الاحتمالات

د. مصطفى بابكر

الاحتمالات (Probabilities)

- يعتبر علم الاحتمالات من أهم علوم الإحصاء وذلك لأن معظم النظريات والطرق الإحصائية بنيت من الأساس عليه. ليس ذلك فحسب بل لعلم الاحتمالات أثر كبير على حياتنا اليومية وذلك لأن كثير من القرارات الفردية والجماعية التي تتخذ يوميا تبنى على توقعات مختلفة لحدوث بعض الأشياء أو عدم حدوثها .

- الاحتمال هو مقياس لحالة عدم اليقين ويعرف على أنه قيمة رقمية بين الصفر والواحد الصحيح لتوقعات حدوث حدث معين حيث الصفر يعني استحالة حدوث الحدث والواحد يعني التيقن من حدوثه. ويرتبط حساب الاحتمالات بما يعرف في علم الإحصاء بالتجارب العشوائية (Random Experiments) والمتغيرات العشوائية (Random Variables) حيث عشوائية تعني كما أسلفنا تساوي فرص الحدوث لقيم المتغير أو التجربة.

- يعرف الاحتمال كلاسيكياً بأنه نسبة حدوث الحدث المعين إذا ما تم إجراء التجربة العشوائية لعدد كبير من المرات. أي:

$$P(A) = \frac{\text{عدد مرات حدوث } A}{\text{عدد المرات التي أعيدت فيها التجربة}}$$

ومثال لذلك احتمال ظهور رقم فردي عند رمي زهرة يساوي $\frac{3}{6}$ أو $\frac{1}{2}$ حيث تمثل 3 عدد مرات ظهور رقم فردي و 6 تمثل عدد الأرقام الممكن حدوثها عند إجراء تجربة رمي الزهرة.

- هنالك قانونين مهمين في حساب الاحتمالات هما قانون الجمع وقانون الضرب أو المضاعفة.

- تبعاً لقانون الجمع فإن احتمال حدوث أحد الحدثين A أو B في تجربة ما هو مجموع احتمالي حدوثهما منفردين ناقصا احتمال حدوثهما مجتمعين، اي:

$$P(A\ or\ B) = P(A) + P(B) - P(A,B)$$

ويسمى الحدثان A و B أحداثٍ مستبعدة (Mutually Exclusive) إذا كان حدوثهما معا غير ممكن ($P(A,B)=0$)
وفي هذه الحالة يكون قانون الجمع:

$$P(A \text{ or } B) = P(A) + P(B)$$

ويسري هذا القانون أيضاً في حالة حدوث 3 أحداث (متغيرات) أو أكثر.

- يستخدم قانون الضرب أو المضاعفة في حساب الاحتمال في حالة الأحداث غير المستبعدة أي أن احتمال حدوثها مجتمعة أكبر من الصفر.
- تبعاً لقانون الضرب أو المضاعفة فإن احتمال حدوث الحدثين A و B معا في تجربة ما هو حاصل ضرب احتمال حدوث أولهما واحتمال حدوث ثانيهما علما بحدوث الأول، أي:

$$P(A,B) = P(A) \times P(B/A)$$

حيث $P(B/A)$ يعرف بالاحتمال المشروط (Conditional Probability) ويقرأ "احتمال حدوث B علما بحدوث A" أي أن الشرط هو حدوث قبل حدوث .

أما إذا كان حدوث أحد الحدثين لا يؤثر على حدوث الآخر فيسمى الحدثين بالحدثين المستقلين وفي هذه الحالة فإن:

$$P(A,B) = P(A) \times P(B)$$

وكمثال لقانوني الجمع والضرب معاً افترض أن هنالك صندوقان بالأول 5 كرات حمراء، 5 سوداء وبالثاني 4 حمراء، 6 صفراء وافترض أن التجربة تحتوي على سحب كرة من كل صندوق فإن احتمال أن تكون إحدى الكرتين حمراء يساوي: احتمال سحب كرة حمراء من الصندوق الأول + احتمال سحب كرة حمراء من الصندوق الثاني – احتمال سحب كرة حمراء من الصندوق الأول وكرة حمراء من الصندوق الثاني، أي:

بتطبيق قانون الجمع:

$$P(R1\ or\ R2) = P(R1) + P(R2) - P(R1, R2)$$

حيث R تعني أن لون الكرة المسحوبة أحمر والرقمين 1 و 2 يعنيان الصندوق الأول والصندوق الثاني.

