الرتب

تستخدم بغض النظر عن توزيع البيانات سواء كان طبيعي أو غير طبيعي ويلجأ اليها الباحث عند عدم توفر الفرضيات الخاصة بالدالة المميزة فيتم إستخدام تحويلات الرتب للبيانات الاصلية.أي إستبدال البيانات الاصلية برتبها وبعدها يتم تطبيق الطرق السابقة (الخطية أو التربيعية).

1. طريقة HUBER

يتم إستبدال البيانات الأصلية ببيانات مشذبة. ولغرض تشذيب البيانات يتم حساب المتوسطات ومصفوفة التباين- التباين المشترك حيث يتم إستبدال $\bar{X}$ و S باوزان جديدة تعتمد على إحصاءة مهالونوبيس (D_i) وحسب الآتي:

$$\bar{X}^r = \frac{\sum w_i x_i}{\sum w_i}$$

$$S^r = \frac{\sum w_i^2 (x_i - \bar{X}^r)(x_i - \bar{X}^r)'}{\sum w_i^2}$$

$$w_i = \begin{cases} \frac{2}{D_i} & if\ D_i > 2 \\ 1 & if\ D_i \le 2 \end{cases}$$

دراسات سابقة:

- دراسة رومبل (Rempel ,1958):

عنوان الدراسة: استخدام التحليل الإحصائي المتعدد لاختبار مينيسوتا المتعدد الأوجه للشخصية في تصنيف الطلبة إلى الفتيان المحتمل جنوحهم وغير المحتمل جنوحهم.

The use of multivariate statistical analysis of Minnesota Multiphasic Personality Inventory scoters in the classification of delinquent and no delinquent high school boys.

هدف الدراسة: هدفت الدراسة إلى تصنيف طلبة الصف التاسع إلى الفتيان المحتمل جنوحهم وغير المحتمل جنوحهم.

عينة الدراسة: تكونت عينة الدراسة من 241 طالبا وطالبة من طلاب المدارس الرسمية.

أداة الدراسة: اختبار مينيسوتا المتعدد الأوجه للشخصية MMPI.

نتائج الدراسة: أظهرت نتائج الدراسة جدوى تطبيق الأساليب الإحصائية لتحليل التباين المتعدد للمتغيرات على اختبار مينيسوتا المتعدد الأوجه للشخصية ومع البيانات المدرسية، وقد أثبتت التقنيات المستخدمة أنها فعالة إلى حد 62.3 في المائة من غير الجانحين و69.5 في المائة من الجانحين تم التعرف بصورة صحيحة عن طريق استخدام تحليل التباين المتعدد المراحل. في حين أن مزيج من بيانات السجلات المدرسية مع نتائج الاختبار مكنت من التحديد الصحيح ل. 74.2 في المائة من الفتيان غير الجانحين و67.5 في المائة من الأولاد عرضة للجنوح.

- دراسة فراي (Frye 1998 ,):

عنوان الدراسة: تقييم الطلاب المراهقين باضطرابات انفعالية خطيرة باستخدام اختبار مينيسوتا المتعدد الأوجه للشخصية.

Assessing Adolescent Students with serious emotional disturbance using the MMPI-A.

هدف الدراسة: هدفت الدراسة إلى تحديد ما إذا كان اختبار مينيسوتا المتعدد الأوجه للشخصية قادرا على التمييز بين الطلاب الذين لا يعانون مشكلات انفعالية خطيرة والطلاب الذين يعانون منها.

عينة الدراسة: تألفت عينة الدراسة من /111/ مشاركا من طلاب المدارس الرسمية الذين شخصوا بأنهم يعانون اضطرابا انفعاليا خطيرا، ومن المراهقين المقيمين في المشافي النفسية و العيادات العلاجية الداخلية، تراوحت أعمار المشاركين بين /14/ و/18/ سنة، وبلغ عدد أفراد المعيارية /38/ مراهقا، أما العينة التي سحبت من المرضى المقيمين في المشافي فقد تألفت من /37/ مشاركا، والعينة التي شخصت باضطراب انفعالي خطير تألفت من /36/ مراهقا، من أعراق مختلفة.

أداة الدراسة: اختبار مينيسوتا متعدد الأوجه للشخصية الخاص بالمراهقين.

