s.s. Bere Anikó Zsuzsanna

Statisztikai módszerek a fizikában

- **statisztika:** adatokon alapuló kísérlettervezési, gyűjtési, rendezési, összesítési, ábrázolási, analizálási, értelmezési és következtetési módszerek összessége
  - ***vizsgálat tervezése***
    - **megfigyeléses vizsgálat:** változtatás nélküli külső megfigyelés.
    - **kísérlet:** valamilyen kezelést végzünk, majd megfigyeljük a hatásait
  - ***időbeli lefolyás***
    - **keresztmetszeti vizsgálat:** az adatokat egy időpontban mérjük, gyűjtjük
    - **utólagos vizsgálat:** múltbéli adatokat használunk
    - **előre tervezett:** az adatokat a jövőben gyűjtjük olyan csoportokból melyek valamilyen közös faktorban megegyeznek
  - ***mintavételezése***
    - **teljesen véletlen:** a populáció minden tagjának ugyan akkora esélye van arra hogy a mintába bekerüljön
    - **szisztematikus:** valamilyen kezdőponttól indulva kiválasztjuk minden K-adik elemet a populációból
    - **kényelmes:** használjuk azt a mintát amit a legkönnyebb beszerezni
    - **rétegzett:** oszd fel a populációt kettő vagy több csoportra, melyeken belül bizonyos fontos tulajdonságok azonosak, vagy hasonlók majd vegyél mintát mindegyik rétegből
    - **klaszter:** oszd a populációt valamilyen természetes módon klaszterekre, véletlenül válassz közülük, használd az összes tagot.
  - ***buktatói***
    - **rossz minták:** melyekre nem igaz a Benford törvény (haranggörbe szerint helyezkednek el a minta elemei)
    - **publikációs elfogultság:** csak vagy főleg olyan mintákat mutatok be amik alátámasztják a megállapításaimat. Az ellenpéldákat nem, vagy nem a valóságos arányban mutatom be.
    - **túl kicsi minták:**
    - **félrevezető grafikák:** mindig figyelnünk kell a megjelenő számokra!
    - **önkéntes választási adatok?!** törekedni kell a jó reprezentálásra
    - **becsapós kérdések, kérdések sorrendje, sugalló kérdések**
    - **részleges ismeretek, hiányzó adatok**
    - **készakart hamisítások**
  - ***korreláció***
    - **A → B; A ← B; A ↔ B**
    - **C → A; B**
- **populáció (alapsokaság):** a tanulmányozandó elemek összessége, teljessége. A gyűjtemény teljes abban az értelemben, hogy tartalmaz minden tanulmányozandó tárgyat.
  - **paraméter:** a populációt jellemző numerikus érték
  - **cenzus:** adatok gyűjteménye a populáció minden eleméről.
- **minta:** a populációból véletlenszerűen kiválasztott elemek halmaza.
  - **statisztika:** a mintát jellemző numerikus érték
- **adatok**: összegyűjtött megfigyelések (mérések, kérdőíves válaszok, felmérések)
  - ***minőség szerinti csoportosítás:***
    - **Kvantitatív adatok**: méréseket vagy leszámlálásokat jellemző számok.
      - ***diszkrét*:** amikor a lehetséges adatok száma véges, vagy legalábbis számlálható.
      - ***folytonos*:** ami végtelen sok lehetséges értéket vehet fel valamilyen folytonos skálán és nincsenek benne lyukak, szakadások.

- **Kvalitatív adatok**: kategóriákra bonthatók, melyeket valamilyen nem-numerikus jellemzők alapján különböztethetünk meg.