وبتطبيق قانون الضرب واستقلالية الحدثين:

$$P(R1\ or\ R2) = P(R1) + P(R2) - P(R1) \times P(R2)$$

وبتطبيق التعريف الكلاسيكي للاحتمال:

$$P(R1\ or\ R2) = \frac{5}{10} + \frac{4}{10} - \frac{5}{10} \times \frac{4}{10} = 0.7$$

- مما سبق نرى أن حساب الاحتمال يعتمد على حساب عدد الأحداث المتوقعة وفي الحالات البسيطة يمكن حصر هذه الأعداد بسهولة ولكنه عند زيادة عدد الأحداث أو عندما ترتبط الأحداث بمتغيرات مستمرة (Continuous Variables) فإنه من الصعب حصرها ولذلك نلجأ لاستخدام الدوال الرياضية في تعريف وحساب الاحتمال وبالتالي لفكرة التوزيعات الاحتمالية (Probability Distributions).

- عادةً ما يميز بين نوعين من الدوال الاحتمالية: دالة الكثافة الاحتمالية (Pdf) وتمثل القيم المباشرة لاحتمال حدوث الحدث عند نقطة أو مدى معين ودالة الكثافة الاحتمالية التراكمية (cdf) وتمثل مجموع القيم التراكمية في دالة الكثافة الاحتمالية لحين حدوث الحدث عند النقطة المعينة.
- أشهر التوزيعات الاحتمالية المستخدمة في وصف توزيعات المعاينة وفي الإحصاء الاستدلالي (أو الاستنتاجي) هي التوزيع الطبيعي، توزيع تي وتوزيع كاي- تربيع.

التوزيع الطبيعي (Normal Distribution)

- يعتبر أكثر التوزيعات الاحتمالية استخداماً وخاصة في مجال الإحصاء التطبيقي وذلك لأن معظم الظواهر الطبيعية لها هذا التوزيع. وأهم صفات هذا التوزيع:

أ. شكل المنحنى التكراري ناقوسي (أي أنه متماثل حول الوسط (Bell Shaped).

ب. يتحدد التوزيع تحديداً كاملاً بمعرفة وسطه الحسابي وانحرافه المعياري فقط.

ج. دالة الكثافة الاحتمالية لهذا التوزيع هي:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

حيث μ هي الوسط الحسابي للمجتمع و σ هي الانحراف المعياري. ويعبر عن هذه العلاقة بـ $x \sim N(\mu,\sigma)$ أي أن x لها توزيع طبيعي ذو وسط حسابي μ وانحراف معياري σ .

- لحساب الاحتمال يستخدم جداول التوزيع الطبيعي وهي تعطي احتمال وقوع المتغير المعياري Z بين صفر وأي قيمة أخرى، حيث:

$$Z = \frac{x - \mu}{\sigma}$$

ويعبر عن التوزيع الاحتمالي للمتغير المعياري z $Z \sim N(0,1)$ ويعرف بالتوزيع الطبيعي المعياري (Standard Normal Distribution) حيث لا يعتمد على μ أو σ .

فمثلاً لحساب احتمال أن يقع المتغير x بين قيمتين x_1 و x_2 فإننا نحسب القيم المعيارية لكل من هاتين القيمتين ثم استخدام جداول التوزيع الطبيعي بعد ذلك للحصول على الاحتمال المطلوب أي أننا نحسب $Z_2 = \frac{x_2 - \mu}{\sigma}$ و $Z_1 = \frac{x_1 - \mu}{\sigma}$ ثم بعد ذلك يصير تعبير الاحتمال:

$$P(x_1 \le x \le x_2) = P(Z_1 \le Z \le Z_2)$$

- كما يتميز التوزيع الطبيعي بالخصائص:

$$P(Z < \infty) = 0$$
$$P(Z < 0) = 0.5$$
$$1 - P(Z < Z_1) = P(Z > Z_1)$$

توزيع تي (t-Distribution)

- يشبه التوزيع تي لحد كبير التوزيع الطبيعي ويستخدم بدلاً عن التوزيع الطبيعي في حالة عدم معرفة الانحراف المعياري للمجتمع (σ) أو حين يقل حجم العينة عن 30.
- من خصائص هذا التوزيع أنه مثل التوزيع الطبيعي له منحنى تكراري متماثل حول الوسط ولكنه أكثر سُمكاً عند الأطراف من المنحنى الطبيعي. هنالك معامل واحد يحدد شكل التوزيع وهي ما يسمى بدرجات الحرية.
- تعرف درجات الحرية على أنها عدد المشاهدات المستقلة التي تستخدم في تعريف دالة الكثافة الاحتمالية للتوزيع أو في تقدير معالم التوزيع.

