المادة: الاحصاء والحاسب

الفرقة: تمهيدي ماجستير

الشعبة: اعداد معلم الحاسب الآلي

استاذ المادة: أ.د محمد عبده راغب عماشة
د. رانيا عميد أبو جلاله

٤- أخطاء البيانات الإحصائية

تتعرض البيانات الإحصائية التي يتم جمعها إلى نوعين من الأخطاء:

١-٤-١ خطأ التميز

وهو ينتج عن مصادر متعددة، منها أخطاء في تصميم البحث أو التجربة أو أخطاء فنيه أثناء جمع البيانات أو خلال العمليات الحسابية التي تتم على البيانات المتجمعة.. أخطاء التميز تزداد بازدياد الفروق بين الإمكانيات (المادية والفنية) اللازم توافرها لضمان أقصى درجة دقة ممكنه وبين الإمكانيات الفعلية المتاحة للباحث.
أخطاء التميز قد توجد في البيانات التي يتم جمعها بأسلوب الحصر الشامل وقد توجد أيضاً في البيانات التي يتم جمعها بأسلوب المعاينة، ولكنها إن وجدت فهي غالباً أكبر في الحالة الأولى (الحصر الشامل) مما هي عليه في الحالة الثانية (المعاينة) باعتبار أن حجم العمل في تلك الحالة يكون أقل وبالتالي قد يسهل توفير الإمكانيات اللازمة وتجنب الأخطاء الفنية.

١-٤-٢ خطأ المعاينة العشوائية أو خطأ الصدفة

وهو الخطأ الناتج عن فروق الصدفة بين مفردات المجتمع التي دخلت العينة وبين تلك المفردات التي لم تشأ الصدفة أن تدخل العينة.. وخطأ الصدفة يمكن تقليل قيمته إذا ما تم اختيار العينة بالطريقة المناسبة وإذا ما كان حجم العينة مناسباً لحجم المجتمع وخصائصه.

١-٥ المعالم والإحصاءات

المقاييس الإحصائية التي تحسب من بيانات مجتمع الدراسة بأكمله يطلق عليها معالم المجتمع **(Parameters of population)**، أما المقاييس الإحصائية التي تحسب من بيانات عينه مسحوبة من مجتمع الدراسة فيطلق عليها إحصاءات **(Statistics)** ويعتبر كل إحصاء منها بمثابة تقدير أو قيمه تقديرية لمعلمة المجتمع المناظر، فيكون المتوسط الحسابي المحسوب من بيانات العينة تقدير لمعلمة المجتمع المناظرة وهي المتوسط الحسابي المسحوب منه هذه العينة وهكذا.. ويجب ألا يغيب عن الأذهان بأن حساب قيمة المتوسط الحسابي من بيانات العينة ليس هدفاً في حد ذاته ولكن وسيلة للتعرف على المتوسط الحسابي للمجتمع موضوع الدراسة.. وهكذا بالحال بالنسبة لباقي المقاييس الإحصائية التي تحسب من العينة.
للتفرقة بين المعالم والإحصاءات يجب أن نرمز لكل منها برموز تختلف عن رموز الأخرى، على سبيل المثال يرمز للمتوسط الحسابي للمجتمع بالرمز μ بينما يرمز للمتوسط الحسابي للعينة بالرمز $\bar{x}$ ، أيضاً للانحراف المعياري للمجتمع بالرمز σ بينما يرمز للانحراف المعياري للعينة بالرمز S وهكذا.

١-٥-١ توزيعات المعاينة: Sampling Distributions

نفرض أننا أخذنا عينه حجمها n من مجتمع ما ، ثم سحبنا منها بعض المقاييس الإحصائية مثل المتوسط الحسابي ، التباين ، ... فإن كل مقياس من هذه المقاييس يعتبر متغير عشوائي في ذاته يختلف من عينه إلى أخرى – هذا المتغير العشوائي يخضع لتوزيع معين – هذا التوزيع يسمى بتوزيع العينة. فمثلاً نقول أن توزيع المعاينة للمتوسط الحسابي وهو عبارة عن توزيع جميع المتوسطات الحسابية للعينات المأخوذة من نفس هذا المجتمع ذات الحجم n ، وكذلك فإن توزيع المعاينة للتباين هو توزيع جميع التباينات المحسوبة من عينات لها نفس الحجم n ومأخوذة من نفس المجتمع ، وهكذا ... .