➢ ***mérési szintek szerinti csoportosítás:***
- **Nominális:** nem rendezhető, kategorikus
- **Ordinális:** rendezhető, de nincs értelmezve az adatok különbsége
- **Intervallumos:** értelmezve van az adatok különbsége és rendezhetősége
- **Arányszintű:** van értelmezhető null pontja az adatoknak

➢ ***tulajdonságai:***
- **Középpont:** egy reprezentáns vagy átlag érték ami megmutatja hogy hol van a közepe az adathalmaznak
  - ***centrum:*** az az érték, ami a közepén van az adathalmaznak
  - ***számtani közép:*** $\bar{x} = \frac{\sum x}{n}$ mintára, vagy $\mu = \frac{\sum x}{N}$ populációra.
  - ***súlyozott átlag:*** $\bar{x} = \frac{\sum w \cdot x}{w}$
  - ***medián:*** középső érték, ha az adatok rendezett sorrendben vannak: $\tilde{x}$
    nincs rá hatással egy-egy kiugró érték
  - ***módusz:*** az az érték ami a leggyakrabban fordul elő: M
    nominális adatokra lehet értelmezni
  - ***középtartomány:*** $\frac{legnagyobb - lekisebb}{2}$
- **Terjedelem:** legnagyobb és a legkisebb elem különbsége
- **Percentibilis:** $x\ per.érték = \frac{x\text{–nél kisebb értékek száma}}{\text{összes érték száma}} \cdot 100$
  $L = \frac{k}{100} n$, ahol n az összes adat, k a percentibilis száma
- **Eloszlás:** az adatok eloszlásának természete, alakja
  - ***gyakorisági eloszlás:*** osztályok bevezetése
  - ***relatív gyakorisági eloszlás:*** $rel.gyak. = \frac{osztálygyakoriság}{\text{összes gyakoriság}}$
  - ***kumulatív gyakorisági eloszlás:*** összegző
- **Szórás:** az adatok átlag körüli variabilitásának mértéke
  $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{n\cdot(\sum x^2)-(\sum x)^2}{n\cdot(n-1)}}$ mintára, vagy $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$ populációra
  az értéke dinamikusan nő egy-egy outliers adattól
  mértékegysége megegyezik az adatok mértékegységével
- **Standard eltérés:** $z = \frac{x-\bar{x}}{s} = \sqrt{\frac{n\cdot(\sum w\cdot x^2)-(\sum w\cdot x)^2}{n\cdot(n-1)}}$ mintára,
  vagy $z = \frac{x-\mu}{\sigma}$ populációra
  ha $z < \bar{x} \rightarrow -z$, megszokott értéke $-2 < z < 2$
  - ***Csebisev tétel:*** Az adatok legalább $1 - \frac{1}{K^2}$-ad része mindig közelebb van az átlaghoz, mint K standard eltérés, ahol $K > 1$ pozitív egész
- **Variancia:** szórásnégyzet
  - ***variációs együttható:*** megadja a szórást az átlag százalékában: CV
    $CV = \frac{s}{x} \cdot 100$ mintára, vagy $CV = \frac{\sigma}{\mu} \cdot 100$ populációra
- **Outliers:** olyan mintaértékek melyek a mintaértékek döntő többségétől messze helyezkednek el
- **Idő:** az adatok időben változó tulajdonságai

➢ ***ábrázolás***
- **Hisztogram:** oszlopdiagram

- **Ogiva:** vonaldiagram, ami kumulatív gyakoriságokat mutat
- **Poligon:** egyenes szakaszokkal köti össze az osztály felezőpontok értékeit
- **Dot Plot:** pontként jeleníti meg a diagram adatainak jellemzését.
- **Levél-ág diagram:** minden adatot két részre bontunk: ág és levél
- **Pareto grafikon**: kvalitatív adatokból képzett oszlopdiagram, melynél az oszlopokat nagyság szerint rendezzük.
- **Kördiagram**: kvalitatív adatokat, mint egy torta szeleteit, mutatják be.
- **Szórásdiagram**: adatpárokba rendezett adatok ábrázolása
- **Idősor grafikon**: különböző időpontokban gyűjtött adatok ábrázolására

