A társadalomkutatás módszerei I.

10. hét

Daróczi Gergely

Budapesti Corvinus Egyetem

2011. november 17.

Notes

Outline

Notes

A mintavételi hiba és konfidencia-intervallum

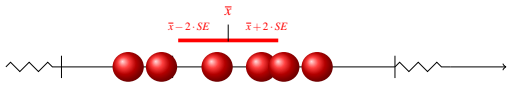
Elmélet

Szükséges képletek:

- **számtani átlag**: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- **korrigált empirikus szórás**: $S^* = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}$ (nem Zh kérdés!)
- **standard/mintavételi hiba**: $SE = \frac{S^*}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} \approx \frac{S^*}{\sqrt{n}}$
- **konfidencia-intervallum**: $\bar{x} \pm z \cdot SE$, ahol legtöbbször $z = 1,96$

Tehát:

- **konfidencia-intervallum**: $= [\bar{x} - 2 \cdot SE; \bar{x} + 2 \cdot SE]$



Notes

A mintavételi hiba és konfidencia-intervallum
Gyakorlat

„Az őszi kutatásban is megkérdezték az autósokat az üzemanyagárak lélektani határáról. A felmérés közben hétről hétre dőltek meg az üzemanyagár csúcsok, ezért a kutatás a 400 és a 450 forint közötti literenkénti ársávot vizsgálta. A gázolaj árának hatását most is rugalmasabban ítélték meg az autósok, még mindig sokan vannak, akik 450 forint feletti áron is ugyanannyit tankolnának, mint most. A benzinnél 420 forintos árnál a válaszadók többsége már nem tankolna annyit mint korábban, s jelentősen csökkentené az autó használatát."

Forensis Autóklub (2011.november)

Notes

A mintavételi hiba és konfidencia-intervallum
Gyakorlat

„Mi az az üzemanyag ár, ahol már hosszútávra leállítanád az autódat és nem tankolnál rendszeresen?"

410, 420, 420, 430, 500, 450, 400, 425, 460

Leíró statisztikák:

- **számtani átlag**: $\bar{x} = \frac{410+420+420+430+500+450+400+425+460}{9} = 435$
- **medián**: 425
- **módusz**: 420
- **minimum érték**: 400
- **maximum érték**: 500
- **terjedelem**: 100
- **szórás/variancia**: *nem Zh kérdés*

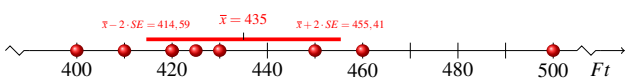
Notes

A mintavételi hiba és konfidencia-intervallum
Gyakorlat

„Mi az az üzemanyag ár, ahol már hosszútávra leállítanád az autódat és nem tankolnál rendszeresen?"

410, 420, 420, 430, 500, 450, 400, 425, 460

- **számtani átlag**: $\bar{x} = \frac{410+420+420+430+500+450+400+425+460}{9} = 435$
- **korrigált empirikus szórás**: $S^* = 30,619$
- **standard/mintavételi hiba**: $SE = \frac{30,619}{\sqrt{9}} = \frac{30,619}{3} = 10,206$
- **konfidencia-intervallum**: $435 \pm 2 \cdot 10,206 = [414,59; 455,41]$



Notes

A mintavételi hiba és konfidencia-intervallum
Példák

„Az "új fizika" lehetőségét vetíti előre az a részecskebomlási "anomália", amelyet az Európai Nukleáris Kutatási Szervezet (CERN) nagy hadronütköztetőjében (LHC) észleltek.

Matthew Charles, az Oxfordi Egyetem fizikusának beszámolója szerint a D-mezon szubatomi részecskék kissé másként bomlanak, mint antirészecskéik. A felfedezés segíthet megérteni, hogy a világegyetemben miért több az anyag, mint az antianyag.

Egyelőre azonban újabb vizsgálatok szükségesek, jelenleg ugyanis statisztikailag mindössze 0,05 százalék a valószínűsége, hogy eredményeik nem véletlenszerűek.”