نتائج الدراسة: بينت النتائج أن /7/ مقاييس من أصل /12/ مقياسا لم تستطع تمييز العينة التي شخصت بالأساس من أنها تعاني اضطرابا انفعاليا جديا وبين المجموعتين الثانية والثالثة (المعيارية والعيادية)، وكان التحليل التمييزي قادرا بدقة على تصنيف المراهقين ضمن مجموعاتهم المناسبة، وفي ضوء هذه النتائج فإن /5/ من مقاييس الاختبار كانت قادرة على تمييز العينة المضطربة من العينة المعيارية، فدرجات الانحراف السيكوباتي والبارانويا وعدم النضج كانت ذات دلالة هامة عند إجراء كل من التحليل الإحصائي والتمييزي للبيانات بين المجموعات.

الجانب التطبيقي

مثال1:

سُحبت عينة عشوائية مؤلفة من 12 رياضي من مجتمع طبيعي وأجرى عليهم اختباري اللياقة والكفاءة وذلك لغرض تصنيفهم إلى مهرة أوغير مهرة وكانت النتائج كما في الجدول أدناه حيث أن:

X_1 تمثل إختبار اللياقة و X_2 تمثل إختبار الكفاءة:

غير المهرة		المهرة	
X_1	X_2	X_1	X_2
57	35	60	33
59	36	61	36
59	38	64	35
61	39	63	38
63	41	65	40
65	43		
59	41		

الحل:

يجب أن نتحقق من شرطي الدالة التميزية الخطية:

1. البيانات تتوزع توزيعاً طبيعياً
2. مصفوفة التباين - التباين المشترك متساوية للمجتمعين.
   البيانات تم أخذها من المجتمع الطبيعي.
3. ومن ثم سوف نقوم بإجراء إختبار لإثبات تساوي مصفوفة التباين - التباين المشترك للمجتمعين من خلال الفرضية

$$H_0: \Sigma_1 = \Sigma_2$$
$$H_1: \Sigma_1 \neq \Sigma_2$$

وأن إحصاءة الإختبار لهذه الفرضية Q وهي:

$$Q = MC^* \sim \chi^2_{(k-1)(p-1)}$$

حيث أن:

$$M = \sum_{i=1}^{2} n_i \ln|S| - \sum_{i=1}^{2} n_i \ln|S_i|$$

$$C^* = 1 - \frac{2p^2+3p-1}{6(p+1)(k-1)}\left[\sum \frac{1}{n_i} - \frac{1}{\sum n_i}\right]$$

$$S = \begin{pmatrix} 7.92 & 5.68 \\ 5.68 & 6.29 \end{pmatrix}$$

$$S_1 = \frac{\begin{pmatrix} S_{11} & S_{12} \\ & S_{22} \end{pmatrix}}{n_1 - 1} = \begin{pmatrix} 7.3 & 4.2 \\ 4.2 & 4.3 \end{pmatrix}$$

$$S_2 = \frac{\begin{pmatrix} S_{11} & S_{12} \\ & S_{22} \end{pmatrix}}{n_2 - 1} = \begin{pmatrix} 8.33 & 6.66 \\ 6.66 & 7.61 \end{pmatrix}$$

$|S| = 17.56 \quad , \quad \ln|S| = 2.865$

$|S_1| = 13.7 \quad , \quad \ln|S_1| = 2.62$

$|S_2| = 19.035 \quad , \quad \ln|S_2| = 2.946$

$$M = 12(2.865) - [5(2.62) + 7(2.946)]$$
$$= 0.685$$

$$C^* = 1 - \frac{2(2)^2+3(2)-1}{6(2+1)(2-1)}\left[\left(\frac{1}{3}+\frac{1}{7}\right) - \frac{1}{12}\right]$$
$$= 0.8131$$

$$MC^* = (0.685)(0.8131) = 0.534$$

$$\chi^2_{1,0.05} = 3.841$$

ومن الواضح أن القيمة المحتسبة اقل بكثيرمن القيمة الجدولية:

$$H_0: \Sigma_1 = \Sigma_2$$

لذا فالقرار هو عدم رفض (قبول) الفرضية، ومعنى ذلك عدم وجود فرق معنوي بين Σ_1 و Σ_2.