توزيع تي (t-Distribution)

- لحساب الاحتمال يستخدم جداول التوزيع تي والمتغير المعياري:

$$t = \frac{x - \mu}{s}$$

حيث s هي الانحراف المعياري للعينة. ولهذا التوزيع t عدد من n-1 درجات حرية.

- تنبع أهمية توزيع تي من إمكانية استعماله في حالات العينات صغيرة الحجم وهي وضع واقعي كثير الحدوث حيث إن المحددات العملية من ضيق للإمكانيات وعدمٍ توفر المِعدات المطلوبة لأخذ العينات تجعل تصغير حجم العينة أمراً ضروريا .

توزيع كاي تربيع (Chi-square X^2 Distribution)

- يختلف توزيع χ^2 عن التوزيع الطبيعي وتوزيع تي بأنه توزيع غير متماثل أو ملتوي (Skewed) وتعتمد درجة التوائه وبالتالي شكله على عدد درجات الحرية. حيث كلما زادت درِجات الحرية كلما قلت درجة التوائه وفي العينات كبيرة الحجم نسبيا يقارب شكله المنحنى الطبيعي.
- يستخدم توزيع χ^2 كثيراً في تحليل البيانات النوعية، اختبارات حسن الموافقة (Goodness of Fit) ، وتحليل التباين.
- للبيانات النوعية في شكل المصفوفة تحسب درجات الحرية لتوزيع كاي تربيع بضرب (عدد الصفوف–1) x (عدد الأعمدة–1) .

- لحساب الاحتمال يستخدم جداول التوزيع χ^2 والمتغير المعياري:

$$\chi^2 = \sum_i Z_i^2 = \sum \frac{(x_i - \mu)^2}{\sigma^2} \sim \chi^2(n)$$

حيث n هي درجات الحرية وذلك في حالة البيانات المستمرة.

أو المتغير المعياري:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(n)$$

حيث O_i التكرار المشاهد للظاهرة في العينة و E_i التكرار المتوقع للظاهرة في المجتمع.

- بعد استعراض الاحتمال والتوزيعات الاحتمالية نعود لتوزيعات المعاينة واستخدام التوزيعات الاحتمالية في قياس درجة تحيز المقدر وكفاءته.
- لوصف تحيز وكفاءة المقدر يستعان بالعزم الأول والثاني لتوزيع المقدر، حيث يمثل العزم الأول $E(\hat{\theta})$ القيمة المتوقعة للمقدر أو وسطه الحسابي والجذر التربيعي للعزم الثاني حول الوسط الحسابي يمثل الإغلاق المعياري $\sqrt{E\left(\hat{\theta}-\theta\right)^2}$.

- يعرف التحيز للمقدر $\hat{\theta}$ بـــ

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

حيث $E(\hat{\theta})$ هي القيمة المتوقعة للمقدر وتحسب باستخدام التوزيع الاحتمالي لـــ $\hat{\theta}$ بالقانون:

$$E(\hat{\theta}) = \int \hat{\theta} f(\hat{\theta}) d\hat{\theta}$$

إذا كان المقدر متغير مستمر وبالقانون:

$$E(\hat{\theta}) = \sum \hat{\theta} f(\hat{\theta})$$

إذا كان المقدر متغير نوعي وحيث $f(\hat{\theta})$ هي التوزيع الاحتمالي للمقدر. ويسمى $\hat{\theta}$ مقدر غير متحيز (Unbiased) في حالة $E(\hat{\theta}) = \theta$. أي أن المقدر غير متحيز إذا تساوى الوسط الحسابي لقيم المقدر $\hat{\theta}$ من كل العينات ذات الحجم n المسحوبة من المجتمع المعني الوسط الحسابي للمجتمع $(\hat{\theta})$.

