

Good-Turing lefedés

Lang Zsolt

2017.03.24.

Bevezetés

Fajok közösségét vizsgáljuk. Sok faj van, az egyedek száma gyakorlatilag végtelen. Az egyedekből véletlen mintát veszünk. Kérdés, a mintában van-e, előfordul-e a közösség minden fajából egyed? Ha nem, akkor mennyire fedik le a mintába került fajok a teljes fajközösséget? Általában, hogyan jellemezhető a minta alapján a közösségben a fajok gyakorisági eloszlása?

A most ismerttetendő eljárást eredetileg a II. VH-ban a német ENIGMA kódjának feltörésére fejlesztették ki. Az ENIGMA-hoz naponta változtatható kódtáblázatok tartoztak, amik 9 „alaptáblázat” speciális átkódolásával jöttek létre. Turing az alaptáblázatban szereplő kódok gyakorisági mintázatával igyekezett beazonosítani az adott napon használt táblázatot (Good 2000).

Elsősorban Good (1953) publikációját követjük.

Néhány jelölés

A minta nagysága legyen N .

A populációban lévő fajok száma S , ezt a számot nem ismerjük, de feltesszük, hogy véges.

A fajok populációs részaránya p_1, p_2, \dots, p_S , ezeket sem ismerjük.

A kérdés pontosabban

Az N elemű minta alapján becsüljük meg a következőt. **Mi annak a valószínűsége, hogy egy újonnan kiválasztott egyed faja már szerepel a mintában?**

Ez a valószínűség **a minta Good-Turing-féle lefedése** (Good-Turing sample coverage).

(Matematikailag könnyebben kezelhetőnek tűnik annak az eseménynek a valószínűsége, hogy az új egyed faja még nem szerepel a mintában.)

A gyakoriságok gyakorisága

Nevezzük **singleton**-nak az olyan fajt, amelyik pontosan egyszer fordul elő a mintában. A mintában lévő singleton-ok számát jelöljük n_1 -gyel.

Legyen **doubleton** a neve az olyan fajoknak, amelyek pontosan kétszer fordulnak elő a mintában. A mintában lévő doubletonok számát jelöljük n_2 -vel.

Általában vizsgáljuk azokat a fajokat, amelyek pontosan r egyeddel vannak jelen a mintában. Legyen a számuk n_r . Ha $r=1$, akkor ezek a singleton-ok, ha $r=2$, akkor a doubleton-ok. Ha $r=0$, akkor **n_0 a mintában nem szereplő fajok száma.**

Két összefüggés

Az egyik $n_0+n_1+\dots+n_r+\dots=S$.

A másik $1\cdot n_1+2\cdot n_2+\dots+r\cdot n_r+\dots=N$

A q_r

Válasszunk ki egy olyan **fajt**, amihez **r számú egyed tartozik a mintában**. Legyen ennek a fajnak a populációs részaránya q_r . Ez a részarány valószínűségi változó, mert függhet a kiválasztott fajtól. Modellezni fogjuk az eloszlását.

A q_r relatív gyakoriságos (maximum likelihood) becslése a mintából

$$r/N.$$

Ez a becslés azonban nem kielégítő. Az $r=0$ esetben pl. 0-t ad.

A fő eredmények

A q_r várható értéke közelítőleg

$$Eq_r \approx r^*/N,$$

ahol

$$r^* = (r+1) \cdot n_{r+1} / n_r.$$

A mintában r egyeddel szereplő fajok együttes populációs részarányának közelítő várható értéke ez alapján

$$n_r \cdot Eq_r = (r+1) \cdot n_{r+1} / N.$$

A mintában nem szereplő fajok együttes populációs részarányának közelítő várható értéke

$$n_1 / N.$$

Általánosítás

q_r m-ik momentuma

$$E(q_r^m) \approx (r+m)^{(m)} / N^m \cdot n_{r+m} / n_r,$$

ahol

$$t(m) = t \cdot (t-1) \cdot \dots \cdot (t-m+1).$$

Itt $r=1,2,3\dots$ és $m=0,1,2,\dots$

Bizonyítás I.

Először megmutatjuk, hogy q_r várható értékét elég megismerni $r \geq 1$ esetén.

A mintában legalább $r \geq 1$ egyeddel szereplő fajok együttes populációs részarányának közelítő várható értéke

$$N^{-1} \cdot \{(r+1)n_{r+1} + (r+2)n_{r+2} + \dots\}.$$

Speciálisan, a mintában legalább 1 egyeddel képviselt fajok együttes populációs részaránya

$$N^{-1} \cdot \{2n_2 + 3n_3 + \dots\} = 1 - n_1/N,$$

mivel

$$1 \cdot n_1 + 2 \cdot n_2 + \dots + r \cdot n_r + \dots = N.$$

Tehát a várható érték képlete $r=0$ -ra is érvényes.

Bizonyítás II.