❖ Valószínűség számítás

➢ *alapfogalmak*

- **esemény:** valamilyen véletlen kísérlet eredményeinek gyűjteménye
- **elemi esemény:** olyan esemény, amit nem lehet egyszerűbb komponensekre bontani
- **esemény tér:** a lehetséges elemi események összessége
- **lehetetlen esemény:** valószínűsége 0
- **bizonyos esemény:** valószínűsége 1
- **független események:** ha az egyik esemény bekövetkezése nem befolyásolja a másik esemény bekövetkezésének valószínűségét.
- **valószínűség:** $\mathrm{P(A)} = \frac{\text{bekövetkezés száma}}{\text{összes kísérlet száma}} = \frac{\text{bekövetkezésének esetei}}{\text{összes elemi esemény száma}}$, $0 \leq \mathrm{P} \leq 1$
- **komplementer ($\bar{A}$):** mindazok az események amikor A nem következik be
  - ***igazi esély*:** $esély = \frac{P(A)}{P(\bar{A})} : 1$
  - ***diszjunkt esemény*:** kölcsönösen kizáró események

➢ *műveletek*

- **összetett esemény:** $P(A+B) = P(A) + P(B) - P(AB)$
- **komplementer események:** $P(A) + P(\bar{A}) = 1$
- **multiplikációs szabály:** $P(AB) = P(A) \cdot P(B)$, ha A és B független
- **feltételes vaslószínűség:** $P(B|A) = \frac{P(AB)}{P(A)}$, B bekövetkezik feltéve, ha A is

➢ *valószínűség eloszlások*

- **véletlen változó:** az egyes számértékeit a véletlen kísérlet véletlenszerű kimenetei határoznak meg
  - ***diszkrét*:** vagy véges sok vagy megszámlálhatóan sok számú értéket vehet fel
    - ♦ várható érték: $E = \sum[x \cdot P(x)]$
  - ***folytonos*:** végtelen sok értéket vehet fel, folytonos skálán megadható mérés eredményeiként adódnak és nem tartalmaznak szakadást
    - ♦ egyenletes eloszlás: ha értékei egyenletesen oszlanak el valamilyen intervallumban; téglalap formájú.

❖ Eloszlások

➢ **Valószínűségi eloszlás:** leírás, ami a véletlen változó minden egyes értékéhez hozzárendeli annak valószínűségét

- ***sűrűség függvény:*** egy folytonos valószínűség eloszlás görbéje.
  - a görbe alatti terület 1.
  - a görbe minden pontja nem negítív.
- ***szokásos érték:*** $max.minimuma = \mu \pm 2\sigma$
- ***ritka esemény szabály:*** x siker n próbálkozásból
  - szokatlanul sok: $P(x + \Delta x) \leq 0{,}05$

- szokatlanul kevés: $P(x - \Delta x) \leq 0{,}05$

▪ ***jellemzői*:**
- $0 \leq P(x) \leq 1;\ \sum P(x) = 1$
- **átlag*:*** $\mu = \sum[x \cdot P(x)]$
- **variancia*:*** $\sigma^2 = \sum[(x-\mu)^2 \cdot P(x)] = [\sum x^2 \cdot P(x)] - \mu^2$
- **szórás*:*** $\sigma = \sqrt{\sum[(x-\mu)^2 \cdot P(x)]} = \sqrt{[\sum x^2 \cdot P(x)] - \mu^2}$

➢ **Binomiális eloszlás**

▪ ***feltételei*:**
- a kimenetelt két csoportra lehet osztani
- fixen rögzített számú kísérletet végzünk
- a kísérletek függetlenek
- a siker valószínűsége állandó a különböző kísérletekben

▪ ***jelölések*: n** kísérletek száma; **x** n-próbálkozás közüli sikeres száma; **p** x-valószínűsége; **q** a sikertelen próbálkozások valószínűsége (1-p); **P(x)** annak a valószínűsége, hogy pontosan x próbálkozás lesz sikeres n próbálkozás közül