- في حالة مقارنة مقدرين غير متحيزين فإن المقدر الأكفأ هو الأقل تبايناً. أي أن الكفاءة النسبية للمقدرات تقاس بمعدلات التباين النسبي وعليه تعرف الكفاءة النسبية للمقــدر $\hat{\theta}_1$ بالمقارنــة مع المقـدر $\hat{\theta}_2$ بالتعبير $\frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$ حيث Var تعني التباين (Variance) ويحسب تباين المقدر باستخدام التوزيع الاحتمالي للمقدر تبعاً للقانون:

$$Var(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

- يمكن استخدام مقدار الخطأ المعاينى المتوقع مسبقاً في تحديد الحجم الأمثل للعينة وذلك من خلال العلاقة العكسية بين الخطا المعياري للمقدر وحجم العينة.

فيما يلي نختم سردنا لخصائص توزيع المعاينة للمقدرات بالتركيز على الوسط الحسابي للعينة $(\bar{x})$ والنسبة (P) :

- يتميز توزيع المعاينة للوسط الحسابي $(\bar{x})$ بالخصائص التالية:

أ. $E(\bar{x}) = \mu$ أي أن الوسط الحسابي للعينة مقدر غير متحيز.

ب. $Var(\bar{x}) = \dfrac{\sigma^2}{n}$ ويعرف انحرافه المعياري (ويسمى أيضاً بالخطأ المعياري "Standard Error") بـ $Se(\bar{x}) = \dfrac{\sigma}{\sqrt{n}}$ حيث σ هي الانحراف المعياري للمجتمع.

ج. $\bar{x} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$ أي أن للوسط الحسابي للعينة توزيع طبيعي وعليه يمكن استخدام جداول التوزيع الطبيعي لحسابات الاحتمالات المتعلقة بـ $\bar{x}$. ويستخدم توزيع تي في حالة $n < 30$ أو σ غير معروفة.

- يتميز توزيع المعاينة للفرق في الأوساط الحاسبية لعينتين بالخصائص التالية:

أ. $E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$ أي أن الفرق بين المقدرين غير متحيز للفرق بين المعلمتين.

ب. $Var(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ والخطأ المعياري للفرق $Se(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1}{\sqrt{n_1}} + \frac{\sigma_2}{\sqrt{n_2}}$ حيث n_1 و n_2 حجمي العينتين و σ_1 و σ_2 الانحراف المعياري لمجتمعي العينتين.

ج. $(\bar{x}_1 - \bar{x}_2) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$ ويستخدم توزيع تي في حالة العينات الصغيرة أو عدم معرفة σ_1 و σ_2 .

- يتميز توزيع المعاينة للنسبة في العينة بالخصائص التالية:

أ. $E(\hat{P}) = P$ حيث $\hat{P}$ هي النسبة في العينة وتعطى بـ $\frac{x}{n}$ حيث x هي عدد مرات حدوث الظاهرة المعينة في العينة مثلاً عدد الأفراد ذوي الدخل المحدود . وتمثل P نسبة الظاهرة في المجتمع.

ويشير التعبير إلى عدم تحيز مقدر النسبة.

ب. $Var(\hat{P}) = \frac{Pq}{n}$ حيث $q = 1 - P$ والخطأ المعياري $Se(\hat{P}) = \sqrt{\frac{Pq}{n}}$.

ج. إذا كان حجم العينة كبير نسبياً فإن توزيع المعاينة للنسبة يقارب التوزيع الطبيعي، أي:

$$\hat{P} \approx N\left(P, \frac{Pq}{n}\right)$$

وعليه يمكن استخدام جداول التوزيع الطبيعي لحساب الاحتمالات المتعلقة بـــ $\hat{P}$ في هذه الحالة. أما في حالة العينات صغيرة الحجم فيمكن استخدام التوزيع χ^2 .

- يتميز توزيع المعاينة للفرق في النسب بين عينتين بالخصائص التالية:

أ. $E(\hat{P}_1 - \hat{P}_2) = P_1 - P_2$ أي مقدار الفرق غير متحيز.

ب. $Var(\hat{P}_1 - \hat{P}_2) = \frac{P_1 q_1}{n_1} + \frac{P_2 q_2}{n_2}$ والخطأ المعياري $Se(\hat{P}_1 - \hat{P}_2) = \sqrt{\frac{P_1 q_1}{n_1}} + \sqrt{\frac{P_2 q_2}{n_2}}$

ج. إذا كان حجم العينتين كبير نسبياً فإن توزيع للفرق بين النسبتين يقارب التوزيع الطبيعي ، أي:

$$(\hat{P}_1 - \hat{P}_2) \approx N\left(P_1 - P_2, \frac{P_1 q_1}{n_1} + \frac{P_2 q_2}{n_2}\right)$$