A törpök és a konfidencia-intervallum

méréstechnikai mese

Hókuszpók feladata

A gonosz Hókuszpók megint foglyul ejtett néhány törpöt. Bezárta őket várának pincéjébe, ahol a helyiségek tele vannak kakukkos órával, pontosan 100 egyforma óra van ott elhelyezve. Mivel Hókuszpók a borait is ott tartja, a hőmérséklet és a páratartalom is szigorúan állandó. Azt mondta, szabadon engedi a törpöket, ha képesek 1 perc pontossággal megmondani, hogy mikor lesz másnap dél. Ez azért nehéz feladat, mert Hókuszpóknak csak egy pontos órája van, de az a saját szobájában. A kakukkos órák össze-vissza járnak, akár 10 percet is eltérhetnek a pontos időtől, mert Hókuszpók már egy hete nem állította be őket. Hókuszpók csak annyit mond nekik kárörvendő vigyorral, hogy az órák a pontos időre vonatkozó torzítatlan becslőt szolgáltatnak, és rájuk zárja az ajtót. Okoska, Ügyi és Nótata más-más stratégiát javasol, a feladat eldönteni, melyikük javaslata a legjobb:

1. Okoska sokat tanult méréstechnikát (is), ezért azt javasolja, számítsák ki az órák által mutatott idő átlagát, és amikor az éppen 12 óra, jelezzenek Hókuszpóknak.

2. Ügyi nem tanult méréstechnikát, de praktikus törp lévén tudja, hogy az órák vagy sietnek vagy késnek, és feltételezte, ez a itt lévő órákra is fele-fele arányban igaz. Mivel az órák nem egyszerre kezdenek kakukkolni, meg lehet számolni, hogy hány óra kezdte már meg a kakukkolást. Ügyi szerint egyszerűen el kell kezdeni számolni másnap délben, hogy hány óra kezdett el kakukkolni, és 50-nél szólni Hókuszpóknak.

3. Nótata inkább művészlélek, de feltűnik neki, hogy amikor az egész óra közeleg (az órák óránként kakukkolnak), először csak egy-egy óra kezd el kakukkolni, majd egyre több, végül már egészen nagy a hangzavar, majd elkezd csökkenni hang intenzitása, és egyre kevesebb óra szól. Egy óra kb. 40 másodpercig kakukkol. Nótata szerint akkor kell szólni Hókuszpóknak, amikor a legnagyobb a hangzavar.

Megoldás

A feladat megfogalmazásából (állandó környezeti feltételek, torzítatlanság) arra lehet következtetni, hogy az órák által mutatott idő a pontos idő mint várható érték körüli normális eloszlást mutat. Ebben az esetben világos, hogy a legjobb stratégia az órák által mutatott idő átlagát kiszámítani, mert ez lesz a várható érték legjobb becslője. Az, hogy milyen valószínűséggel szabadulnak ki, attól függ, milyen az átlag eloszlása.

A pincében elég sok óra van, azaz elég sok minta áll rendelkezésünkre. Mivel a normális eloszlás csak 0.3% valószínűséggel vesz fel olyan értéket, amelynek eltérése a várható értéktől a szórás 3-szorosánál nagyobb, a szórásról azt állíthatjuk, hogy:

$$\sigma \approx \frac{10}{3}\text{perc} \tag{1}$$

Mivel a várható értéket átlagolással becsüljük:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{2}$$

kérdés, mi ennek az eloszlása. A centrális határeloszlás tétel értelmében, ha N elég nagy, x_i tetszőleges eloszlása esetén $\hat{\mu}$ eloszlása normális lesz. Most is igaz ez, amikor abból indultunk ki, hogy x_i eloszlása is normális. Amennyiben az x_i minták függetlenek (márpedig az órák egymástól függetlenül mérik az időt), igaz az, hogy:

$$\sigma' = \frac{\sigma}{\sqrt{N}} \tag{3}$$

ahol σ' μ szórása. Mivel $N = 100$,

$$\sigma' = \frac{\sigma}{\sqrt{100}} \approx \frac{1}{3}\text{perc} \tag{4}$$

$\hat{\mu}$ eloszlása tehát olyan normális eloszlás, amelynek szórása $\sigma' = 1/3$ perc. A Hókuszpók által megadott 1 perces pontosság ± 1 perces konfidencia-intervallumnak felel meg. Mivel ez σ' 3-szorosa, az átlagérték kiszámítása $p = 99.7\%$-os konfidencia-szintű becslőt szolgáltat. (A helyzet azonban nem ennyire szép, erre majd még visszatérünk.)