١ـ٥ـ٢ توزيعات المعاينة للأوساط: Sampling Distributions of Means

نفرض أننا سحبنا عينه حجمها n من مجتمع لانهائي ، القيمة المتوقعة له تساوي μ والانحراف المعياري هو σ فإن المتوسط الحسابي $\bar{X}$ يخضع لتوزيع ما ، متوسط هذا التوزيع وانحرافه المعياري هما

$$\mu_{\bar{X}} = \mu \quad , \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{4-1}$$

وفي الحالة التي يكون فيها المجتمع الأصلي المسحوبة منه العينة مجتمع طبيعي (ويرمز له بالرمز $N(\mu, \sigma^2)$ فإن توزيع المعاينة للمتوسط الحسابي $\bar{X}$ يكون في هذه الحالة توزيع طبيعي أيضاً له نفس المتوسط الأصلي μ ولكن انحرافه المعياري يساوي $\sigma/\sqrt{n}$ ، أي بمعنى أن

$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad \bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \tag{4-2}$$

ومن ثم يكون

$$z = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \sim N(0,1) \tag{4-3}$$

أما إذا كان المجتمع غير طبيعي فإن $\bar{X}$ لا تخضع للتوزيع الطبيعي ولكنها تتوزع توزيع يكون قريباً من التوزيع الطبيعي لقيم n الكبيرة $(n \geq 30)$ حيث أن

$$z = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \xrightarrow[n \to \infty]{as} N(0,1) \tag{4-4}$$

وتعتبر النتيجة السابقة الهامة جداً في الإحصاء وخاصة في التطبيقات العلمية وتسمى نظرية النهاية المركزية Central Limit Theorem والتي تنص على أنه في حالة العينات الكبيرة الحجم فإن المتوسط الحسابي $\bar{X}$ يخضع للتوزيع الطبيعي بالمعاملات μ و $\frac{\sigma^2}{n}$ ، حيث أن μ, σ^2 هما متوسط وتباين المجتمع الأصلي بغض النظر عن شكل توزيع المجتمع الأصلي. ومن ثم فإنه لقيم n الكبيرة تتحقق العلاقة (4-3) بصرف النظر عن توزيع المجتمع الأصلي.

كذلك فإنه إذا كان $\bar{X}_1$ هو المتوسط الحسابي لعينه عشوائية مسحوبة من مجتمع لانهائي متوسطه هو μ_1 وانحرافه المعياري هو σ_1 ، وكان $\bar{X}_2$ هو المتوسط الحسابي لعينة عشوائية مسحوبة من مجتمع لانهائي آخر متوسط μ_2 وانحرافه المعياري σ_2 وكانت العينتين مستقلتين فإن المجموع الجبري لمتوسط العينتين يخضع لتوزيع المعاينة بالمعاملات

$$\mu_{(\bar{x} \pm \bar{x}_2)} = \mu_1 \pm \mu_2 \quad \text{and} \quad \sigma^2_{(\bar{x}_1 \pm \bar{x}_2)} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{4-5}$$

حيث n_1, n_2 هما حجم العينة الأولى والثانية.

وإذا كان المجتمعين الأصليين طبيعيين فإن $(\bar{x}_1 \pm \bar{x}_2)$ يخضع لتوزيع طبيعي أيضاً بالبارامترات المعطاة في (4-5) وعليه فإنه في هذه الحالة

$$z = \frac{(\bar{x}_1 \pm \bar{x}_2) - (\mu_1 \pm \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \tag{4-6}$$

أما إذا كان أحد المجتمعين أو كليهما لا يتوزع توزيعاً طبيعياً فإن $(\bar{x}_1 \pm \bar{x}_2)$ لا يتوزع توزيعاً طبيعياً كذلك ، ولكن لقيم n_1, n_2 الكبيرة فإنه طبقاً لنظرية النهاية المركزية السابقة فإن $(\bar{x}_1 \pm \bar{x}_2)$ يتوزع توزيعاً قريباً من التوزيع الطبيعي وبذلك يمكننا استخدام نفس العلاقة (4-6)في حالة العينات الكبيرة.

٣-٥-١ توزيع المعاينة للتباين: Sampling Distribution of The Variance

إذا كان $s^2 = \dfrac{\sum (x_i - \bar{x})^2}{n-1}$ هو تباين عينه عشوائية حجمها n مأخوذة من مجتمع متوسطه μ وتباينه σ^2 وعزمه الرابع حول المتوسط هو μ_4 فإن

$$\mu_{S^2} = \sigma^2 \quad \text{and} \quad \sigma^2_{S^2} = \frac{\mu_4 - \sigma^4}{n-1} \tag{4-7}$$

وإذا كان المجتمع طبيعي فإن $\mu_4 = 3\sigma^2$ وبالتالي فإن

$$\sigma^2_{s^2} = \left(\frac{2}{n-1}\right)\sigma^2 \tag{4-8}$$

نلاحظ هنا أن S^2 لا تتوزع طبيعي حتى ولو كان المجتمع طبيعي ، ولكنه يتوزع توزيع قريب من التوزيع الطبيعي وذلك لقيم n الكبيرة $(n \geq 100)$. أما إن كان المجتمع الأصلي يخضع للتوزيع الطبيعي فإن المتغير $(n-1)s^2/\sigma^2$ يخضع لتوزيع يسمى توزيع مربع كاي χ^2 بعدد درجات حرية يساوي $n-1$. أي أن

