Becslések, statisztikai próbák

1. Maximum likelihood módszer

Egy ismert típusú eloszlás ismeretlen paraméterét becsüljük úgy. Olyan becslést szeretnénk adni ami a mintáknak a lehető legjobban (*maximálisan*) megfelel.

Az ismeretlen paraméter tartalmazó függvényt L-el jelöljük. Ez diszkrét esetben jelölheti magát a valószínűséget is, de általában csak egy viszonyszámot ad, aminek minél nagyobb az értéke, annál jobbnak tekintjük a becsült paraméter értéket.

A maximumokat a szokásos módon keressük, vagyis megnézzük, hogy hol lesz a függvény deriváltjának az értéke egyenlő 0-val.

Bizonyos eloszlások esetében L helyett egyszerűbb $\ln L$-el számolni. A log-likelihood függvényt l-el jelöljük. $l = lnL$.

1.1. Exponenciális eloszlás paraméterének becslése

Tegyük fel, hogy van n darab mért értékünk egy folyamatról, amiről feltételezzük, hogy exponenciális eloszlású. Azt szeretnénk meghatározni, hogy mennyi lesz az eloszlás λ paramétere.

A mintákat jelöljük $x_1, \ldots, x_n$ formában. Az exponenciális eloszlás sűrűségfüggvénye

$$f_\xi(x) = \begin{cases} \lambda \cdot e^{-\lambda x}, & x \leq 0, \\ 0, & \text{egyébként.} \end{cases}$$

Ez a sűrűségfüggvény tipikus megadása. Fontos azonban észrevenni, hogy a sűrűségfüggvény értéke függ a λ értéktől is, tehát jelölhető $f_\xi(x_i; \lambda)$ formában is. Ez utóbbi azért lesz itt praktikusabb, mert a megfelelő λ értéket becsüljük.

A sűrűségfüggvény csak jelzi, hogy egy-egy érték milyen valószínűséggel fordulhat elő egy pont környezetében. Feltételezzük, hogy a minták azonos eloszlásból, egymástól függetlenül elvégzett kísérletekből származnak. Ekkor a likelihood függvényt a következő alakban írhatjuk fel.

$$L(\lambda; x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\xi(x_i; \lambda) = \prod_{i=1}^{n} \lambda \cdot e^{-\lambda x} = \lambda^n \cdot e^{-\lambda \sum_{i=1}^{n} x_i}$$

Azon λ értéket keressük, ahol az L értéke maximális lesz. Ehhez az alábbi egyenletet kell megoldani λ-ra:

$$\frac{\mathrm{d}L}{\mathrm{d}\lambda} = 0.$$

A likelihood függvény alakja olyan, hogy itt célszerű inkább a log-likelihood függvénnyel számolni:

$$l(\lambda; x_1, \ldots, x_n) = \ln L(\lambda; x_1, \ldots, x_n) = n \cdot \ln(\lambda) - \lambda \cdot \sum_{i=1}^{n} x_i$$

λ szerint deriválva az alábbi összefüggést kapjuk:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} l(\lambda; x_1, \ldots, x_n) = n \cdot \frac{1}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

Ezt átrendezve a λ becsült értékére ($\hat{\lambda}$-val jelölve) azt kapjuk, hogy

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}.$$

1.2. Poisson eloszlás paraméterének becslése

Egy kísérlet esetén tudjuk, hogy a bekövetkezések/előfordulások száma Poisson eloszlást követ. Van egy konkrét mérésünk (mintánk), és az alapján szeretnénk megbecsülni az eloszlás λ paraméterét.

Ebben az esetben valóban valószínűséget szeretnénk maximalizálni, mivel a likelihood függvényt az alábbi módon adhatjuk meg:

$$L(\lambda; k) = P(\xi = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}.$$

A derivált ebből egyszerűen számolható (itt nem szükséges loglikelihood függvényt felírni):

$$\frac{\mathrm{d}L}{\mathrm{d}\lambda} = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\lambda^k}{k!} \cdot e^{-\lambda}\right) = \frac{1}{k!} \cdot k \cdot \lambda^{k-1} \cdot e^{-\lambda} + \frac{\lambda^k}{k!} \cdot (-1) \cdot e^{-\lambda} = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \cdot \left(\frac{k}{\lambda} - 1\right) = 0.$$