Most értékeljük a törpök megoldásait!

1. Okoska megoldása tökéletes, legfeljebb nehezen kivitelezhető. Érdemes elgondolkodni, hogyan lehetne megoldani a várható érték képzését, ha nem áll rendelkezésre semmilyen technikai segítség, egyedül papírt és íróeszközt használhatunk az átlag számításához. A megmenekülés valószínűsége (feltéve, hogy sikerült kiszámítani az átlagot) 99.7%. Láthatóan elég csekély valószínűség adódik arra, hogy nem menekülnek meg. Viszont a szórást csak "szemre" állapítottuk meg, így aztán helyesebb, ha azt mondjuk, Okoska eljárásával majdnem 100% valószínűséggel megmenekülnek.

2. Ügyi megoldása is kihasználja, hogy elég sok mintát vettünk. Ennyi minta alapján felrajzolható az eloszlás hisztogramja, ami a valószínűség-sűrűségfüggvény becslője. A minták feléhez tartozó érték pedig az eloszlás mediánját becsli, amely normális eloszlás esetén megegyezik a várható értékkel. Ez a megoldás a gyakorlatban is egyszerűen kivitelezhető. A megmenekülés valószínűsége ugyanakkora, mint Okoska esetében.

3. Nótata megoldása is akkor működik, ha elég sok óra van, de ez a mi feladatunkban adott. Vizsgáljuk meg, hogy hogyan számítható ki, hogy egy adott óra egy adott t időpontban kakukkol! Egy óra akkor kakukkol, ha legfeljebb T idővel korábban (T a kakukkolás időtartama) kezdte meg a kakukkolást. Ehhez integrálni kell a valószínűség-sűrűségfüggvényt a $[t - T, t]$ intervallumon. Szemléletesen nyilvánvaló, hogy normális eloszlás esetén ez az integrál, és ezzel együtt a kérdéses valószínűség akkor a legnagyobb, ha $t = T/2$, ekkor ugyanis a normális eloszlás valószínűség-sűrűségfüggvényéből a várható értékre szimmetrikus T hosszúságú intervallumot vágjuk ki. Nótata ezt a valószínűséget becsli a minták alapján, és az eloszlás móduszát határozza meg. Ez azonban az eredeti eloszlás várható értékére vonatkozóan torzított becslőt eredményez, a torzítás mértéke $T/2$. Mivel a feladat szerint $T = 40$ másodperc, ami 2/3 perc, a torzítás 1/3 perc, ami éppen az átlagérték szórása, azaz: $\sigma' = T/2$. Ebben az esetben Hókuszpók intervalluma a becslőre aszimmetrikusan helyezkedik el: az alsó határ a becslőtől $4\sigma'$, a felső $2\sigma'$ távolságra. Ezt figyelembe véve a valószínűségek:

$$\begin{aligned} p(\hat{\mu} > 11:59) &\cong 0.5 \\ p(\hat{\mu} < 12:01) &\cong 0.477 \end{aligned} \tag{5}$$

azaz a megmenekülés valószínűsége $p \cong 97.7\%$. Ez kicsit rosszabb, mint az előző két esetben, de még mindig elég jó. 0.1% pontosság persze itt sem jogos, a variancia pontatlan becslése miatt, de a tendenciát jól mutatja az eredmény.

Megjegyzés

A medián, illetve a módusz alapján történő becslés a valószínűség-sűrűségfüggvény becslésével van kapcsolatban, amelynek hibái részletesebb vizsgálatokat indokolnának. Konkrét hibaszámításra azonban csak a minták ismeretében lenne mód, amikor pedig a méréstechnikailag releváns átlagérték-számítást kell alkalmazni. Ehelyütt elégedjünk meg annyival, hogy elegendően nagy számú mintánk áll rendelkezésre, ezért volt jogos az ismertetett 2. és 3. megoldás.