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \tag{4-9}$$

ويعتبر توزيع مربع كاي من التوزيعات الهامة في الإحصاء التطبيقي ودالة كثافته هي

$$f(y) = y^{\frac{v-1}{2}} e^{-y/2}, \quad y > 0 \tag{4-10}$$

حيث v هي عدد درجات الحريه للتوزيع وتعتبر هي المعامل الوحيد له ويتضح من شكل الدالة أنها داله متصلة وتقع بأكملها فوق النصف الموجب لمحور السينات ، منحنى هذه الداله غير متماثل ويعتبر من المنحنيات موجبة الالتواء ويقل التواؤه (وبالتالي يقترب من التماثل) كما زادت درجات الحريه v. وتكون القيمة المتوقعة لهذا التوزيع هي v و تباينه هو $2v$ أي بمعنى أن

$$E(y) = \mu_y = v$$

$$V(y) = \sigma^2 = 2v \tag{4-11}$$

فإذا كان s_1^2 هو تباين عينه عشوائية حجمها n_1 مسحوبة من مجتمع طبيعي $N(\mu_1, \sigma_1^2)$ ، وكان s_2^2 هو تباين عينه عشوائية أخرى حجمها n_2 ومسحوبة من مجتمع طبيعي آخر $N(\mu_2, \sigma_2^2)$ وكانت العينتان مستقلتان فإن المتغير:

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1) \tag{4-12}$$

حيث أن $F(n_1-1,n_2-1)$ تسمى بتوزيع F بدرجتي الحريه n_1-1 و n_2-1 و دالة الكثافة الإحتماليه للمتغير y الذي يخضع لتوزيع F بدرجتي الحريه v_1,v_2 تعطى بالصورة:

$$f(y) = \frac{y^{\frac{v_1}{2}-1}}{(v_1y+v_2)^{\frac{v_1+v_2}{2}}} \quad , \quad y>0 \tag{4-13}$$

وكما يتضح من الداله في (4-13) أن المنحنى يقع بالكامل في النصف الموجب لمحور السينات كما في حالة توزيع χ^2 ، وهو أيضاً غير متماثل وموجب الالتواء ولكن يقترب من التماثل كلما زادت درجات الحريه v_1,v_2.

ذكرنا سابقاً أنه إذا كان $\bar{X}$ هو المتوسط الحسابي لعينه حجمها n مأخوذة من مجتمع طبيعي بالمعاملات μ,σ^2 فإن

$$z=\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} \quad \sim \quad N(0,1)$$

هذا إذا كانت σ معلومة ، ولكن في حالة ما إذا كانت قيمة σ غير معلومة فإننا نستخدم بدلا منها الانحراف المعياري للعينة S ولكن في هذه الحالة يصبح المتغير $\frac{\sqrt{n}(\bar{x}-\mu)}{S}$ يخضع لتوزيع يعرف بتوزيع t ستيودنت $t-\text{student}$ بدرجات حريه $n-1$ ، أي أن

$$t=\frac{\sqrt{n}(\bar{x}-\mu)}{s} \quad \sim \quad t(n-1) \tag{4-14}$$

دالة الكثافة لتوزيع t بدرجات حريه v تعطي بالصورة:

$$f(t)= \quad (1+\frac{t^2}{v})^{-\frac{v+1}{2}} \tag{4-15}$$

وهو توزيع متماثل حول محور y وهو يشبه في ذلك المنحنى الطبيعي القياسي $N(0,1)$ ولكنه أقل تحدباً من التوزيع الطبيعي القياسي ولكنه يقترب من التوزيع الطبيعي كلما زادت درجات الحريه.

وإذا كان $\bar{X}_2$ و S_1^2 هما المتوسط الحسابي والتباين لعينه حجمها n_1 مأخوذة من مجتمع طبيعي متوسط هو μ_1 وكان $\bar{X}_2$ و S_2^2 هما المتوسط الحسابي والتباين لعينه أخرى حجمها n_2 ومأخوذة من مجتمع طبيعي آخر له المتوسط μ_2 وكانت العينتان مستقلتان فإن المتغير

$$t=\frac{(\bar{X}_2 \pm \bar{X}_2)-(\mu_1 \pm \mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \quad \sim \quad t(n_1+n_2-2) \tag{4-16}$$

حيث أن $S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ يسمى بالتباين المشترك للعينتين.The Pooled Variance