

# Matematikai statisztika összefoglaló zárószigorlatra

Dr. Tómacs Tibor  
Eszterházy Károly Egyetem  
Matematikai és Informatikai Intézet

## Statisztikai mező és minta

A valószínűségszámításban tárgyalt feladatokban mindig szerepel valamilyen információ bizonyos típusú véletlen események valószínűségére vonatkozóan. Például:

- *Mi a valószínűsége annak, hogy két szabályos kockával dobva a kapott számok összege 7?*

Itt a szabályosság azt jelenti, hogy a kocka bármely oldalára  $\frac{1}{6}$  valószínűséggel eshet.

- *Egy boltban az átlagos várakozási idő 2 perc. Mi a valószínűsége, hogy 3 percen belül nem kerülünk sorra, ha a várakozási idő exponenciális eloszlású?*

Itt az adott információk alapján  $1 - e^{-\frac{x}{2}}$  annak a valószínűsége, hogy a várakozási idő kevesebb mint  $x$  perc.

Ha egy hasonló feladatban a megoldáshoz szükséges információk nem mindegyike ismert, akkor azokat nekünk kell tapasztalati úton meghatározni. A *matematikai statisztika* ilyen jellegű problémákkal foglalkozik.

A statisztikai feladatokban tehát az események rendszere, pontosabban az  $(\Omega, \mathcal{F})$  mérhető tér adott, de a valószínűség nem.

**Definíció (statisztikai mező).** Legyen  $\mathcal{P}$  azon  $P: \mathcal{F} \rightarrow \mathbb{R}$  függvények halmaza, melyekre  $(\Omega, \mathcal{F}, P)$  valószínűségi mező. Ekkor az  $(\Omega, \mathcal{F}, P)$  rendezett hármast *statisztikai mezőnek* nevezzük.

A statisztikai feladatok mindig megfogalmazhatók valószínűségi változók segítségével, így egy valószínűségi változóra vonatkozólag kell információkat

gyűjteni. Jelöljük ezt  $\xi$ -vel. Az adatgyűjtésnek a statisztikában egyetlen módja van, a  $\xi$ -t meg kell figyelni (mérni) többször, egymástól függetlenül. Az  $i$ -edik megfigyelés eredményét jelölje  $\xi_i$ , amely egy véletlen érték, vagyis valószínűségi változó.

**Definíció (minta).** A  $\xi$  valószínűségi változóra vonatkozó  $n$  elemű minta alatt a  $\xi$ -vel azonos eloszlású  $\xi_1, \dots, \xi_n$  független valószínűségi változókat értünk. A  $\xi_k$ -t  $k$ -adik mintaelemnek,  $n$ -et pedig a mintaelemek számának nevezzük.

Ha több valószínűségi változóra is szükségünk van, akkor mindegyikre kell megfigyeléseket végezni, így több mintánk is lesz.

A gyakorlatban nem mintával dolgozunk, hanem konkrét számokkal, melyek a mintaelemek lehetséges értékei.

**Definíció (mintarealizáció).** Ha  $\xi_1, \dots, \xi_n$  a  $\xi$  valószínűségi változóra vonatkozó minta és  $\omega \in \Omega$ , akkor a  $\xi_1(\omega), \dots, \xi_n(\omega)$  értékeket  $\xi$ -re vonatkozó mintarealizációnak nevezzük.

Statisztikai feladatokban mintarealizáció alapján számolunk. Az így meghozott döntés nem biztos, hogy megfelel a valóságnak, csak annyit mondhatunk róla, hogy nem mond ellent a mintarealizációnak.

## A matematikai statisztika alaptétele

A vizsgált  $\xi$  valószínűségi változó  $F_\xi$  eloszlásfüggvényét ismerve rengeteg kérdés megválaszolható. Így a statisztikában alapvető feladat az eloszlásfüggvény becslése. Ehhez azt kell tudni, hogy  $F_\xi(x) = P(\xi < x)$  minden  $x \in \mathbb{R}$  esetén. Tehát egy esemény valószínűségét kell megbecsülni. A valószínűség definícióját a relatív gyakoriság tulajdonságai sugallták, így az a sejtésünk, hogy egy esemény valószínűségét a relatív gyakoriságával lenne érdemes becslni. A  $\xi < x$  esemény relatív gyakorisága a  $\xi$ -re vonatkozó minta alapján, az  $x$ -nél kisebb mintaelemek és a összes mintaelemek számának hányadosa.

**Definíció (tapasztalati eloszlásfüggvény).** Legyen  $\xi_1, \dots, \xi_n$  egy  $\xi$  valószínűségi változóra vonatkozó minta. Ekkor azt a függvényt, amely minden  $x$  valós számhoz az

$$F_n^*(x) := \frac{x\text{-nél kisebb mintaelemek száma}}{n}$$

valószínűségi változót rendel, a  $\xi$ -re vonatkozó  $n$  elemű mintához tartozó tapasztalati eloszlásfüggvényének nevezzük.

