Matematikai statisztika
Témakör 5: Konfidenciaintervallumok

Elek Péter

1. Alapok

Konfidenciaintervallumok

- Eddig *pontbecslések* szerepeltek a kurzus során (azaz a becslés eredménye egy konkrét numerikus érték vagy vektor volt).
  - Az (aszimptotikusan torzítatlan) pontbecslés mintavételi szórása közvetett információt ad a becslés jóságáról (azaz hogy várhatóan milyen messze van a becslés a tényleges paramétertől).
- *Konfidenciaintervallum* esetén a becslés eredménye egy intervallum.

Példa

- Megmértük 16 doboz gabonapehely tömegét, és a kapott eredmények:

$$506, 508, 499, 503, 504, 510, 497, 512,$$
$$514, 505, 493, 496, 506, 502, 509, 496.$$

- Tegyük fel, hogy a gyártási folyamat eredményeképpen a tömeg normális eloszlású ismeretlen μ várható értékkel és $\sigma = 5$ szórással. Adjunk 90%-os konfidenciaintervallumot μ-re!

Megoldás

- Tudjuk, hogy $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, azaz $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Ezért

$$\Pr\left(z_{0.05} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{0.95}\right) = 0.9.$$

- Ez a kifejezés $\overline{X}$-ra (illetve pontosabban $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$-re) vonatkozik.

- Ekvivalens felírás (mivel $z_{0.05} = -z_{0.95}$):

$$\Pr\left(\overline{X} - z_{0.95}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{0.95}\frac{\sigma}{\sqrt{n}}\right) = 0.9.$$

- Tehát a $\left[\overline{X} - z_{0.95}\frac{\sigma}{\sqrt{n}}; \overline{X} + z_{0.95}\frac{\sigma}{\sqrt{n}}\right]$ *véletlen intervallum* 90% valószínűséggel tartalmazza μ-t.

Megoldás (folyt.)

- Példánkban $\overline{x} = 503.8$, $\sigma = 5$, $n = 16$, $z_{0.05} = -1.645$, $z_{0.95} = 1.645$,
- tehát $\overline{x} - z_{0.95}\frac{\sigma}{\sqrt{n}} = 501.7$ és $\overline{x} + z_{0.95}\frac{\sigma}{\sqrt{n}} = 505.8$.
- Ezért a 90%-os konfidenciaintervallum $[501.7; 505.8]$.

Megjegyzések

- A klasszikus statisztikában *nem mondhatjuk*, hogy 90% annak valószínűsége, hogy μ a $[501.7; 505.8]$ intervallumban van, mivel μ nem valószínűségi változó (μ vagy benne van az intervallumban, vagy nem)
- Pontos értelmezés: az $\{X_i\}$ minta vétele *előtt* 90% annak valószínűsége, hogy a véletlen intervallum tartalmazza μ-t.
- Később: Bayes-statisztikában az intervallumbecslések értelmezése természetesebb, hiszen beszélhetünk μ (a posteriori) eloszlásáról.

2. Várható érték becslése

μ konfidenciaintervalluma nagy minta esetén

- Tudjuk, hogy f.a.e. minta esetén $\frac{\overline{X}-\mu}{s^*/\sqrt{n}} \overset{A}{\sim} N(0,1)$.
- Tehát a korábbiakhoz hasonlóan:

$$\Pr\left(z_{\alpha/2} < \frac{\overline{X} - \mu}{s^*/\sqrt{n}} < z_{1-\alpha/2}\right) = 1 - \alpha$$

- azaz (felhasználva, hogy $z_{\alpha/2} = -z_{1-\alpha/2}$),

$$\Pr\left(\overline{X} - z_{1-\alpha/2} * \frac{s^*}{\sqrt{n}} < \mu < \overline{X} + z_{1-\alpha/2} * \frac{s^*}{\sqrt{n}}\right) = 1 - \alpha.$$

- A μ paraméter $1-\alpha$ szintű konfidenciaintervalluma:

$$\left[\overline{x} - z_{1-\alpha/2} * \frac{s^*}{\sqrt{n}}; \overline{x} + z_{1-\alpha/2} * \frac{s^*}{\sqrt{n}}\right].$$

- Szokásos szabály a nagymintás közelítésre: $n > 120$ (vagy akár $n > 60$)
  - de ez természetesen nem egy precíz állítás

Konfidenciaintervallum normális mintában: ismert σ

- *Tegyük fel*, hogy $X \sim N\left(\mu, \sigma^2\right)$.
- Ekkor $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.
- Ezért

$$\begin{aligned} \Pr\left(z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) &= 1-\alpha \\ \Pr\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1-\alpha. \end{aligned}$$

- Tehát a μ paraméter $1-\alpha$ szintű konfidenciaintervalluma:

$$\left[\overline{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Konfidenciaintervallum normális mintában: ismeretlen σ

