BEVEZETÉS

Ez a könyv a matematika egyik ága alapjainak ismertetésével foglalkozik, ezt az ágat matematikai statisztikának nevezik. Ez utóbbit a rövidség kedvéért gyakorta egyszerűen statisztikának nevezik. Ugyanakkor ügyelni kell arra, hogy ez a rövidítés csak akkor megengedett, ha félreértéstől szó sem lehet, ugyanis maga a statisztika szó rendszerint egy kicsit más fogalmat takar.

Mi is az a matematikai statisztika? Sokféle leíró „meghatározását" lehetne megadni, melyek többé-kevésbé fednék a matematika ezen ágának tartalmát. Az egyik legegyszerűbb és legdurvább az általános sokaságból történő mintavétel fogalmával, és a valószínűségszámítási kurzusok elején gyakran tárgyalt hipergeometrikus eloszlást definiáló feladattal kapcsolatos összehasonlításon alapul. Ott a véletlenül választott elemek összetételének eloszlását vizsgálják, ismerve a sokaság összetételét. Ez a tipikus valószínűségszámítási feladat. Ugyanakkor gyakran meg kell oldani a fordított feladatot is, amikor ismert a minta összetétele, és ebből kell meghatározni, hogy milyen maga a sokaság. Az ilyenfajta fordított feladatok alkotják, képletesen szólva, a matematikai statisztika tárgyát.

Kicsit pontosítva ezt az összehasonlítást, azt mondhatjuk: a valószínűségszámításban kiderítjük – ismerve bizonyos jelenségek viselkedését –, hogyan viselkednek (hogyan oszlanak meg) egy és más általunk tanulmányozott, a kísérletekben megfigyelhető jellemzők. A matematikai statisztikában éppen fordítva – a kísérleti adatok a kiinduló pont (rendszerint ezek valószínűségi változók megfigyelései), és ebből kell a vizsgált jelenség természetére vonatkozó ilyen-olyan állításokat és döntéseket levezetni. Ily módon itt az emberi tevékenység egyik legfontosabb válfajába ütközünk – a megismerés folyamatába. Az az állítás, miszerint az „igazság kritériuma a gyakorlat" a legközvetlenebb kapcsolatban van a matematikai statisztikával, mivel éppen ez a tudomány tanulmányozza azokat az eljárásokat (a pontos matematikai modellek keretein belül), amelyek lehetővé

teszik, hogy válaszolhassunk a kérdésre, megfelel-e a – kísérleti eredmények formájában jelentkező – tapasztalat a jelenség természetéről alkotott hipotézisnek, vagy sem.

Eközben feltétlenül ki kell emelni, hogy – ugyanúgy, mint a valószínűségszámítás esetében – nem azok a kísérletek érdekelnek minket, amelyek alapján a vizsgált jelenségekre vonatkozóan egyértelmű, determinisztikus következtetésekre juthatunk, hanem azok a kísérletek, amelyek eredményei véletlen események. A tudomány fejlődésével az ilyenfajta feladatok szerepe egyre nagyobb lesz, mivel a kísérletek pontosságának növelésével együtt egyre nehezebb lesz elkerülni a mérési és számítási lehetőségeink korlátaiból és nehézségeiből származó „véletlen tényezőket".

A matematikai statisztika a valószínűségszámítás része abban az értelemben, hogy minden egyes matematikai statisztika feladat lényegében (néha teljesen sajátságos) valószínűségszámítási feladat. Ugyanakkor maga a matematikai statisztika önálló helyet foglal el a tudományok rendszerében. A matematikai statisztikát úgy lehet tekinteni, mint azt a tudományágat, melynek tárgya az ember (és nemcsak az ember) olyan feltételek melletti indukciós viselkedése, amikor a saját nem determinisztikus tapasztalatai alapján kényszerül a számára legkevesebb veszteséggel járó döntést meghozni.*

A matematikai statisztikát a statisztikus döntések elméletének is nevezik, mivel úgy is lehet jellemezni, mint a statisztikus (kísérleti) adatokon alapuló optimális döntések (e két utóbbi szót meg kell magyarázni) tudománya. A feladatok pontos megfogalmazását később, a könyv főrészében fogjuk megadni. Most csak arra korlátozódunk, hogy bemutassuk a statisztikai feladatok három egyszerű és tipikus példáját.

1. példa. Sok termék esetében a minőségét jellemző alapvető paraméterek egyike az élettartama. Azonban egy termék élettartama (mondjuk egy rádiócsőé) rendszerint véletlenszerű, előre meghatározni nem lehetséges. A tapasztalat azt mutatja, hogy ha a gyártási folyamat az ismert értelemben homogén, akkor az 1., 2., ... termék $\xi_1, \xi_2, \ldots$ élettartamát független, azonos eloszlású valószínűségi változóknak kell tekinteni. A minket érdeklő paramétert, mely meghatározza az élettartamot, természetes módon azonosíthatjuk a $\theta = \mathsf{E}\xi_i$ értékkel. Az egyik standard feladat abban áll, hogy tisztázzuk, vajon mivel egyenlő θ. Ahhoz, hogy meghatározzuk ezt az értéket, vegyünk n késztermékét és ellenőrizzük őket. Legyenek $x_1, x_2, \ldots, x_n$ ezen ellenőrzött termékek élettartamai. Tudjuk, hogy

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i \xrightarrow[\text{m.m.}]{} 0,$$

ha $n \to \infty$. Ezért természetes azt várni, hogy az $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ érték elég nagy n

* Részletesebben erről lásd [56].