A megoldás finomítása

Pontosabb számításra akkor van lehetőség, ha ismerjük az órák által mutatott időt egy konkrét időpontban. A várható értéket és a szórást ekkor ki lehet számolni a következőképpen:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i;\ \ s^* = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2} \tag{6}$$

ahol $\hat{\mu}$ a várható érték becslője az x_i minták alapján, s^* a tapasztalati szórás, N pedig a minták száma.

Megoldás a szórást ismertnek feltételezve

Először tételezzük fel, hogy:

$$s^* = \sigma \tag{7}$$

azaz a tapasztalati szórás megegyezik az eloszlás szórásával. Azt tudjuk, hogy N független, σ szórású mérési eredmény átlagolásakor az átlag szórása:

$$\sigma' = \frac{\sigma}{\sqrt{N}} \tag{8}$$

tehát a:

$$z = \frac{\hat{\mu} - \mu}{\sigma'} \tag{9}$$

valószínűségi változó standard normális eloszlást mutat. Ennek segítségével megadható egy olyan intervallum, amelyben a z valószínűségi változó p valószínűséggel benne van:

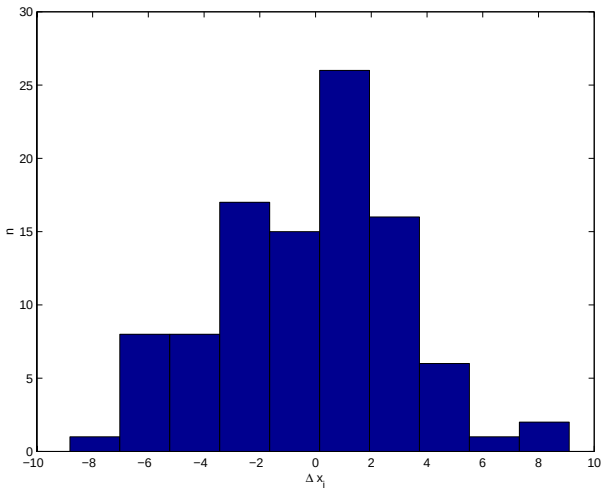
$$p\left[\mu - z_{(b/2)}\sigma' < \hat{\mu} < \mu + z_{(b/2)}\sigma'\right] = 1 - b \tag{10}$$

ahol b rendszerint egy kicsiny szám, hiszen szeretnénk olyan intervallumot kapni, amelyben a mérési eredmény igen nagy valószínűséggel megtalálható. A $z_{(b/2)}$ kifejezés a standard normális eloszlás $1 - b/2$ valószínűséghez tartozó változója. Mivel az eloszlás szimmetrikus, a táblázatok szokásosan csak a Gauss-görbe pozitív fele alatti területet adják meg. A $b/2$ jelölés a le nem fedett területet jelenti.

A fenti kifejezéssel csak az a probléma, hogy mérési eredményre (a mért értékekből számított átlagra) ad egy intervallumot és egy valószínűséget, amelyet az ismeretlen várható érték segítségével fejeztünk ki. A valóságban ez fordítva van: a várható értéket nem ismerjük, de az átlagot, a valószínűségi változót igen. A kifejezés azonban átrendezhető ennek megfelelően:

$$p\left[\hat{\mu} - z_{(b/2)}\sigma' < \mu < \hat{\mu} + z_{(b/2)}\sigma'\right] = 1 - b \tag{11}$$

Azt az intervallumot, amit ez a kifejezés határoz meg, konfidencia-intervallumnak nevezzük, az $1 - b$ értéket pedig konfidenciaszintnek. Figyeljük meg, hogy itt a várható értéket, amely nem valószínűségi



1. ábra. Eltérések a pontos időtől

változó, de ismeretlen, fogjuk közre egy valószínűségi változóval, az átlagértékkel meghatározott intervallummal.