Tudjuk, hogy $\frac{\lambda^k}{k!} \cdot e^{-\lambda} > 0$ biztosan teljesül. Így tehát

$$\frac{k}{\lambda} - 1 = 0 \quad \Rightarrow \quad \frac{k}{\lambda} = 1 \quad \Rightarrow \quad \hat{\lambda} = k.$$

2. Egymintás u-próba

- Feltételezzük, hogy a minta normális eloszlású.
- Ismerjük a szórást. (A felírásban a jobbalsó sarokban lévő 0-ás index jelzi, hogy ismerjük.)
- A várható értéket szeretnénk ellenőrízni egy adott minta alapján.

$$\xi_1, \ldots, \xi_n \sim \mathcal{N}(\mu, \sigma_0^2)$$

Arról szeretnénk meggyőződni, hogy a μ (várható) érték megegyezik-e egy μ_0 feltételezett várható értékkel. Ez hipotézisek formájában a következőképpen néz ki:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Tudjuk, hogy a következőképpen számított u érték normális eloszlást követ:

$$u = \frac{\overline{\xi} - \mu_0}{\sigma_0} \cdot \sqrt{n} \sim \mathcal{N}(0, 1).$$

Jelölje a szignifikancia szintet az $1 - \alpha$ kifejezés. A hipotézis eldöntéséhez a következő összefüggést kell megvizsgálni:

$$P\left(|u| \leq u_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

A számítások során ez annyit jelent, hogy a minták alapján számított értéket kell összehasonlítani a táblázatban találhatóval,

$$|u| \leq u_{\frac{\alpha}{2}} \quad \Rightarrow \quad \text{Elfogadjuk a } H_0 \text{ hipotézist.}$$

A szignifikancia szint általában adott szokott lenni, tipikusan 0.95-nek választják, és van hogy százalékos értékként (95%-os szignifikancia szint) hivatkoznak rá. Az $u_{\frac{\alpha}{2}}$-t a következő összefüggés alapján számíthatjuk ki:

$$\Phi\left(u_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}.$$

2.1. Számítási példa

Elvégeztünk 25 mérést, amely a következő eredményeket adta:

```
518.9598, 501.1553, 495.0711, 520.8532, 510.0095,
484.6808, 482.6422, 491.6349, 497.0269, 492.8886,
495.2472, 486.5123, 489.5820, 493.8927, 490.1868,
487.7767, 510.8099, 499.7295, 497.5016, 500.1564,
514.2096, 495.3246, 505.3583, 491.7084, 515.3049
```

Feltételezzük, hogy ez egy $\sigma = 10$ szórású, normális eloszlású valószínűségi változó adta. Vizsgáljuk meg, hogy a várható érték egyenlő-e 500-al!

Hipotézisek:

$$H_0 : \mu = 500$$
$$H_1 : \mu \neq 500$$

A számértékek átlaga $\overline{\xi} = 498.7289273514852$.

$$u = \frac{\overline{\xi} - \mu_0}{\sigma_0} \cdot \sqrt{n} \approx -0.6355363242574015$$

Az $u_{\frac{\alpha}{2}}$ értékének számítása $\alpha = 0.05$ esetén:

$$\Phi(u_{\frac{\alpha}{2}}) = 1 - \frac{0.05}{2} = 0.975 \quad \Rightarrow \quad u_{\frac{\alpha}{2}} \approx 1.96$$

Ez alapján

$$P(|u| \leq 1.96) = 0.95,$$

vagyis elfogadhatjuk a H_0 hipotézist.

3. Kétmintás u-próba

- Két eloszlást szeretnék összehasonlítani. Jelölje az ezekhez tartozó valószínűségi változókat ξ és η. Feltételezzük, hogy mindkettő normális eloszlású.
- Ismerjük mindkettő szórását.
- Azt szeretnénk megállapítani, hogy a várható értékeik megegyeznek-e a rendelkezésre álló minták alapján.