A következő tétel szerint a tapasztalati eloszlásfüggvény nagyon jó becslése a valódi eloszlásfüggvénynek.

**Tétel (matematikai statisztika alaptétele).** 1 valószínűséggel teljesül, hogy a tapasztalati eloszlásfüggvény a minta elemszámának végtelenbe tartása esetén, a valós számok halmazán egyenletesen konvergál a valódi eloszlásfüggvényhez, azaz

$$P\left(\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |F_n^*(x) - F_\xi(x)| = 0\right) = 1.$$

Tehát, ha a minta elemszáma elegendően nagy, akkor  $F_n^*(x)$  1 valószínűséggel egyformán jól közelíti az  $F_\xi(x)$  értékét minden  $x \in \mathbb{R}$  esetén.

## Statisztikák

Tegyük fel, hogy egy ismeretlen eloszlású  $\xi$  valószínűségi változó várható értékét kell meghatározni. Mivel az eloszlást nem ismerjük, ezért a minta alapján kell becslést adni. A későbbiekben látni fogjuk, hogy bizonyos szempontból jó becslése a várható értéknek a  $\xi$ -re vonatkozó  $\xi_1, \dots, \xi_n$  minta elemeinek a számtani közepe, azaz  $\frac{1}{n}(\xi_1 + \dots + \xi_n)$ . Általánosan fogalmazva itt egy olyan függvényt definiáltunk, amely egy valószínűségi változókból álló rendezett  $n$ -eshez egy valószínűségi változót rendel.

**Definíció (statisztika).** Az olyan függvényeket, melyek a mintához egy valószínűségi változót rendelnek, *statisztikának* nevezzük.

**Definíció (nevezetes statisztikák).** Legyen  $\xi_1, \dots, \xi_n$  a  $\xi$  valószínűségi változóra vonatkozó minta. Ekkor

1. a *mintaátlag*  $\bar{\xi} := \frac{1}{n}(\xi_1 + \dots + \xi_n)$ ;
2. a *tapasztalati szórásnégyzet*  $S_n^2 := \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2$ ;
3. a *korrigált tapasztalati szórásnégyzet*  $S_n^{*2} := \frac{1}{n-1} \sum_{i=1}^n (\xi_i - \bar{\xi})^2$ .

## Pontbecslések

Tegyük fel, hogy a vizsgált valószínűségi változó eloszlásának típusa már ismert (egyenletes, exponenciális, normális, stb.), de az eloszlás valamely paramétere vagy paraméterei ismeretlenek. A pontbecslés feladata, hogy ezeknek az ismeretlen paramétereknek egy valós értékű függvényét becsljük meg egy statisztikával. Például, ha  $\xi$  egyenletes eloszlású az  $[a, b]$  intervallumon, ahol  $a$  és  $b$  ismeretlenek, akkor becsljük meg egy statisztikával a  $\xi$  várható értékét. Itt a várható érték  $\frac{a+b}{2}$ , azaz a paraméterek egy valós értékű függvénye.

Fontos kérdés, hogy milyen szempontok szerint válasszuk ki a becslést megadó statisztikát. A következő természetesnek tűnő feltételeket adjuk:

- ingadozzon a becslendő érték körül;
- szórása a lehető legkisebb legyen;
- a minta elemszámának végtelenbe tartása esetén konvergáljon a becslendő értékhez.

A következőkben ezeket a feltételeket fogalmazzuk meg pontosabban. Legyen  $\xi_1, \xi_2, \dots$  az előbbi  $\xi$  valószínűségi változóra vonatkozó végtelen elemszámú minta (azaz  $\xi_1, \xi_2, \dots$  független  $\xi$ -vel azonos eloszlású valószínűségi változók). Jelölje  $\vartheta_1, \dots, \vartheta_v$  az ismeretlen paramétereket, legyen  $\vartheta := (\vartheta_1, \dots, \vartheta_v)$  és jelöljük  $g(\vartheta)$ -val a becslendő értéket.

**Definíció (torzítatlan becslés).** A  $T$  statisztika  $g(\vartheta)$  *torzítatlan becslése*, ha  $T$  várható értéke minden lehetséges  $\vartheta$  esetén  $g(\vartheta)$ -val egyezik meg.

Például

- a mintaátlag torzítatlan becslése a várható értéknek;
- egy esemény relatív gyakorisága torzítatlan becslése az esemény valószínűségének;
- a korrigált tapasztalati szórásnégyzet torzítatlan becslése a szórásnégyzetnek.