نلاحظ بأن شرطي الدالة التميزية الخطية متحقق. ولإيجاد دالة التميز الخطية وهي:

علينا أن نحسب القيم $\underline{\bar{x}}_1$ و $\underline{\bar{x}}_2$ و S لكل مجموعة واحتساب النتائج بموجبها وكما يلي:

المجموعة الأولى:

$$\sum X_1 = 182 \ , \ \sum X_1^2 = 6654$$
$$\sum X_2 = 313 \ , \ \sum X_2^2 = 19611$$
$$\sum X_1 X_2 = 11410$$
$$\underline{\bar{X}}_1 = \begin{pmatrix} 36.4 \\ 62.6 \end{pmatrix}$$
$$S_{11} = \sum x_1^2 - \frac{\left(\sum x_1\right)^2}{n} = 6654 - \frac{(182)^2}{5} = 29.2$$
$$S_{22} = \sum x_2^2 - \frac{\left(\sum x_2\right)^2}{n} = 19611 - \frac{(313)^2}{5} = 17.2$$
$$S_{12} = \sum x_1 x_2 - \frac{\left(\sum x_1\right)\left(\sum x_2\right)}{n} = 11410 - \frac{(182)(313)}{5} = 16.8$$

المجموعة الثانية:

$$\sum X_1 = 273 \ , \ \sum X_1^2 = 10697$$
$$\sum X_2 = 423 \ , \ \sum X_2^2 = 25607$$
$$\sum X_1 X_2 = 16537$$
$$\underline{\bar{X}}_2 = \begin{pmatrix} 39 \\ 60.42 \end{pmatrix}$$
$$S_{11} = \sum x_1^2 - \frac{\left(\sum x_1\right)^2}{n} = 10697 - \frac{(273)^2}{7} = 50$$
$$S_{22} = \sum x_2^2 - \frac{\left(\sum x_2\right)^2}{n} = 25607 - \frac{(423)^2}{7} = 45.71$$
$$S_{12} = \sum x_1 x_2 - \frac{\left(\sum x_1\right)\left(\sum x_2\right)}{n} = 16537 - \frac{(273)(423)}{7} = 40$$

ومن هذه النتائج نستخرج القيم التالية والتي تمثل التباينات المدمجة:

$$V_{11} = \frac{S_{11}(1) + S_{11}(2)}{n_1 + n_2 - 2} = \frac{29.2 + 50}{5 + 7 - 2} = 7.92$$
$$V_{22} = \frac{S_{22}(1) + S_{22}(2)}{n_1 + n_2 - 2} = \frac{17.2 + 45.71}{5 + 7 - 2} = 6.291$$
$$V_{12} = \frac{S_{12}(1) + S_{12}(2)}{n_1 + n_2 - 2} = \frac{16.8 + 40}{5 + 7 - 2} = 5.68$$

$$M = 12(2.865) - [5(2.62) + 7(2.946)]$$
$$= 0.685$$
$$C^* = 1 - \frac{2(2)^2 + 3(2) - 1}{6(2+1)(2-1)}\left[\left(\frac{1}{3} + \frac{1}{7}\right) - \frac{1}{12}\right]$$
$$= 0.8131$$
$$M\,C^* = (0.685)(0.8131) = 0.534$$

$$\chi^2_{1,0.05} = 3.841$$

ومن الواضح أن القيمة المحتسبة اقل بكثير من القيمة الجدولية:

$$H_0 : \Sigma_1 = \Sigma_2$$

لذا فالقرار هو عدم رفض (قبول) الفرضية، ومعنى ذلك عدم وجود فرق معنوي بين Σ_1 و Σ_2.
نلاحظ بأن شرطي الدالة التميزية الخطية متحقق. ولإيجاد دالة التميز الخطية وهي:

$$y = \underline{x}' S^{-1}\left(\underline{\bar{x}}_1 - \underline{\bar{x}}_2\right)$$

وبالتالي فإن مصفوفة التباين الكلي تكون:

$$|S| = 17.5623$$

$$S^{-1} = \frac{adj(S)}{|S|} = \frac{\begin{pmatrix} 6.291 & -5.68 \\ -5.68 & 7.92 \end{pmatrix}}{17.5623} = \begin{pmatrix} 0.358 & -0.323 \\ -0.323 & 0.451 \end{pmatrix}$$

