Дәріс 2. Биологиялық деректер мен олардың түрлері

Биологиялық мәліметтер сапалық (мысалы, жынысы, түрі) және сандық (мысалы, өсім ұзындығы, массасы) болып бөлінеді. Сандық мәліметтер: Дискретті (мысалы, тұқым саны, дара саны) Үздіксіз (мысалы, масса, ұзындық, концентрация). Деректерді жинау ережелері: репрезентативтілік, қателіктерді азайту, бақылаулар санының жеткіліктілігі.

Биологиялық деректер – тірі организмдер мен олардың ортасында жүріп жатқан процестерді сипаттайтын, бақылау мен тәжірибелер нәтижесінде алынған сандық немесе сапалық ақпарат. Бұл деректер биология, экология, физиология, генетика және биохимия сияқты көптеген ғылым салаларында қолданылады. Биологиялық деректердің басты ерекшелігі – олардың жоғары өзгергіштігі мен күрделілігі. Әрбір ағза, популяция немесе экожүйе көптеген факторлардың (генетикалық, экологиялық, климаттық, антропогендік және т.б.) әсерінен қалыптасады, сондықтан жиналған мәліметтер табиғи вариацияларды көрсетеді (Sokal & Rohlf, 2012).

Биологиялық деректерді шартты түрде екі үлкен топқа бөлуге болады: сапалық (qualitative) және сандық (quantitative). Бұл жіктеу зерттеудің мақсатына, объектісіне және алынған нәтижелердің сипаттамасына байланысты.

Сапалық деректер — объектілердің қасиеттерін, категорияларын және айырмашылықтарын сипаттайды. Олар көбінесе «иә/жоқ», «бар/жоқ» түрінде немесе категориялық (номиналды) түрде жазылады. Мысалы, өсімдіктің түрі (Berberis integerrima, Rosa laxa, Thymus serpyllum), жынысы (еркек, ұрғашы), жапырақ түсі (жасыл, қызғылт, сары), немесе тозаңның пішіні (дөңгелек, сопақ, үшкір). Бұл деректер көбіне сапалық сипаттамаларға жататындықтан, оларды статистикалық өңдеу кезінде жиілік (frequency) және пайыздық үлес (percentage) сияқты көрсеткіштер қолданылады. Егер категориялар арасында белгілі бір тәртіп бар болса (мысалы, даму кезеңдері: тұқым – өскін – гүл – жеміс), онда мұндай деректер реттік (ordinal) деректер деп аталады (Zar, 2010).

Сандық деректер — өлшенетін, нақты мәндермен көрсетілетін белгілер. Олар биологияда ең жиі қолданылады және зерттелетін процестерді сандық тұрғыдан сипаттауға мүмкіндік береді. Сандық деректер екі түрге бөлінеді: дискретті (discrete) және үздіксіз (continuous).

Дискретті деректер тек бүтін сандар түрінде болады. Мысалы, тұқым саны, гүл саны, даралар саны, микроб колонияларының саны, т.б. Мұндай деректер санауға негізделеді және аралық мән қабылдамайды (мысалы, 3,5 дара болмайды). Үздіксіз деректер, керісінше, кез келген нақты мәнді қабылдай алады және өлшеуге негізделген. Мысалы, өсімдіктің биіктігі (см), массасы (г), жапырақ ауданы ($см^2$), фотосинтез жылдамдығы (мг $CO_2/см^2$/сағ), фермент белсенділігі, ДНҚ концентрациясы және т.б. Үздіксіз деректерді талдау үшін орташа арифметикалық, дисперсия, стандартты ауытқу сияқты статистикалық көрсеткіштер есептеледі (Fisher, 1935; Magurran, 2004).

Биологиялық деректердің сапасы олардың жинау тәсіліне, өлшеу дәлдігіне және таңдалған үлгінің (sample) репрезентативтілігіне байланысты. Репрезентативтілік (representativeness) — алынған үлгінің зерттелетін популяцияның нақты қасиеттерін дәл бейнелеу дәрежесі. Егер үлгі кездейсоқ емес, біржақты таңдалса, онда ол жалпы популяцияны сипаттай алмайды, нәтижесінде зерттеу қорытындылары бұрмаланады. Мысалы, өсімдіктердің биіктігін тек көлеңкелі жерде өлшеген жағдайда, бүкіл популяция үшін орташа биіктік төмен бағалануы мүмкін. Сондықтан зерттеу дизайнында үлгілерді іріктеу кезінде кездейсоқтық принципі (randomization) және тең мүмкіндіктер ережесі (equal probability sampling) сақталуы қажет (Quinn & Keough, 2002; Fowler et al., 2013).