**Definíció (hatásos becslés).** A  $g(\vartheta)$  összes véges szórású torzítatlan becslése közül a legkisebb szórásút, a  $g(\vartheta)$  *hatásos becslésének* nevezzük.

Hatásos becslés nem minden esetben létezik!

**Tétel (hatásos becslés egyértelműsége).** A hatásos becslés 1 valószínűséggel egyértelmű, azaz, ha  $T_1$  és  $T_2$  a  $g(\vartheta)$ -nak hatásos becslései, akkor minden lehetséges  $\vartheta$  esetén 1 a valószínűsége, hogy  $T_1 = T_2$ .

Például egy esemény relatív gyakorisága hatásos becslése az esemény valószínűségének.

**Definíció (erősen konzisztens becsléssorozat).** A  $T_n$  statisztikasorozat (ahol  $n \in \mathbb{N}$  a minta elemszámát jelenti)  $g(\vartheta)$ -nak *erősen konzisztens becsléssorozata*, ha minden lehetséges  $\vartheta$  esetén  $\lim_{n \rightarrow \infty} T_n = g(\vartheta)$  teljesül 1 valószínűséggel.

Például

- a mintaátlag erősen konzisztens becsléssorozata a várható értéknek;
- egy esemény relatív gyakorisága erősen konzisztens becsléssorozata az esemény valószínűségének;
- a tapasztalati szórásnégyzet erősen konzisztens becsléssorozata a szórásnégyzetnek.

Hogyan lehet megtalálni az ilyen tulajdonsággal rendelkező becsléseket? Több módszer is ismert, az egyik ezek közül az ún. *maximum likelihood módszer*, amelynek az a lényege, hogy adott mintarealizációhoz az ismeretlen paramétereknek olyan becslését adjuk meg, amely mellett az adott mintarealizáció a legnagyobb valószínűséggel következik be.

## Intervallumbecslések

Tegyük fel, hogy a vizsgált valószínűségi változó eloszlásának típusa már ismert (egyenletes, exponenciális, normális, stb.), de az eloszlás valamely paramétere vagy paraméterei ismeretlenek. Jelölje  $\vartheta_1, \dots, \vartheta_v$  az ismeretlen paramétereket, és legyen  $\vartheta := (\vartheta_1, \dots, \vartheta_v)$ . Az intervallumbecslés feladata valamely ismeretlen  $\vartheta_k$  paraméter becslése egy olyan intervallummal, amelybe ez az ismeretlen paraméter nagy valószínűséggel beleesik. Az intervallumot majd konfidenciaintervallumnak nevezzük, amelynek végpontjait két statisztikával adjuk meg.

**Definíció (konfidenciaintervallum).** Legyen  $T_1$  és  $T_2$  statisztikák. Azt mondjuk, hogy a  $[T_1, T_2]$  intervallum  $1 - \alpha$  *biztonsági szintű konfidenciaintervallum* a  $\vartheta_k$  paraméterre, ha minden lehetséges  $\vartheta$  esetén legalább  $1 - \alpha$  valószínűséggel teljesül, hogy  $T_1 \leq \vartheta_k \leq T_2$ . Ha ennek a valószínűségét minden lehetséges  $\vartheta$  esetén meghatározzuk, akkor az így kapott valószínűségekből álló halmaznak az infimumát a  $[T_1, T_2]$  konfidenciaintervallum *pontos biztonsági szintjének* nevezzük. Ha minden lehetséges  $\vartheta$  esetén a  $\vartheta_k < T_1$  és  $\vartheta_k > T_2$  események valószínűségei megegyeznek, akkor a  $[T_1, T_2]$  intervallumot *centráltnak* konfidenciaintervallumnak nevezzük  $\vartheta_k$ -ra nézve.

Legyen például  $\xi$  normális eloszlású ismeretlen  $m$  várható értékkel és ismert  $\sigma$  szórással. Ekkor

$$T_1 = \bar{\xi} - \frac{\sigma}{\sqrt{n}} \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \quad \text{és} \quad T_2 = \bar{\xi} + \frac{\sigma}{\sqrt{n}} \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right)$$

választással  $[T_1, T_2]$  az  $m$ -re nézve  $1 - \alpha$  a pontos biztonsági szintű centráltnak konfidenciaintervallum, ahol  $\Phi^{-1}$  a standard normális eloszlás eloszlásfüggvényének inverzét jelenti.

## Hipotézisvizsgálatok

Hogyan lehet dönteni a mintarealizáció alapján arról, hogy egy a statisztikai mezőre vonatkozó feltételezést, más szóval *hipotézist* elfogadjuk-e igaznak vagy sem? Ez a hipotézis lehet például az, hogy a vizsgált valószínűségi változó normális eloszlású, vagy a valószínűségi változó várható értéke megfelel

az előírásnak, vagy két valószínűségi változó független, vagy várható értékeik megegyeznek stb.