Az órák esetében tehát szükség van a konkrét időértékekre. Az 1. ábrán egy 100 elemű, normális eloszlású mintasorozat hisztogramját látjuk. (A mintákat számítógéppel generáltuk.) A várható érték itt nulla, ez fejezi ki, hogy az órák a várható értékre vonatkozó torzítatlan becslőt szolgáltatnak. A várható érték és a szórás becslői:

$$\hat{\mu} = -0.1809;\ s^* = 3.3535 \tag{12}$$

Az eltérés a pontos időtől 1 perc lehet, így akár a (10), akár a (11) egyenletek alapján:

$$z_{(b/2)} = \frac{\sqrt{N}}{s^*} = \frac{10}{s^*} = 2.9820 \tag{13}$$

ehhez $b/2 = 0.499$ tartozik, tehát a megmenekülés valószínűsége $p = 99.8\%$.

A tapasztalati szórás mint valószínűségi változó

A fenti eredmény várakozásainkkal összhangban van. Az azonban, amit a (7) egyenlettel kikötöttünk, nem igaz. s^* kifejezésében ugyanis az x_i minták szerepelnek, amelyek valószínűségi változók, ebből következik, hogy maga s^* is valószínűségi változó. A négyzetre emelés miatt a kezdetben normális eloszlású valószínűségi változóink eloszlása nem lesz normális.

Tekintsük először az n független, normális eloszlású valószínűségi változó négyzetének összegét:

$$\chi_n^2 = \sum_{i=1}^{n} z_i^2 \tag{14}$$

Ezt n szabadságfokú chinégyzet-eloszlásnak nevezzük, amelynek sűrűségfüggvénye a következő:

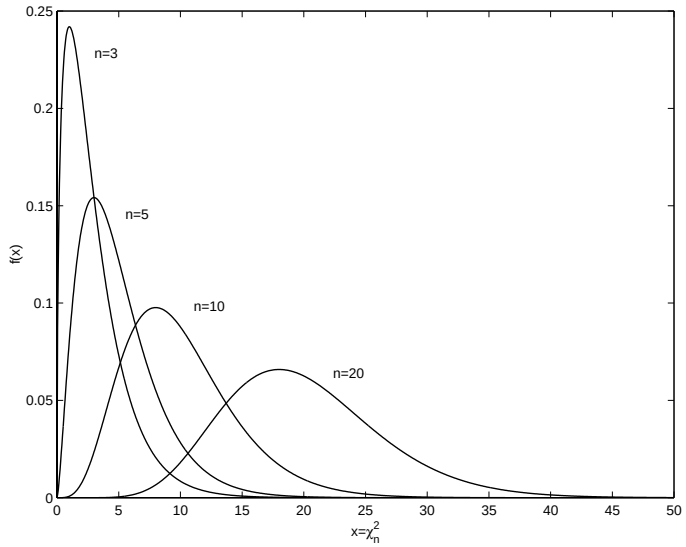
$$f(y) = \left[2^{\frac{n}{2}}\Gamma(\frac{n}{2})\right]^{-1} y^{\frac{n}{2}-1}e^{-\frac{y}{2}};\ y = \chi_n^2 \tag{15}$$

ahol $\Gamma(.)$ az ún. Gamma-függvény:

$$\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt \tag{16}$$

A Gamma-függvény a faktoriális függvény kiterjesztése. Egész számok esetén:

$$\Gamma(k+1) = k! \tag{17}$$



2. ábra. Chinégyzet-eloszlás

A sűrűségfüggvény alakja néhány n-re a 2. ábrán látható. A sűrűségfüggvény aszimmetrikus, bár $n \to \infty$-re – a centrális határeloszlás-tétel miatt – normális eloszláshoz tart. A chinégyzet-eloszlás várható értéke és szórása:

$$E\left\{\chi_n^2\right\} = n; \ \text{var}\left\{\chi_n^2\right\} = 2n \tag{18}$$

Belátható, hogy rögzített $n = N$-re:

$$s^{*2} = \frac{1}{N-1}\chi_{N-1}^2\sigma^2 \tag{19}$$

azaz a tapasztalati szórásnégyzet $N-1$-edfokú chinégyzet-eloszlást követ. Most már meghatározható az s^{*2}-re vonatkozó valószínűségi intervallum:

$$p\left[\frac{\sigma^2}{N-1}\chi_{N-1,(b/2)}^2 < s^{*2} < \frac{\sigma^2}{N-1}\chi_{N-1,(1-b/2)}^2\right] = 1-b \tag{20}$$