Forrás: index.hu

Notes

A mintavételi hiba és konfidencia-intervallum
Példák

A módszertan haszna. EP választások 2009: „Hajszálpontos mérés”

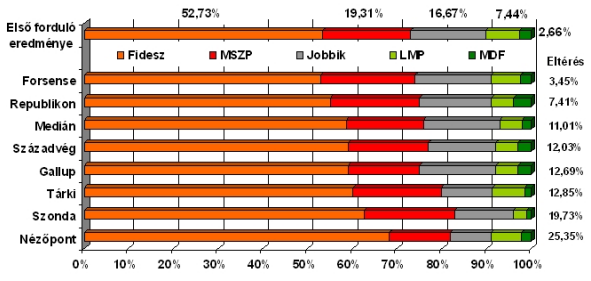
	Nézőpont		Tárki	Medián	*NRC*	
	BSZ	BSZP	BSZP	??	??	**eredmény**
Fidesz	54%	**66%**	**70%**	60%	50%	**56,4%**
MSZP	12%	14%	17%	21%	**26%**	**17,4%**
Jobbik	6%	7%	4%	7%	**13%**	**14,8%**
MDF	5%	6%	**1%**	4%	4%	**5,3%**
SZDSZ	3%	4%	3%	4%	3%	**2,2%**

	Nézőpont	TÁRKI	Medián	*NRC*
Kutatás ideje	V. 20-22.	V. 7-20	V. 22-26.	n.a.
Módszer	Telefonos lekérdezés	Személyes lekérdezés (?)	Személyes lekérdezés	Online kérdőív
Megkérdezettek száma	1000	1000	1200	1000

Forrás: Dr. Bartus Tamás előadásanyagai

Notes

A mintavételi hiba és konfidencia-intervallum
Példák



Forrás: spss.hu

Notes

A mintavételi hiba és konfidencia-intervallum

Példák

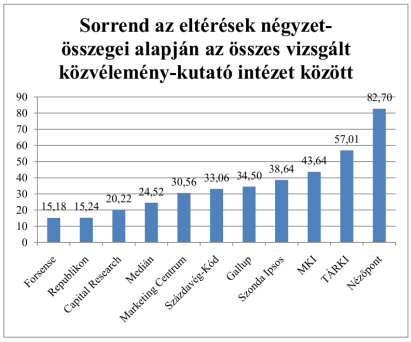
Kutatóintézet (mintanagyság)[49]	A vonatkozó részminta konfidencia intervalluma[50]	Fidesz	MSZP	Jobbik	MDF	LMP	Összesen
Forsense (N=530)	+/-4,743	7,27	0,31	4,67	0,34	1,44	**14,03**
Medián (N=n/a)	n/a	7,27	2,31	0,33	0,66	2,44	**11,01**
Századvég-Kód (N=520>n)	+/-6,6%<	6,27	1,31	1,67	0,34	2,44	**12,03**
Gallup (N=1014>n)	+/-4,4<	3,31	3,31	0,33	0,34	2,44	**12,69**
Szonda (N=795>n)	+/-3,825<	10,69	0,69	3,67	1,66	4,44	**19,73**
Nézőpont (N=1000>n)	+/-3,2<	4,31	4,31	5,67	1,34	0,44	**25,35**
Átlag	+/-	**6.52**	**2,04**	**2,72**	**0,78**	**2,27**	**15,8**

Forrás: Metz Rudolf Tamás – A 2010-es országgyűlési választások előrejelzései és azok eltérései

Notes

A mintavételi hiba és konfidencia-intervallum

Példák

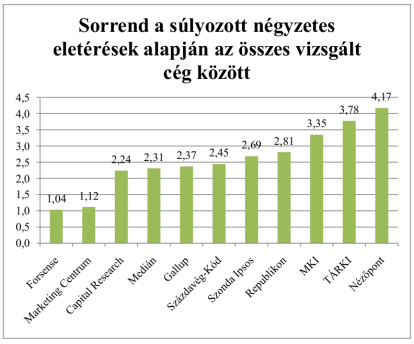


Forrás: Kópházi Dániel – A politikai közvélemény-kutatások megbízhatósága

Notes

A mintavételi hiba és konfidencia-intervallum

Példák



Forrás: Kópházi Dániel – A politikai közvélemény-kutatások megbízhatósága

Notes

A mintavételi hiba dichotóm változók esetében

Elmélet

Bernoulli-eloszlás:

- diszkrét, dichotóm valószínűségi változó
- p valószínűséggel 1, q $(= 1 - p)$ valószínűséggel 0 értéket vesz fel
- **átlag**: p
- **medián**: –
- **módusz**: $\begin{cases} 0 & \text{if } q > p \\ 0,1 & \text{if } q = p \\ 1 & \text{if } q < p \end{cases}$
- **szórás**: $\sqrt{p(1-p)}$
- **variancia**: $p(1-p)$
- **standard/mintavételi hiba**: $SE = \frac{S^*}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} \approx \frac{S^*}{\sqrt{n}} \approx \frac{\sqrt{p(1-p)}}{\sqrt{n}}$
- **konfidencia-intervallum**: $\bar{x} \pm z \cdot SE$, ahol legtöbbször $z = 1,96$