ξ-hez n darab, η-hoz m darab mért érték áll rendelkezésre. A paramétereket jelöljük a következőképpen:

$$\xi \sim \mathcal{N}(\mu_1, \sigma_1), \quad \eta \sim \mathcal{N}(\mu_2, \sigma_2).$$

A hipotézisek a következők:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Az egymintás u-próbához hasonlóan számolható. Itt viszont

$$u = \frac{\overline{\xi} - \overline{\eta}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0, 1).$$

3.1. Számítási példa

A ξ változóhoz tartozó értékek:

```
105.8206, 96.1440, 96.4462, 102.1284, 98.4490,
103.5946, 99.7446, 105.5324, 101.2205, 93.5363,
97.5556, 93.3587, 99.1626, 94.2575, 104.4654,
98.9614, 100.6020, 102.9229, 98.7940, 101.6148
```

Az η változóhoz tartozó értékek:

```
89.6774, 92.2259, 98.4799, 90.5277, 94.3893,
87.4341, 95.9473, 97.5764, 93.4602, 98.6512
```

Azt feltételezzük, hogy $\sigma_1 = 4, \sigma_2 = 3$.

$n = 20, m = 10, \overline{\xi} = 99.71557543010641, \overline{\eta} = 93.83694642431972$

$$u = \frac{\overline{\xi} - \overline{\eta}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \approx 4.508702629861019$$

$\alpha = 0.05 \Rightarrow u_{\frac{\alpha}{2}} = 1.96$

$$P(|u| \leq 1.96) = 0.95$$

Nem teljesül, ezért a H_0 hipotézist elutasítjuk, és a H_1-et fogadjuk el, vagyis hogy a két változó várható értéke nem egyezik meg.

4. F-próba

- Adott ξ és η normális eloszlású valószínűségi változó.
- A rendelkezésre álló minták alapján el szeretnénk dönteni, hogy a szórásuk megegyezik-e.

Hipotézisek:

$$H_0 : \sigma_1 = \sigma_2$$
$$H_1 : \sigma_1 \neq \sigma_2$$

Ki kell számítani hozzá az alábbi F értéket:

$$F = \max\left\{ \frac{(n-1)s_n^{*2}}{(m-1)s_m^{*2}}; \frac{(m-1)s_m^{*2}}{(n-1)s_n^{*2}} \right\}.$$

(A nagyobb korrigált tapasztalati szórásnégyzetet osztjuk a kisebbel, úgy hogy közben a mintaszámok arányát is figyelembe vesszük.) Az $F \geq 1$ biztosan teljesül. Dönteni az F-eloszlás adott α értéke szerint lehet:

$$P(F > F_\alpha) = 1 - \alpha.$$

4.1. Számítási példa

ξ által adott értékek:

```
-14.0780, 2.8105, 11.6229, 0.1691, -23.6753,
23.5296, 1.1141, -6.9857, -11.7706, 3.9790
```

η által adott értékek:

```
0.9428, -2.9558, -2.2356, -9.9934, -17.2261,
-10.9656, 26.2531, 3.7662, 4.0430, 6.1016,
17.2619, -1.2116, 0.0059, -13.5192, 22.2694
```

Vizsgáljuk meg, hogy a ξ és η szórása megegyezhet-e!

$n = 10, m = 15, s_n^{*2} \approx 182.46, s_m^{*2} \approx 159.48$
$F \approx \dfrac{14 \cdot 159.48}{9 \cdot 182.46} \approx 1.3596$
Az $m-1, n-1$ szabadságfokú F-eloszláshoz tartozó érték 3.025,

$$P(F > 3.025) = 1 - \alpha$$

vagyis elutasítjuk a H_0 hipotézist, a két eloszlás szórását nem tekintjük egyenlőnek.

5. Egymintás T-próba

- Egy ξ, feltételezés szerint normális eloszlású valószínűségi változó várható értékét becsüljük.
- A szórást nem ismerjük, azt is magából a mintából becsüljük.

$$\xi_1, \ldots, \xi_n \sim \mathcal{N}(\mu, \sigma^2)$$

Hipotézisek:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

A döntéshez, hogy melyik hipotézist fogadjuk el, meg kell határozni egy t értéket:

$$t = \frac{\overline{\xi} - \mu_0}{s_n^*} \cdot \sqrt{n} \sim t_{n-1} \quad (n-1 \text{ szabadságfokú Student eloszlás}).$$

A szignifikanciaszint kiválasztása után:

$$P\left(|t| < t_{n-1, \frac{\alpha}{2}}\right) = 1 - \alpha.$$

6. Kétmintás T-próba

- A próba azt vizsgálja, hogy ξ és η független, normális eloszlású valószínűségi változóknak a minták alapján megegyezhet-e a várható értéke.
- A próba feltételezi, hogy a ξ és η szórása megegyezik. Ezt F-próba elvégzésével kell belátni.