A szórásnégyzetre vonatkozó konfidencia-intervallum ezek után:

$$p\left[\frac{s^{*2}(N-1)}{\chi_{N-1,(1-b/2)}^2} < \sigma^2 < \frac{s^{*2}(N-1)}{\chi_{N-1,(b/2)}^2}\right] = 1-b \tag{21}$$

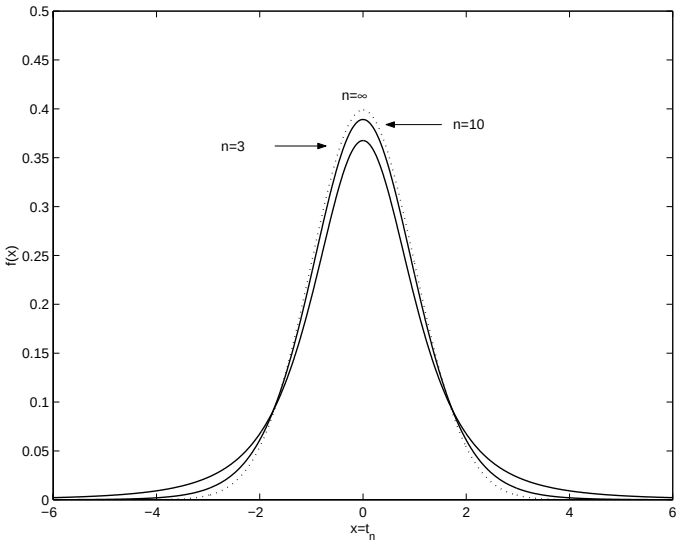
Figyeljük meg, hogy az aszimmetria miatt a két oldalon nem ugyanaz a χ_{N-1}^2 érték van. Mivel a konfidencia-intervallum alulról mindenképpen korlátos (a szórás nem lehet negatív), gyakran egyoldali konfidencia-intervallumot adnak meg:

$$p\left[\sigma^2 < \frac{s^{*2}(N-1)}{\chi_{N-1,(b)}^2}\right] = 1-b \tag{22}$$

Visszatérve az eredeti feladatra, kiszámíthatjuk a szórásnégyzetre, illetve a szórásra vonatkozó egyoldali konfidencia-intervallumot. Mivel erre vonatkozóan a feladatban semmi sem szerepel, vegyük fel a konfidenciaszintet önkényesen 99%-ra! A számítást a (12) helyen adott adatokkal elvégezve:

$$p\left[\sigma^2 < \frac{99s^{*2}}{\chi_{99,(0.01)}^2} = 16.089\right] = 99\% \tag{23}$$

azaz a szórás 99%-os konfidenciaszinttel kisebb, mint 4.0111. A megadott adatok alapján tehát 1% a valószínűsége annak, hogy a valódi szórás ennél nagyobb.



3. ábra. Student t-eloszlás

Megoldás ismeretlen szórás esetén

Kérdés, hogy ezek után hogyan számítható ki a várható értékre vonatkozó konfidencia-intervallum. Megtehetnénk, hogy a (11) egyenletbe valamely igen magas konfidenciaszinttel meghatározott maximális szórást helyettesítünk be, a worst case hibaösszegzés mintájára, de az így kiszámított valószínűségek nem lennének pontosak, hiszen a nagyobb szórásnak a valószínűsége is kisebb, márpedig a konfidenciaszámítással éppen a korrekt valószínűségek megadása a célunk.