Notes

A mintavételi hiba dichotóm változók esetében

Pesszimista megközelítés

Bernoulli-eloszlás:

- a várható legnagyobb mintavételi hibával számolunk,
- a mérési hiba a szórás és a mintaelemszám függvénye,
- a mintaelemszám növelésével csökkenthető a mintavételi hiba,
- ha egy mintában magas a szórás, magas lesz a mintavételi hiba.

Milyen p érték mellett lesz a lehető legmagasabb a szórás?

$$S^* = \sqrt{p(1-p)}$$

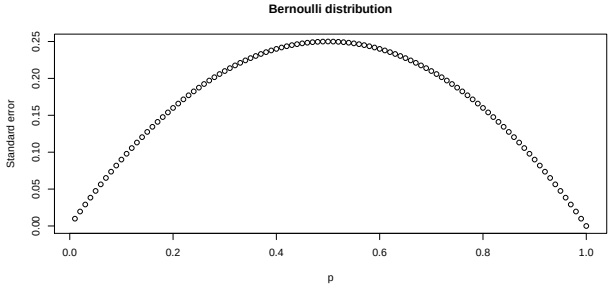
$$p = 0.5$$

$$VAR(x) = 0.5 \cdot (1 - 0.5) = 0.5^2 = 0.25$$

Notes

A mintavételi hiba dichotóm változók esetében

Pesszimista megközelítés



standard/mintavételi hiba: $SE = \frac{S^*}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} \approx \frac{S^*}{\sqrt{n}} \approx \frac{\sqrt{p(1-p)}}{\sqrt{n}}$

Notes

A mintavételi hiba dichotóm változók esetében

Mintanagyság meghatározása

Mekkora mintára van szükségem ahhoz, hogy egy párt támogatottságát plusz/mínusz 2 százalék pontossággal mérjem?

- 2 százalék pontosság 95 %-os döntési szinten: $SE = 1$,
- várható legnagyobb szórásnégyzet százalékos értékeknél: $50 \cdot (100 - 50) = 2500$
- $SE = \frac{S^*}{\sqrt{n}}$

$$\Downarrow$$

- $1 = \frac{\sqrt{2500}}{\sqrt{n}}$

$$\Downarrow$$

- $1 \cdot \sqrt{n} = \sqrt{2500}$
- $n = 2500$

Notes

Mintanagyság meghatározása általános esetben

Példa

Mekkora mintára van szükségem ahhoz, hogy 5 perc pontosság meg tudjam állapítani a napi tévénézésre fordított idő hosszát a felnőtt magyar lakosság körében?

- 5 perc pontosság 95 %-os döntési szinten: $SE = 2.5$,
- becsült szórás: 10
- $SE = \frac{S^*}{\sqrt{n}}$

$$\Downarrow$$

- $2,5 = \frac{10}{\sqrt{n}}$

$$\Downarrow$$

- $2,5 \cdot \sqrt{n} = 10$
- $\sqrt{n} = 4$
- $n = 16$

Notes

Mintanagyság meghatározása általános esetben

Példa

Mekkora mintára van szükségem ahhoz, hogy 1 perc pontosság meg tudjam állapítani a napi tévénézésre fordított idő hosszát a felnőtt magyar lakosság körében?

- 1 perc pontosság 95 %-os döntési szinten: $SE = 0.5$,
- becsült szórás: 10
- $SE = \frac{S^*}{\sqrt{n}}$

$$\Downarrow$$

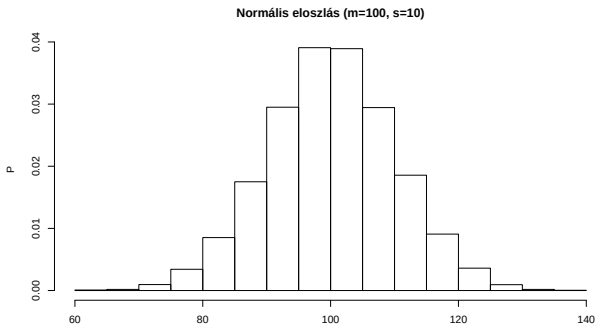
- $0,5 = \frac{10}{\sqrt{n}}$

$$\Downarrow$$

- $0,5 \cdot \sqrt{n} = 10$
- $\sqrt{n} = 20$
- $n = 400$

Notes

Mintanagyság meghatározása általános esetben
Példa



Notes

Mintanagyság meghatározása általános esetben
Példa

Mekkora mintára van szükségem ahhoz, hogy 5 perc pontosság meg tudjam állapítani a napi tévénézésre fordított idő hosszát a felnőtt magyar lakosság körében?