érték esetén közel lesz θ-hoz, és ez lehetővé teszi, hogy valamilyen mértékben feleljünk a feltett kérdésre. Eközben világos, hogy mi érdekeltek vagyunk abban, hogy a szükséges megfigyelések n száma a lehetőség szerinti legkisebb legyen, ugyanakkor a θ érték becslése pedig a lehetőség szerint minél pontosabb legyen (a θ paraméter túlságos növelése, illetve csökkentése anyagi veszteségekhez vezet).

2. példa. Egy radar a $t_1, t_2, \ldots, t_n$ időpillanatokban végigpásztázza a légtér egy adott részét abból a célból, hogy bizonyos tárgyak jelenlétét felfedje. Jelölje $x_1, x_2, \ldots, x_n$ a műszer által felfogott, visszavert jel értékét. Ha az adott térrészben nincsen számunkra érdekes objektum, akkor az x_i értékeket tekinthetjük független valószínűségi változóknak, amelyek eloszlása ugyanolyan, mint egy ξ valószínűségi változóé, amelynek viselkedése különféle zavaró tényezők természetétől függ. Ha a megfigyelési periódus folyamán valamilyen objektum található a látótérben, akkor az x_i értékek a zavarok értékeivel együtt egy a „hasznos" jelet is fognak tartalmazni, és így x_i eloszlása ugyanolyan lesz, mint $\xi + a$ eloszlása. Ily módon, ha az első esetben az x_i eloszlásfüggvénye $F(x)$, akkor a második esetben az eloszlásfüggvényük $F(x-a)$ alakú lesz. Az $x_1, x_2, \ldots, x_n$ minta alapján kell dönteni arról, hogy a két eset közül éppen melyik a helytálló, azaz létezik-e az adott helyen számunkra érdekes objektum, vagy sem.

Ebben a feladatban lehetségesnek látszik, hogy megadjunk egy bizonyos értelemben „optimális döntési szabályt", amely minimális hibával oldja meg a kitűzött feladatot. A megfogalmazott feladatot a következő módon lehet megnehezíteni. Az objektum először nincs jelen, majd a megfigyelés kezdetétől számított ismeretlen θ időpontban megjelenik. A lehető legpontosabban meg kell határozni az objektum megjelenésének θ időpontját. Ez az úgynevezett „riasztási feladat", amelynek egész sor, az alkalmazások szempontjából fontos interpretációja van.

3. példa. Valamilyen kísérletet először az „A" feltételek mellett elvégeznek n_1-szer, majd a „B" feltételek mellett n_2-ször. Jelölje $x_1, \ldots, x_{n_1}$ és $y_1, \ldots, y_{n_2}$ az A és B feltételek mellett kapott kísérleti eredményeket. Kérdés: vajon az eredmények alapján fel lehet-e ismerni a kísérleti körülmények megváltozását. Más szavakkal, ha $\mathbf{P}_A$ jelöli az x_i, $1 \leqslant i \leqslant n_1$ és $\mathbf{P}_B$ az y_i, $1 \leqslant i \leqslant n_2$ eloszlását, akkor a kérdés lényege az, hogy teljesül-e a $\mathbf{P}_A = \mathbf{P}_B$ összefüggés, vagy nem.

Ha például azt kell megállapítani, hogy valamilyen preparátum befolyásolja-e a fejlődést, mondjuk növények vagy állatok fejlődését, akkor párhuzamosan két sorozat kísérletet végeznek el (preparátum nélkül vagy azzal), és ezek eredményeit kell tudni összehasonlítani.

Gyakran fellépnek ennél bonyolultabb feladatok is, amikor az ennek megfelelő kérdést sok, különböző feltételek mellett végzett megfigyeléssorozat esetén kell feltenni. Ha a kísérletek eredménye függ a feltételektől, akkor általában meg kell vizsgálni a függőség jellemzőit is (az úgynevezett regressziós feladat).

A 3. példa és az említett bonyolultabb problémák is a *két vagy többmintás* statisztikai feladatok osztályába tartoznak. Ezeket a feladatokat egy külön könyvben fogjuk vizsgálni (lásd az Előszót).

A bonyolultsági fok és a tartalmuk szerint különböző tipikus statisztikai feladatok listáját tovább lehetne folytatni. Azonban mindegyikükben közös az alábbi két körülmény:

1. Semmilyen probléma sem lenne előttünk, ha a megfigyelések eredményeinek eloszlása, amelyek a feladatokban szerepelnek, ismertek lennének.

2. Mindegyik feladatban a kísérletek eredményei alapján kell a megfigyelések eloszlásaira vonatkozó valamiféle döntést hozni (innen származik a korábban már említett „Statisztikus döntések elmélete").

Ezzel a két megjegyzéssel összefüggésben minden további és speciálisan a példaként említett feladatokban is alapvető jelentősége van a következő ténynek. A ξ valószínűségi változó $x_1, \ldots, x_n$ megfigyelései alapján nagy n értékek esetén tetszőleges pontossággal helyre lehet állítani a szóban forgó valószínűségi változó ismeretlen $\mathbf{P}$ eloszlását. Ugyanez az állítás igaz az ismeretlen eloszlás tetszőleges $\theta = \theta(\mathbf{P})$ funkcionáljára.

Ez a tény a matematikai statisztika alapköve. Erről, illetve még pontosabb állításokról szól az 1. Fejezet.