Mivel a szórást nem ismerjük, a (9) egyenlettel adott z helyett a

$$z' = \frac{\hat{\mu} - \mu}{s^*/\sqrt{N}} \tag{24}$$

változót kell tekintenünk. A számláló normális eloszlású, a nevező chinégyzet-eloszlást követ. Együtt egy új eloszlást határoznak meg, ehhez tekintsük az alábbi valószínűségi változót:

$$t_n = \frac{z}{\sqrt{\frac{y}{n}}} \tag{25}$$

ahol z normális eloszlású, y pedig n-edfokú chinégyzet-eloszlást követ. Ezt az eloszlást n szabadságfokú Student t-eloszlásnak nevezzük, amelynek sűrűségfüggvénye a következő:

$$f(t_n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})}\left(1 + \frac{t_n^2}{n}\right)^{-\frac{n+1}{2}} \tag{26}$$

A sűrűségfüggvény alakja néhány n-re a 3. ábrán látható. A sűrűségfüggvény szimmetrikus, és $n \to \infty$-re – a centrális határeloszlás-tétel miatt – ez is normális eloszláshoz tart. A görbe alakja már n viszonylag kis értékeire is a normális eloszlásra hasonlít. A Student t-eloszlás várható értéke és szórása:

$$E\{t_n\} = 0;\ \mathrm{var}\{t_n\} = \frac{n}{n-2} \tag{27}$$

Belátható, hogy rögzített $n = N$-re:

$$z' = t_{N-1} \tag{28}$$

azaz a (24) egyenlettel definiált z' valószínűségi változó $N-1$-edfokú Student t-eloszlást követ. Most már meghatározható a z'-re vonatkozó valószínűségi intervallum:

$$p\left[-t_{N-1,(b/2)} < z' < t_{N-1,(b/2)}\right] = 1-b \tag{29}$$

A várható értékre vonatkozó konfidencia-intervallum behelyettesítés és rendezés után:

$$p\left[\hat{\mu} - t_{N-1,(b/2)}\frac{s^*}{\sqrt{N}} < \mu < \hat{\mu} + t_{N-1,(b/2)}\frac{s^*}{\sqrt{N}}\right] = 1-b \tag{30}$$

Visszatérve ismét az eredeti feladatra, kiszámíthatjuk a várható értékre vonatkozó konfidencia-intervallumot, most már a (7) feltétel nélkül. Az eltérés a pontos időtől most is 1 perc lehet, így a (30) egyenlet alapján:

$$t_{(b/2)} = \frac{\sqrt{N}}{s^*} = \frac{10}{s^*} = 2.9820 \tag{31}$$

ehhez $b/2 = 0.4984$ tartozik, tehát a megmenekülés valószínűsége $p = 99.7\%$.

Jól látható, hogy a normális eloszlást, illetve a Student t-eloszlást feltételező megoldás csaknem megegyezik, bár – várakozásainkkal összhangban – az utóbbi ugyanarra az intervallumra kisebb konfidenciaszintet ad meg. Ez azonban olyan csekély eltérés, amely akár a számítási pontatlanságokból is eredhet.

Megoldás 10 kakukkos óra esetén

A kétféle megoldás összevetésére tegyük most fel, hogy mindössze 10 kakukkos óra áll a törpök rendelkezésére. (Ismét számítógéppel generáltunk mintákat.) Ebben az esetben a hisztogram nem túl informatív, a normális eloszlású mintákra a várható érték és a szórás becslői:

$$\hat{\mu} = -0.0873;\ s^* = 3.2466 \tag{32}$$

A normális eloszlás feltételezésével számolt megmenekülési valószínűség (lásd (13) egyenlet) itt most $p \approx 67\%$. A szórásra vonatkozó egyoldali konfidencia-intervallum (lásd (23) egyenlet):

$$p\left[\sigma^2 < \frac{9s^{*2}}{\chi^2_{9,(0.01)}} = 45.389\right] = 99\% \tag{33}$$

azaz a szórás 99%-os konfidenciaszinttel kisebb, mint 6.7371, ami természetesen tágabb intervallum, mint 100 minta feldolgozása esetén. A Student-t eloszlással számolt megmenekülési valószínűség (lásd (31) egyenlet) ebben az esetben $p \approx 64.5\%$, azaz egyértelműen kisebb, mint normális eloszlást feltételezve.

A feladat fix konfidencia-intervallumot határozott meg, és az ehhez tartozó szintek voltak különbözők. Ezzel megegyező komplexitású feladat az adott konfidenciaszintekhez tartozó intervallumok meghatározása. Ilyenkor a normális eloszlás feltételezésével szűkebb intervallumot (kisebb hibát) kapunk, mintha Student-t eloszlással számoltunk volna.