Az n_r várható értéke felírható az

$$E_N(n_r) = C_{N,r} \cdot \sum_k p_k^r \cdot (1-p_k)^{N-r}$$

alakban.

Speciálisan

$$E_N(n_0) = \sum_k (1-p_k)^N.$$

Nem ismerjük azonban a fajok p_1, p_2, \dots, p_S populációs részarányát.

Bizonyítás III.

A mintabeli r megfigyelt előfordulás likelihoodja

$$P(r \text{ előfordulás a mintában} | q_r = p_k) = C_{N,r} \cdot p_k^r \cdot (1-p_k)^{N-r}.$$

A faj kiválasztásának a priori valószínűségét **modellezzük** $1/S$ -sel (nem informatív prior).

Ebben a modellkörnyezetben q_r posterior eloszlása arányos a likelihooddal

$$P(q_r = p_k | r \text{ előfordulás a mintában}) \propto p_k^r \cdot (1-p_k)^{N-r}.$$

Bizonyítás IV.

A posterior eloszlással ki tudjuk számolni q_r posterior momentumait:

$$E(q_r^m | r \text{ előfordulás}) = \frac{\sum_k p_k^{r+m} \cdot (1-p_k)^{N-r}}{\sum_k p_k^r \cdot (1-p_k)^{N-r}}.$$

Vegyük észre, hogy ennek részei nagyon hasonlítanak az n_r várható értékére:

$$E_N(n_r) = C_{N,r} \cdot \sum_k p_k^r \cdot (1-p_k)^{N-r}.$$

Behelyettesítéssel kapjuk, hogy

$$E(q_r^m | r \text{ előfordulás}) = \frac{(r+m)^{(m)}}{(N+m)^{(m)}} \cdot \frac{E_{N+m}(n_{r+m})}{E_N(n_r)}.$$

A várható értékeket a mintában megfigyelt n_{r+m} és n_r értékekkel közelítve megkapjuk a bizonyítandó összefüggéseket.

Heurisztika

Képzeljük el, hogy az N elemű mintát az egyedek egymás után történő kisorsolásával kaptuk. Ha a mintához hozzáveszünk egy új egyedet és ennek faja a mintában addig nem fordult elő, akkor az új mintában az egyed faja singleton lesz.

A mintában nem szereplő fajú egyed kiválasztásának eseménye egy singleton létrejöttét jelenti. Ennek valószínűségét a relatív gyakorisággal becsülve jutunk a mintában nem szereplő fajok részarányának

$$n_1/N$$

becsléséhez.

Megjegyzés: a singletonok tartósan csak nagy N mellett, a mintavétel vége felé maradnak meg, amikor mind N , mind n_1 már „elég stabil”. (Az elején egy singleton gyorsan doubletonná változik).

Megjegyzések

- ❖ Ha $n_1=0$, azaz nincs a mintában singleton, akkor a Good-Turing elmélet szerint azt jósoljuk, hogy **már minden faj szerepel a mintában**.
- ❖ Az n_1/N becslés átlagnégyzetes hibája nagy N -re $\leq 1/N$ (Robbins 1968).
- ❖ Az ökológiai mellett sok lingvisztikai alkalmazás is létezik.
- ❖ Az n_1, n_2, \dots gyakoriságokat szokás simítani
- ❖ Chao és mtsai (1987, 2012) javítottak a becslésen:

$$n_1/N \cdot (N-1) \cdot n_1 / [(N-1) \cdot n_1 + 2 \cdot n_2],$$

ha $n_1 > 0$ és $n_2 > 0$.

Kitekintés, általánosítás

- ❖ Az elméletet és finomított változatait alkalmazzák a közösség fajeloszlásának pontosabb becslésére. Chao és mtsai a fajok rangszám szerint csökkenő módon rendezett relatív gyakoriságait (rank abundance distribution of species) modellezték a módszerrel (Chao et al. 2015)
- ❖ A fajok részarányából származtatott mennyiségek, pl. a Shannon entrópia becslése is javítható a mintában nem előforduló fajok figyelembe vételével (Chao et al. 2013).
- ❖ Annak ellenére, hogy régi ötletről van szó, a témakört ma is intenzíven kutatják, fejlesztik.

Hivatkozások

Chao A (1987). Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. *Biometrics* 43(4), 783-791.

Chao A et al. (2015). Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory. *Ecology* 96(5), 1189–1201.

Chao A, Jost L (2012). Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology* 93(12), 2533–2547.

Chao A, Wang YT, Jost L (2013). Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution* 4, 1091–1100.

Hivatkozások

Good IJ (1953). The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika* 40(3/4), 237-264.

Good IJ (2000). Turing's anticipation of empirical bayes in connection with the cryptanalysis of the naval enigma. *Journal of Statistical Computation and Simulation* 66:2, 101-111.

Robbins HE (1968). Estimating the total probability of the unobserved outcomes of an experiment. *The Annals of Mathematical Statistics* 39, 256–257.