- *Tegyük fel*, hogy $X \sim N\left(\mu, \sigma^2\right)$.
- Ekkor $\frac{\overline{X}-\mu}{s^*/\sqrt{n}} \sim t_{n-1}$.
- Tehát

$$\begin{aligned} \Pr\left(t_{\alpha/2;n-1} < \frac{\overline{X}-\mu}{s^*/\sqrt{n}} < t_{1-\alpha/2;n-1}\right) &= 1-\alpha \\ \Pr\left(\overline{X} - t_{1-\alpha/2;n-1}\frac{s^*}{\sqrt{n}} < \mu < \overline{X} + t_{1-\alpha/2;n-1}\frac{s^*}{\sqrt{n}}\right) &= 1-\alpha. \end{aligned}$$

- Tehát a μ paraméter $1-\alpha$ szintű konfidenciaintervalluma:

$$\left[\overline{x} - t_{1-\alpha/2;n-1} * \frac{s^*}{\sqrt{n}}; \overline{x} + t_{1-\alpha/2;n-1} * \frac{s^*}{\sqrt{n}}\right].$$

Feladat

- A túlfoglalás elfogadható mértékének meghatározása érdekében egy légitársaság a járatain el nem foglalt helyek számára szeretne konfidenciaintervallumot adni.
- Ezért véletlenszerűen kiválasztanak $n = 225$ hasonló járatot, és azt találják, hogy ezeken az el nem foglalt helyek átlaga $\overline{x} = 11.6$ fő, korrigált tapasztalati szórása pedig $s^* = 4.1$ fő.
- Határozzuk meg az el nem foglalt helyek várható számának 90%-os, 95%-os és 99%-os konfidenciaintervallumát!

3. Sokasági arány nagymintás becslése

p **konfidenciaintervalluma nagy minta esetén**

- Legyen $X \sim \text{Bernoulli}(p)$.
- Nagy mintában ekkor $\frac{\overline{X}-\mu}{s^*/\sqrt{n}} \overset{A}{\sim} N(0,1)$ és $\frac{\overline{X}-\mu}{s/\sqrt{n}} \overset{A}{\sim} N(0,1)$.
- Mivel $s^2 = \overline{X}(1-\overline{X})$, ezért $\frac{\overline{X}-p}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}} \overset{A}{\sim} N(0,1)$.
- Ezért

$$\Pr\left(z_{\alpha/2} < \frac{\overline{X}-p}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}} < z_{1-\alpha/2}\right) = 1-\alpha.$$

- Tehát a (közelítően) $1-\alpha$ szintű konfidenciaintervallum:

$$\left[\overline{x} - z_{1-\alpha/2} * \sqrt{\frac{\overline{x}(1-\overline{x})}{n}}; \overline{x} + z_{1-\alpha/2} * \sqrt{\frac{\overline{x}(1-\overline{x})}{n}}\right].$$

- Szokásos szabály a nagymintás közelítésre: $np > 10$ és $n(1-p) > 10$
  - bár ennél nagyobb "szerencsétlen" paraméterek is vannak.

Feladat

- A KSH (vagy más országok statisztikai hivatalainak) munkaerő-felmérése (MEF, LFS) a foglalkoztatás becslésére szolgál.
- Magyarországon 2013 első negyedévében, 54.000 fő megkérdezése alapján, a munkaképes korú (15-74 éves) korosztály foglalkoztatási rátája 51,5%-nak adódott.

- Adjuk meg a foglalkoztatási ráta 95%-os konfidenciaintervallumát!
  - Feltételezzünk f.a.e. mintát, bár a valóságban rétegzett mintavétel van.
- Magyarország munkaképes korú lakossága körülbelül 7.630.000 fő. Adjunk 95%-os konfidenciaintervallumot a foglalkoztatottak létszámára 2013 első negyedévében!

4. Sokasági variancia becslése

σ^2 konfidenciaintervalluma normális mintában

- Normális eloszlású mintában $\frac{(n-1)s^{*2}}{\sigma^2} = \frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}$.
- Ekkor

$$\begin{aligned}
\Pr\left(k_{\alpha/2;n-1} < \frac{ns^2}{\sigma^2} < k_{1-\alpha/2;n-1}\right) &= 1-\alpha \\
\Pr\left(\frac{ns^2}{k_{1-\alpha/2;n-1}} < \sigma^2 < \frac{ns^2}{k_{\alpha/2;n-1}}\right) &= 1-\alpha.
\end{aligned}$$

- A σ^2 paraméter $1-\alpha$ szintű konfidenciaintervalluma:

$$\left[\frac{ns^2}{k_{1-\alpha/2;n-1}}; \frac{ns^2}{k_{\alpha/2;n-1}}\right].$$

Feladat

- Tudjuk, hogy a férfiak illetve a nők testmagassága normális eloszlású.
- Egy $n = 100$ elemű f.a.e. mintában a korrigálatlan tapasztalati szórásnégyzet $s^2 = 169$-nek adódott.
- Határozzuk meg a sokasági variancia 95%-os konfidenciaintervallumát!
- Határozzuk meg a sokasági variancia (közelítő) konfidenciaintervallumát normális közelítést használva is!

Tananyag

Tananyag

- Wooldridge Appendix C.5
- Amemiya 8.1-8.2