Деректер жинау процесінде қателіктерді азайту ерекше маңызды. Қателіктер екіге бөлінеді: жүйелі (systematic) және кездейсоқ (random). Жүйелі қателік — өлшеу құралының дұрыс калибрленбеуінен немесе әдістемедегі жүйелі ауытқудан туындайды. Кездейсоқ қателік — табиғи вариация мен сыртқы факторлардың әсерінен болады. Бұл қателіктерді азайту үшін өлшеулер бірнеше рет қайталанып, орташа мән есептеледі. Сонымен қатар, барлық өлшеулер бірдей жағдайда жүргізілуі тиіс (мысалы, температура, жарық, уақыт параметрлері бірдей болу керек) (Crawley, 2013).

Бақылаулар санының жеткіліктілігі де аса маңызды. Егер бақылаулар тым аз болса, статистикалық талдаудың сенімділігі төмендейді, ал егер тым көп болса, зерттеу артық ресурстарды талап етеді. Көптеген биологиялық зерттеулерде әрбір топ немесе емдік нұсқа бойынша кемінде 20–30 қайталау (replicate) ұсынылады. Бұл сан стандартты қателікті (standard error) төмендетіп, нәтижелердің нақтылығын арттырады (Zar, 2010).

Биологиялық деректердің жинақталуы мен ұйымдастырылуы зерттеу нәтижесін тиімді талдауға мүмкіндік береді. Деректерді жинау кезінде әр жазбаға міндетті түрде метадеректер (metadata) тіркелуі қажет: бақылау күні, орны (координаталар), зерттеуші аты, әдіс түрі, температура, ылғалдылық және басқа факторлар. Мұндай тәсіл FAIR Data Principles (Findable, Accessible, Interoperable, Reusable) стандарттарына сай келеді және кейінгі талдау мен деректер алмасуға мүмкіндік береді (GO FAIR, 2016).

Бүгінгі күні биологиялық деректердің түрлері мен көлемі өте кеңейді. Қазіргі замандағы зерттеулер тек далалық бақылаумен шектелмейді — оларды қашықтықтан зондтау (remote sensing), GIS-картография, молекулалық талдау және биоинформатикалық деректер базаларымен біріктірілген цифрлық жүйелер арқылы жинауға болады. Мысалы, өсімдіктердің фотосинтетикалық белсенділігін спутниктік Sentinel-2 немесе MODIS деректері негізінде NDVI (Normalized Difference Vegetation Index) арқылы бағалауға болады. Генетикалық зерттеулерде Illumina немесе Oxford Nanopore секвенирлеу технологиялары арқылы алынған нуклеотид тізбектері FASTQ форматында сақталып, кейін NCBI GenBank немесе ENA деректер қорларына енгізіледі. Мұндай деректер көлемі

терабайттармен өлшеніп, оларды өңдеу үшін биостатистикалық және биоинформатикалық әдістер (мысалы, PCA, AMOVA, Random Forest, Machine Learning) пайдаланылады (Hastie et al., 2021; Bolstad & Curran, 2020).

Биологиялық деректерді дұрыс классификациялау мен талдау ғылыми нәтижелердің сапасын айқындайды. Егер сапалық және сандық деректер дұрыс ажыратылмаса, статистикалық әдіс қате таңдалып, қате қорытындыға әкелуі мүмкін. Мысалы, үздіксіз деректерге χ^2 тестін қолдану немесе категориялық деректерге ANOVA қолдану жарамсыз. Сондықтан әрбір зерттеуші деректер түрлерін, олардың қасиеттерін және оларды өңдеудің тиісті статистикалық тәсілдерін жақсы білуі қажет (Fowler et al., 2013).

Осылайша, биологиялық деректер – тірі табиғаттағы заңдылықтарды сандық тұрғыда сипаттаудың негізі. Оларды дұрыс жинау мен талдау биологиялық, экологиялық және медициналық зерттеулердің дәлдігін арттырады. Биостатистикалық тұрғыдан өңделген деректер ғылыми гипотезаларды тексеруге, модель құруға және нақты шешім қабылдауға мүмкіндік береді.

Пайдаланылған әдебиеттер:

Zar, J. H. Biostatistical Analysis. Pearson, 2010.

Sokal, R. R., & Rohlf, F. J. Introduction to Biostatistics. Freeman, 2012.

Fisher, R. A. The Design of Experiments. Edinburgh: Oliver & Boyd, 1935.

Magurran, A. E. Measuring Biological Diversity. Blackwell, 2004.

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.

GO FAIR Initiative. FAIR Data Principles. 2016.

Bolstad, W. M., & Curran, J. M. Introduction to Bayesian Statistics. Wiley, 2020.

Hastie, T., Tibshirani, R., & Friedman, J. The Elements of Statistical Learning. Springer, 2021.