- 5 perc pontosság 95 %-os döntési szinten: $SE = 2.5$,
- becsült szórás: 100
- $SE = \frac{S^*}{\sqrt{n}}$
- $2,5 = \frac{100}{\sqrt{n}}$
- $2,5 \cdot \sqrt{n} = 100$
- $\sqrt{n} = 40$
- $n = 1600$

Annál nagyobb minta kell, . . .

- minél nagyobb pontosságra törekszem,
- minél nagyobb a vizsgált változó szórása a populációban.

Notes

Torzítatlanság és reprezentativitás
Elmélet

Amennyiben

- X: az a változó, amiről meg akarunk tudni valamint (vizsgálati változó),
- és Y: tetszőleges NEM vizsgálati változó, melynek paramétere ismert,

akkor:

A minta torzítatlan
ha X mintabeli átlaga = X valós átlaga.

A minta reprezentatív
ha Y mintabeli átlaga = Y valós átlaga.

De:

- ezek közül melyik megismerhető?

Notes

Torzítatlanság és reprezentativitás
A torzítatlanság típusai

Szelekciós torzítás:

- a mintába kerülés a vizsgált változó,
- vagy azzal összefüggő, látens dimenzió függvénye.

A reprezentativitás hiányából fakadó torzítás:

- a mintába kerülés valószínűsége összefügg egy megfigyelt, de nem vizsgált változóval,
- amely változó eloszlása eltér a populációbeli ismert eloszlástól.

Kérdés:

- A vizsgált változó összefügg-e az említett változóval?

L. az NRC eredményeit az EP választással kapcsolatban!

Notes

Torzítatlanság és reprezentativitás
Példa

Miért nem tudta előrejelezni a Literary Digest 1936-ban Roosevelt újraválasztását?

Torz mintavételi keret használata:

- A *Digest* mintavételi kerete: gépkocsi tulajdonosok és telefon-előfizetők listája, ahol
- nagyobb arányban fordulnak elő konzervatív (jómódú) szavazók.

Szelektív válaszmegtagadás:

- A *Digest* által kiküldött kérdőíveknek „csak” 22 százaléka érkezett vissza!
- És a visszaküldés a pártpreferencia függvénye: a kérdőívet alacsonyabb arányban küldték vissza a demokrata szavazók.

Notes

Torzítatlanság és reprezentativitás
Példa

Vizsgáljuk a magyar felnőtt lakosság jövedelmi viszonyait!

Torz mintavételi keret használata:

- Mintavételi keret legyen a mobiltelefon-előfizetők listája, ahol
- a kevéssé tehetősek nem jelennek meg, ill.
- a keret akkor is torzított, ha a leggazdagabbak titkosítják számukat.

Szelektív válaszmegtagadás:

- Kisebb eséllyel készül interjú azokkal, akik sokat dolgoznak,
- és akik sokat dolgoznak, valószínűleg sokat is keresnek.

Notes

Egy elrettentő példa

„A szavazás lezárult, kiderült tehát, hogy a Nemzeti Sport SMS-ben szavazó olvasói kit láttak a világ legjobbjának az elmúlt esztendőben. Három kategóriában viaskodtak a legek, harmadszorra a csapatok versengésének végeredményét ismertetjük. A szavazók szerint 2005-ben a Barcelona labdarúgócsapata volt a legjobb!

Reprezentatív a minta, elvégre lapunk olvasói csak elenyésző részét képezik a labdarúgásról véleményt formálók táborának, ám él bennünk a gyanú, hogy ha a Nemzeti Sport globálisan hirdette volna meg szimpátiaszavazását, akkor is a Barcelona érdemelte volna ki «A világ legjobb csapata» címet."

Forrás: *Nemzeti Sport* (2006. 01. 06.)

Notes

Köszönöm a figyelmet!

Daróczi Gergely
daroczi.gergely@btk.ppke.hu

Notes

Notes