Azt a feltételezést, amelyről döntést akarunk hozni, *nullhipotézisnek* nevezzük és  $H_0$ -val jelöljük. Ha  $H_0$ -t elutasítjuk, akkor egy azzal ellentétes állítást fogadunk el, melyet *ellenhipotézisnek* nevezünk, és  $H_1$ -gyel jelölünk. Döntésünk lehet helyes, vagy helytelen az alábbiak szerint:

	$H_0$ -t elfogadjuk	$H_0$ -t elutasítjuk
$H_0$ igaz	<i>helyes döntés</i>	<i>elsőfajú hiba</i>
$H_0$ nem igaz	<i>másodfajú hiba</i>	<i>helyes döntés</i>

A döntést egy ún. *próba* segítségével fogjuk meghozni. Ennek a próbának a *terjedelme*  $\alpha$ , ha minden lehetséges esetben az elsőfajú hiba valószínűsége legfeljebb  $\alpha$ . A próbát *torzítatlannak* nevezzük, ha a  $H_0$ -t nagyobb valószínűséggel utasítjuk el, ha  $H_1$  igaz, mint amikor  $H_0$  igaz.

A próbához két dologra lesz szükségünk, egy statisztikára és egy intervallumra. A statisztikát *próbastatisztikának*, az intervallumot *elfogadási tartománynak* nevezzük. Az elfogadási tartomány komplementerét *kritikus tartománynak* nevezzük. Amennyiben a mintarealizációt behelyettesítve a próbastatisztikába egy olyan számot kapunk, amely benne van az elfogadási tartományban, akkor elfogadjuk a nullhipotézist. Ha ez nem teljesül, azaz a kritikus tartományba esik a próbastatisztika értéke, akkor az ellenhipotézist fogadjuk el.

Adott feltételekkel és adott null- illetve ellenhipotézis esetén az a célunk, hogy egy olyan torzítatlan próbát találjunk, amely rögzített terjedelem esetén a lehető legkisebb valószínűségű másodfajú hibát eredményezi.

Előfordulhat, hogy még a legkisebb másodfajú hiba valószínűsége is túlságosan nagy. Ekkor célszerű az előző kívánalmak mellett azt is megkövetelni a próbától, hogy a minta elemszámának végtelenbe tartásával a másodfajú hiba valószínűsége konvergáljon nullához. Ezt a tulajdonságot úgy nevezzük, hogy a próba *konzisztens*. Ilyen próba esetén, amennyiben a legkisebb másodfajú hiba valószínűsége is túlságosan nagy, csak annyit kell tennünk, hogy növeljük a minta elemszámát, amivel tetszőlegesen kicsivé tehető a másodfajú hiba valószínűsége.

### Egymintás u próba

Legyen  $\xi$  normális eloszlású valószínűségi változó ismeretlen  $m$  várható értékkel és ismert  $\sigma$  szórással, továbbá legyen  $\xi_1, \dots, \xi_n$  a  $\xi$ -re vonatkozó minta.

A

$$H_0: m = m_0$$

$$H_1: m \neq m_0$$

hipotézisekre kell adni  $\alpha$  terjedelmű próbát, ahol  $m_0 \in \mathbb{R}$  rögzített. Bizonyítható, hogy az előző kívánalmaknak megfelelő  $\alpha$  terjedelmű próbát kapunk, ha a próbastatisztika

$$u := \frac{\bar{\xi} - m_0}{\sigma} \sqrt{n}$$

és az elfogadási tartomány a

$$\left[ -\Phi^{-1} \left( 1 - \frac{\alpha}{2} \right), \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \right]$$

intervallum.

### **Kétmintás u-próba**

Legyen  $\xi$  normális eloszlású valószínűségi változó  $m_1$  és  $\sigma_1$  paraméterekkel,  $\eta$  normális eloszlású valószínűségi változó  $m_2$  és  $\sigma_2$  paraméterekkel, ahol  $m_1, m_2$  ismeretlenek és  $\sigma_1, \sigma_2$  ismertek. Legyen  $\xi_1, \dots, \xi_{n_1}$  a  $\xi$ -re vonatkozó, illetve  $\eta_1, \dots, \eta_{n_2}$  az  $\eta$ -ra vonatkozó minta. A

$$H_0: m_1 = m_2$$

$$H_1: m_1 \neq m_2$$

hipotézisekre kell adni adott  $\alpha$  terjedelmű próbát. Bizonyítható, hogy a kívánalmaknak megfelelő próbát kapunk, ha a próbastatisztika

$$u := \frac{\bar{\xi} - \bar{\eta}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

és az elfogadási tartomány a

$$\left[ -\Phi^{-1} \left( 1 - \frac{\alpha}{2} \right), \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \right]$$

intervallum.