A hipotézisek a következők:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

A vizsgálathoz ki kell számítani az alábbi t értéket, amiről tudjuk, hogy $n+m-2$ szabadsági fokú Student eloszlású valószínűségi változó:

$$t = \frac{\overline{\xi} - \overline{\eta}}{\sqrt{\dfrac{m+n}{m \cdot n}}} \Bigg/ \sqrt{\frac{(n-1) \cdot s_n^{*2} + (m-1) \cdot s_m^{*2}}{n+m-2}} \sim t_{n+m-2}.$$

A $|t| < t_{n+m-2, \frac{\alpha}{2}}$ állítás *erősségét* vizsgáljuk, vagyis hogy

$$P\left(|t| < t_{n+m-2, \frac{\alpha}{2}}\right) = 1 - \alpha$$

adott $(1-\alpha)$ szignifikancia szinten teljesül-e.

7. χ^2-próba

- Illeszkedésvizsgálathoz használhatjuk. Van egy mintánk, és azt szeretnénk megállapítani, hogy adott, konkrét eloszláshoz tartozhat-e.
- $\xi_1, \ldots, \xi_n$ azonos eloszlású, független valószínűségi változók.
- F a valódi eloszlás (amit nem ismerünk), F_0 a feltételezett eloszlás.

Hipotézisek:

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

A próbához ξ felvehető értékeinek tartományát intervallumokra kell bontani. Diszkrét esetben ez egyszerűbb, lehet például minden felvehető érték egy-egy külön intervallumban.

$$\chi^2 = \sum_{i=1}^{k} \frac{(\nu_i - n \cdot p_i)^2}{n \cdot p_i} \sim \chi^2_{k-1},$$

ahol

- k: az intervallumok száma,
- ν_i: az i-edik intervallumban előforduló értékek száma,
- p_i: annak a valószínűsége, hogy egy véletlenszerű bekövetkezés az i-edik intervallumba essen.

Az érték azért csak $k-1$ szabadságfokú, mert $k-1$ intervallum meghatározása után a k-adik már nem lehet független. Az $(1-\alpha)$ szignifikancia szint meghatározása után a

$$P\left(|\chi^2| \leq \chi^2_{k-1,\alpha}\right) = 1 - \alpha$$

alapján lehet dönteni.

7.1. Számítási példa

Kaptunk egy tetraéder alakú *dobókocka*-félét. Meg szeretnénk bizonyosodni arról, hogy valóban szabályos, és egyenletes eloszlás szerint fordulnak elő az $1, 2, 3, 4$ értékek. Elvégeztünk 100 dobást, amely a következő eredményeket adta:

dobás	1	2	3	4
darabszám	21	18	30	31

. Mit mondhatunk a kocka szabályosságával kapcsolatban 95%-os szignifikancia szinten?

Az intervallumokat megválaszthatjuk úgy, hogy $(-\infty, 1.5), [1.5, 2.5), [2.5, 3.5), [3.5, +\infty)$. (A számítások során diszkrét esetben nincs jelentősége, hogy a felvett értékek között hogy adjuk meg a részintervallumokat.)

Egyenletes eloszlást feltételezve:

$$P(\xi = 1) = P(\xi = 2) = P(\xi = 3) = P(\xi = 4) = \frac{1}{4}.$$

A χ^2 értékét az alábbi módon kapjuk:

$$\chi^2 = \sum_{i=1}^{k} \frac{(\nu_i - n \cdot p_i)^2}{n \cdot p_i} = \frac{(21-25)^2}{25} + \frac{(18-25)^2}{25} + \frac{(30-25)^2}{25} + \frac{(31-25)^2}{25} = 5.04$$

A χ^2 táblázat alapján, ha $k = 3, \alpha = 0.05$, akkor $\chi^2_{k-1,\alpha} = 7.815$, amiből

$$P\left(|\chi^2| \leq 7.815\right) = 1 - 0.05 = 0.95$$

eredményt kapjuk, vagyis a dobásaink 95%-os szignifikancia szinten egyenletes eloszlás szerintinek tekinthetők.