الدالة المميزة

$$y = \underline{x}' S^{-1}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2) = \underline{x}' C^*$$

$$C^* = S^{-1}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)$$

$$= \begin{pmatrix} 0.358 & -0.323 \\ -0.323 & 0.451 \end{pmatrix}\begin{pmatrix} -2.6 \\ 2.18 \end{pmatrix} = \begin{pmatrix} -1.634 \\ 1.822 \end{pmatrix} = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}$$

$$y = (x_1 \quad x_2)\begin{pmatrix} -1.634 \\ 1.822 \end{pmatrix}$$

$$y = -1.634x_1 + 1.822x_2$$

نقطة الفصل

$$\bar{y}_1 = \underline{\bar{x}}_1' S^{-1}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2) = (36.4 \quad 62.6)\begin{pmatrix} -1.634 \\ 1.822 \end{pmatrix} = 54.5796$$

$$\bar{y}_2 = \underline{\bar{x}}_2' S^{-1}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2) = (39 \quad 60.42)\begin{pmatrix} -1.634 \\ 1.822 \end{pmatrix} = 46.359$$

$$z = \frac{\bar{y}_1 + \bar{y}_2}{2} = \frac{54.5796 + 46.359}{2} = 50.469$$

قاعدة التصنيف

المشاهدة x تعود للمجتمع الأول إذا كانت $y - z > 0$

المشاهدة x تعود للمجتمع الثاني إذا كانت $y - z \leq 0$

فلو افترضنا المشاهدة $\underline{X} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ لنرى لأي مجتمع تعود، وهنا يجب أن نجد قيمة الدالة المميزة:

$$y = -1.634x_1 + 1.822x_2$$

$$= -1.634(1) + 1.822(1) = 0.188$$

$$y - z = 0.188 - 50.469 = -50.281 < 0$$

إذا هذه المشاهدة يتم تصنيفها بأنها تعود للمجتمع الثاني.

أهمية كل متغير

بالإمكان إظهار أهمية كل متغير من خلال تطبيق المقياس الآتي:

$$C_i^* = C_i\sqrt{V_{ii}}$$

$$C_1^* = C_1\sqrt{V_{11}} = -1.634\sqrt{7.92} = -4.598$$

$$C_2^* = C_2\sqrt{V_{22}} = 1.822\sqrt{6.291} = 4.569$$

ونرى بأن المتغيرين لهما نفس الأهمية إذ الفرق قليل جداً.

إيجاد خطأ التصنيف

$$D^2 = (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)' S^{-1}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)$$

$$= (-2.6 \quad 2.18)\begin{pmatrix} 0.358 & -0.323 \\ -0.323 & 0.451 \end{pmatrix}\begin{pmatrix} -2.6 \\ 2.18 \end{pmatrix} = 8.22$$

$$2.867 = D$$

وبذلك فإن إحتمال خطأ التصنيف أعلاه هو صغير جداً ويعطي إنطباعاً عن كفاءة عالية لدالة التمييز.

وبذلك فإن إحتمال خطأ التصنيف أعلاه هو صغير جدا ويعطي إنطباعاً عن كفاءة عالية لدالة التمييز.

وللوقوف على مستوى هذه الكفاءة من جانب إستنتاجي إحصائي، نقوم بإختبار الفرضية التالية:

$$H_0 : \underline{\mu_1} = \underline{\mu_2}$$
$$H_1 : \underline{\mu_1} \neq \underline{\mu_2}$$

ويتم الإختبا بإستخدام إختبار F الذي يعتمد على إحصاءة هوتلنك Hotteling T^2 والتي تكون:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2 = \frac{(5)(7)}{5+7}(8.22) = 23.975$$

وبالتالي، فإن إختبار F سيكون:

$$F = \frac{n_1 + n_2 - p - 1}{n_1 + n_2 - 2} T^2$$
$$= \frac{5+7-2-1}{5+7-2}(23.975) = 21.577$$

الخاتمة:

لقد تمت دراسة التحليل التمييزي لاختبار سوانسون للمعالجة المعرفية وبينت نتائج الدراسة أن هناك فروق دالة إحصائيا بين التخصصات التطبيقية والنظرية في جميع الاختبارات الفرعية، لصالح التخصصات التطبيقية في التنظيم المكاني والترافق اللفظي والمتتالية غير اللفظية، ولصالح التخصصات النظرية في المتتالية السمعية الرقمية وإعادة رواية القصة والتصنيف اللفظي، كما أنه يوجد فروق دالة إحصائيا بين الذكور والإناث ذوي التخصصات التطبيقية في إعادة رواية القصة والتنظيم المكاني والمتتالية غير اللفظية، لصالح مجموعة الإناث في إعادة رواية القصة ولصالح مجموعة الذكور في التنظيم المكاني والمتتالية غير اللفظية.