

Дәріс 8. Корреляциялық және регрессиялық талдау

Корреляция – екі айнымалы арасындағы байланыс дәрежесін анықтау. Пирсон және Спирмен коэффициенттері. Регрессия – бір айнымалының басқасына тәуелділігін сипаттайтын теңдеу. Мысал: температура мен фотосинтез жылдамдығының арасындағы байланыс.

Биологиялық жүйелер мен процестердің көпшілігі бір-бірімен тығыз байланыста және өзара тәуелді болып келеді. Өсімдіктердің өсуі, дамуы мен өнімділігі, жануарлардың мінез-құлқы мен физиологиялық реакциялары, популяциялардың саны мен экологиялық жағдайлар – бұлардың барлығы түрлі факторлардың әсерімен өзгереді. Осы факторлар арасындағы өзара байланыстарды сандық тұрғыда анықтау және олардың дәрежесін сипаттау үшін биостатистикада корреляциялық және регрессиялық талдау әдістері кеңінен қолданылады. Бұл әдістер зерттелетін айнымалылар арасындағы тәуелділікті анықтап, бірінің өзгерісі екіншісіне қалай әсер ететінін есептеуге мүмкіндік береді.

Корреляция – екі немесе одан да көп айнымалылардың өзара байланысын сипаттайтын статистикалық көрсеткіш. Егер бір айнымалы өзгергенде екіншісі де белгілі бір бағытта өзгерсе, онда олардың арасында корреляциялық байланыс бар деп саналады. Мысалы, температура артқанда фотосинтез қарқыны да өссе, бұл оң корреляция; ал температура артқанда тұқымның өну пайызы азайса, бұл теріс корреляция болып есептеледі. Егер бір айнымалының өзгерісі екіншісіне әсер етпесе, корреляция жоқ деп айтылады. Корреляция байланысының сандық көрсеткіші – корреляция коэффициенті (r). Ол -1 мен $+1$ аралығында мән қабылдайды: егер $r = +1$ болса, айнымалылар арасында толық оң байланыс бар; егер $r = -1$ болса, толық теріс байланыс бар; ал $r = 0$ болса, айнымалылар тәуелсіз дегенді білдіреді. Мысалы, өсімдіктің биіктігі мен жапырақ саны арасындағы $r = 0,85$ болса, бұл өте күшті оң байланыс, ал тұздылық деңгейі мен өну пайызы арасындағы $r = -0,7$ болса, күшті теріс байланыс бар екенін көрсетеді.

Корреляция себеп-салдарлық байланысты білдірмейді, ол тек екі айнымалының бірге өзгеру дәрежесін сипаттайды. Мысалы, күн сәулесінің ұзақтығы мен температура арасында күшті оң корреляция болуы мүмкін, бірақ бұл температура күн ұзақтығының нәтижесі дегенді білдірмейді. Сондықтан корреляциялық талдау көбіне алғашқы зерттеу сатысында, яғни факторлар арасындағы ықтимал байланыстарды анықтау үшін қолданылады.

Корреляцияны есептеу әдісі деректердің сипатына байланысты таңдалады. Егер айнымалылардың үлестірілімі қалыпты болса және олардың арасындағы байланыс сызықтық сипатта болса, онда Пирсонның корреляция коэффициенті қолданылады. Бұл коэффициент екі айнымалы арасындағы сызықтық байланыстың дәрежесін анықтайды және ол келесі формуламен есептеледі:

$$r = \frac{\Sigma[(x - \bar{x})(y - \bar{y})]}{\sqrt{[\Sigma(x - \bar{x})^2 \times \Sigma(y - \bar{y})^2]}}$$

Мұндағы x және y – айнымалылар мәндері, ал \bar{x} және \bar{y} – олардың орташа шамалары. Егер алынған r оң болса, айнымалылар бір бағытта өзгереді; ал теріс болса – кері бағытта өзгереді.

Егер деректер қалыпты үлестірілмеген болса немесе байланыс сызықтық емес сипатта болса, онда Спирменнің рангілі корреляция коэффициенті (ρ) қолданылады. Бұл әдісте айнымалылардың нақты мәндері емес, олардың реттік орындары (рангтері) пайдаланылады. Спирмен корреляциясы экологиялық және биологиялық зерттеулерде жиі қолданылады, себебі табиғи деректердің басым бөлігі қалыпты үлестірілмейді және сызықтық емес сипатта болады.

Корреляция коэффициентінің мәнін интерпретациялау кезінде оның шамасы байланыс күшін сипаттайды. Егер r немесе ρ шамасы 0 мен 0,3 аралығында болса – әлсіз байланыс, 0,3 пен 0,6 аралығында болса – орташа байланыс, 0,6 пен 0,8 аралығында болса – күшті, ал 0,8 пен 1,0 аралығында болса – өте күшті байланыс бар деп есептеледі. Мысалы, температура мен фотосинтез арасындағы $r = 0,83$ болса, бұл екі айнымалының арасында өте күшті оң байланыс бар екенін білдіреді.

Регрессиялық талдау корреляциядан айырмашылығы, бір айнымалының (тәуелді айнымалы – y) басқа айнымалыға (тәуелсіз айнымалы – x) нақты түрде тәуелділігін анықтайды. Егер корреляция «байланыс бар ма, жоқ па?» деген сұраққа жауап берсе, регрессия «егер бір айнымалы өзгерсе, екіншісі қаншалықты өзгереді?» деген сұраққа жауап береді. Регрессиялық талдау нәтижесінде айнымалылар арасындағы тәуелділік математикалық теңдеу түрінде сипатталады, бұл теңдеу арқылы болжау (прогноз) жасауға болады.

