الانحدار والارتباط البسيط Simple Regression and Correlation

التوزيع ذو متغيرين Bivarite Distribution ويرمز لهذين المتغيرين x و y فمثلا قد يكون x هو عدد نباتات القطن في وحدة المساحة بينما المتغير y هو كمية المحصول الناتج ... من ذلك يتضح بان كل فرد من افراد العينة له قياسان احدهما للمتغير x والاخر للمتغير y .

ان الغاية الرئيسية من دراسة توزيع ذو متغيرين هي:

1- لتحديد العلاقة بين المتغيرين x و y ووضعها بشكل معادلة بحيث يمكن التنبؤ من x بدلالة y (وهذا ما يسمى بالانحدار Regression).
2- لقياس درجة العلاقة بين المتغيرين (x و y) وهذا ما يسمى بالارتباط (Correlation) او بمعنى اخر قياس مدى التلازم او الترابط بين متغيرين مستقلين.

الانحدار الخطي البسيط Simple linear regression

هو خط مستتقيم يمر بمجموعة من النقاط تجعل مجموع مربعات النقاط المتبقية من النموذج (اي المسافات الراسية بين النقاط المتبقية والخط) اقل ما يمكن.

والنحدار بصورة عامة ينقسم الى قسمين:

1- انحدار بسيط Simple regression

ويشمل على متغيرين فقط احدهما مستقل x والاخر معتمد y .

2- انحدار متعدد Multiple regression

ويشمل على اكثر من متغيرين احدهما معتمد y والاخرى متغيرات مستقلة ($x_1, x_2, ..., x_n$)

وسوف يقتصر شرحنا هنا فقط على الانحدار الخطي البسيط اي ان المعادلة تصف العلاقة بين x و y هي معادلة خطية (linear) وتمثل بيانيا بخط مستقيم.

نفرض باننا نريد التنبؤ بالدرجات النهائية لطالب في مادة الاحصاء (y) معتمدين بذلك على معدل درجته الفصلية في الاحصاء x لذا فان (x,y) تمثل النتيجة لاي طالب في المجنمع (population).

ان عينة عشوائية من المجتمع ذات حجم n يمكن تمثيلها بالمجموعة

$$(xi\ ,\ yi)\ where\ i = 1, 2, 3, \ldots, n$$

ان قيمة y ستكون متغيرة من عينة الى اخرى لذا فان yi هي قيمة من قيم المتغير العشوائي. وللسهولة سنعني بالرمز $y|x_i$ بانه يمثل المتغير العشوائي y التابع لقيمة ثابت x_i وسنرمز لتوزيعه الاحتمالي بالرمز $f_{y|x}$. وسنرمز للوسط الحسابي لتوزيع جميع الدرجات النهائية في مادة الاحصاء عند كل معدل فصلي معلوم x بالرمز $\mu_{y|x}$ اي ان:

$$\mu_{\mathrm{yIx}} = E(y|x)$$

بينما سنرمز لتباين هذا التوزيع بالرمز $\sigma^2_{y|x}$

هذا وسنفترض بان: $y_1, y_2, \ldots, y_n$

لها توزيع طبيعي اي ان $y \sim N(\mu_{y|x}, \sigma^2_{y|x})$

وان تبايناتها متساوية اي $\sigma^2_{y|x} = \sigma^2$

لجميع قيم x

ان المعلمة $\mu_{y|x}$ هي ثابتة لاي قيمة معينة من x ولكنها تتغير للقيم المختلفة من x.

ان ما سبق ذكره يمكن تلخصه بيانيا بثلاثة قيم من x (x_1, x_2, x_3) وكما في الشكل التالي:

ان المنحني الذي يصل الاوساط الحسابية لجميع التوزيعات يسمى بمنحني الانحدار Regression curve فاذا كان هذا المنحني يمثل خطا مستقيما كما في الشكل اعلاه فان الانحدار خطي linear يمكن تمثيله بالمعادلة التالية:

$\mu_{y|x} = \alpha + \beta x$ Where α and β are constants

وتسمى هذه المعادلة بمعادلة خط الانحدار للمجتمع وهي تمثل معدل العلاقة بين x and y في المجتمع.

الشكل

ان الثابت α (a) هو معدل قيمة y عندما تكون x=0 (الشكل 333)

وتسمى نقطة تقاطع خط الانحدار بالمحور الراسي (y- Intercept) بينما الثابت β (b) فهو ميل خط انحدار المجتمع ويسمى ايضا بمعامل انحدار y على x ويعرف بانه معدل التغير في y عندما تتغير قيمة x وحدة واحدة.

$$y_x = a + bx$$

β تكون موجبة	β تكون سالبة

تقدير الثوابت او المعالم Estimation of Parameters

1- تقدير النقطة Point Estimation

ان خطالانحدار لعينة عشوائية حجمها n من المشاهدات

$$(xi\ ,\qquad yi)\ where\ i = 1, 2, 3, \ldots, n$$

يمثل بالمعادلة التالية

$$\bar{y}_x = a + bx$$

والذي هو تقدير لقيمة خط انحدار المجتمع $(\mu_{y|x})$ هذا وان كل زوج من المشاهدات $(xi\ ,\ yi)$ تتمثل بالعلاقة التالية

$$y_i = a + bx_i + e_i$$

حيث ان e_i يدعى بالخطا العشوائي

وباستخدام طريقة المربعات الصغرى Method of the least squares فيمكن ايجاد قيمة كل من a و b اللذان هما تقديران لكل من α و β على التوالي وذلك بجعل مجموع مربعات الانحرافات (عن الخط المستقيم) اقل ما يمكن اي:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (yi - a - bx_i)^2$$

وباستخدام التفاضل الجزئي بالنسبة لكل من a و b ثم جعل النلتج مساويا لصفر نحصل على المعادلاتين الطبيعيتين (Normal equations)

$$\sum yi = na + b\sum xi \text{.........(1)}$$

$$\sum xi\ yi = a\sum xi + b\sum xi^2 \text{..........(2)}$$

وبحل هاتين المعاداتين نحصل على:

$$b = \frac{\sum xi\ yi - \frac{(\sum xi)(\sum yi)}{n}}{\sum xi^2 - \frac{(\sum xi)^2}{n}}$$

$$a = \bar{y} - b\bar{x}$$

الارتباط البسيط