Ең қарапайым түрі – сызықтық регрессия, ол $y = a + bx$ теңдеуі арқылы өрнектеледі. Мұндағы a – регрессияның бос мүшесі, яғни $x = 0$ болғандағы y мәні; b – регрессия коэффициенті, яғни x бір бірлікке өзгергенде y қаншалықты өзгеретінін көрсететін сан. Егер b оң болса, онда x артқанда y да артады, яғни байланыс оң; ал егер b теріс болса, онда x артқанда y кемиді, яғни байланыс кері бағытта. Мысалы, теңдеу $y = 2 + 0,4x$ болса, бұл x шамасы (мысалы, температура) 1°C артқанда y шамасы (мысалы, фотосинтез жылдамдығы) 0,4 бірлікке өседі дегенді білдіреді.

Биологияда регрессиялық талдау арқылы көптеген процестерді модельдеуге болады. Мысалы, өсімдіктің биомассасының өсуін уақытқа тәуелді түрде сипаттау, жануарлардың салмағын жасына қарай болжау, фотосинтездің қарқындылығын температура немесе жарық деңгейіне қарай есептеу сияқты тапсырмаларда регрессия тиімді қолданылады. Егер бірнеше фактор әсер етсе (мысалы, температура, ылғал және жарық), онда көптік регрессия (multiple regression) әдісі қолданылады, ол бірнеше айнымалының бір мезгілде әсерін бағалауға мүмкіндік береді.

Мысал ретінде фотосинтез жылдамдығы мен температура арасындағы байланысты қарастырайық. 10°C -тан 35°C -қа дейінгі аралықта температура артқан сайын фотосинтез де күшейіп, шамамен 25°C маңында ең жоғары

деңгейге жетіп, содан кейін төмендей бастайды. Бұл деректер бойынша Пирсон корреляциясы $r = 0,82$, яғни күшті оң байланыс бар екені анықталды. Бірақ жоғары температура кезінде фотосинтез төмендейтін болғандықтан, бұл байланыс толық сызықтық емес. Осы жағдайда Спирмен коэффициенті есептеліп, $\rho = 0,79$ мәні алынды. Регрессиялық талдау нәтижесінде фотосинтез жылдамдығы мен температура арасындағы тәуелділік $y = -0,06x^2 + 2,5x - 8$ теңдеуімен сипатталды. Бұл теңдеу өсімдіктің фотосинтезі үшін оңтайлы температура шамамен $24\text{--}25^\circ\text{C}$ екенін көрсетті.

Корреляция мен регрессияның арасындағы негізгі айырмашылықты түсіну маңызды. Корреляция тек айнымалылардың арасындағы өзара байланыс бар екенін көрсетеді, бірақ себептілікті түсіндірмейді. Ал регрессия бір айнымалының екіншісіне нақты тәуелділігін сандық түрде сипаттайды және болашақ мәндерді болжауға мүмкіндік береді. Яғни корреляция – байланыс бар ма деген сұраққа, ал регрессия – бұл байланыс қандай және қаншалықты күшті деген сұраққа жауап береді.

Бұл әдістер биологияның барлық салаларында қолданылады. Физиологияда – фермент белсенділігі мен температураның, экологияда – популяция тығыздығы мен қорек қорының, агробиологияда – тыңайтқыш мөлшері мен өнімділіктің, ал генетикада – белгілер арасындағы тұқымқуалаушылық қатынастардың байланысын анықтау үшін пайдаланылады. Корреляция мен регрессия табиғи жүйелердің ішкі құрылымын түсінуге, факторлардың әсерін сандық тұрғыда дәлелдеуге және ғылыми болжамдар жасауға мүмкіндік береді.

Нәтижелерді интерпретациялауда зерттеуші бірнеше маңызды қағиданы есте сақтауы тиіс. Біріншіден, жоғары корреляция себептілікті білдірмейді – бір айнымалының өзгерісі екіншісіне тікелей әсер етеді деген қорытынды жасауға болмайды. Екіншіден, деректер көлемі неғұрлым үлкен болса, нәтиже соғұрлым сенімді болады. Үшіншіден, шектен тыс мәндер (outliers) нәтижені бұрмалауы мүмкін, сондықтан оларды алдын ала тексеру қажет. Төртіншіден, деректерді графикалық түрде көрсету (scatter plot, яғни нүктелік диаграмма) байланыс сипатын көзбен көруге және қателерді анықтауға мүмкіндік береді.

Қорытындылай келе, корреляциялық және регрессиялық талдаулар биостатистиканың ең маңызды құралдарының бірі болып табылады. Олар биологиялық деректер арасындағы байланыстарды анықтап қана қоймай, табиғи процестердің заңдылықтарын модельдеуге және ғылыми тұрғыда түсіндіруге мүмкіндік береді. Корреляция айнымалылар арасындағы өзара қатынасты сипаттайды, ал регрессия олардың арасындағы нақты тәуелділікті және болжамдық қатынасты көрсетеді. Осы екі әдісті дұрыс қолдану биология, экология, физиология және агрономия салаларында алынған нәтижелердің дәлдігі мен сенімділігін арттырады.

Пайдаланылған әдебиеттер:

- Zar, J. H. Biostatistical Analysis. Pearson, 2010.
- Sokal, R. R., & Rohlf, F. J. Introduction to Biostatistics. W. H. Freeman, 2012.
- Quinn, G. P., & Keough, M. J. Experimental Design and Data Analysis for Biologists. Cambridge University Press, 2002.
- Fowler, J., Cohen, L., & Jarvis, P. Practical Statistics for Field Biology. Wiley-Blackwell, 2013.
- Crawley, M. J. The R Book. Wiley, 2013.
- Magurran, A. E. Measuring Biological Diversity. Blackwell, 2004.