▪ ***képlet*:** $P(x) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$
- **átlag*:*** $\mu = np$
- **variancia*:*** $\sigma^2 = npq$
- **szórás*:*** $\sigma = \sqrt{npq}$

▪ ***közelítés*:** normális eloszlással
- **ha** $np \geq 5$ és $nq \geq 5$, **akkor** az átlag és a szórás adatok felhasználhatók a normális eloszlás kiszámításához.
- **folytonossági korreláció!** [x-0,5; x+0,5] intervallumot kell nézni

➢ **Poisson eloszlás**

▪ ***feltételei*:**
- ritka események eloszlások leírására használják
- egy adott intervallumra vonatkoztatott *x* események előfordulásának számára
- az előfordulásoknak véletlenszerűnek, függetlennek és egyenletes eloszlásúnak kell lenniük

▪ ***jelölések*: x** a (véletlen változó) események előfordulási száma;

▪ ***képlet*:** $P(x) = \frac{\mu^x e^{-\mu}}{x!}$
- **átlag*:*** μ adott
- **szórás*:*** $\sigma = \sqrt{\mu}$

▪ ***megjegyzés*:** a binomiális eloszlás jól közelíthető Poisson eloszlással ha n > 100 és np < 10 → $\mu = np$.

➢ **Standard normális eloszlás:**

▪ folytonos eloszlású függvény (haranggörbe) – pl: hőmérséklet

▪ ***jellemzője*:**
- **átlaga***: 0*
- **szórása***: 1*
- **sűrűség függvénye alatti terület***: 1*

▪ ***jelölések*: z** távolság a függvényben 0-tól (feltétel); **terület** a valószínűséget jelöli (görbe alatti)

➢ **Normális eloszlás:**

▪ ha nem $\mu = 0$ és $\sigma = 1$ akkor standardizálni kell! $z = \frac{x-\mu}{\sigma}$

▪ ***normalitás vizsgálata:***

- hisztogram készítés
- 1 outliers megengedhető, több nem
- szimmetrikusnak kell lennie

➢ **Statisztika eloszlása**
- a statisztika minden lehetséges értékének eloszlása abban az esetben, amikor értékét a populáció minden lehetséges n elemszámú mintájára kiszámítjuk
- ***az arány eloszlása:*** mintabeli arány eloszlása a populáció minden lehetséges elemszámú mintájára
  - minta arányának átlaga = populáció arányával
  - normális eloszlással közelíthető
- ***az átlag eloszlása:*** mintabeli átlag eloszlása, ha a populációból vett összes lehetséges ne elemszámú mintát vesszük
  - általában képlettel, táblázattal prezentálják
  - változik a bevitt adatok értékeitől → **minta variabilitás**
- ***interpretáció:***
  - **torzítatlan becslés:** a populációs paraméterekhez tartó statisztikák (átlag, variancia, részarány)
  - **torzító becslés:** nem tartanak a populáció paramétereihez (medián, terjedelem, szórás)
- ***központi határeloszlás tétele:***
  - a minta méretét növelve (>30) a minta normáleloszláshoz tart.
  - $\mu_{\bar{x}} = \mu$ és $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
  - ha véges a populáció akkor: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

➢ **Aránybecslés**
- ***feltételei***
  - a minta egy egyszerű véletlen minta
  - binomiális eloszlás feltételei fennállnak
  - van legalább 5 sikeres és 5 sikertelen eset
- ***pontbecslés***
  - **jelölések: p** populáció aránya; $\hat{p} = \frac{x}{n}$ minta aránya; $\hat{q} = 1 - \hat{p}$ minta arány.
  - a mintaarány ($\hat{p}$) a legjobb pontbecslése a populáció aránynak (p).
- ***intervallumbecslés*** (konfidencia intervallum)
  - **megbízhatóság** (konfidencia szint): az az $1 - \alpha$ valószínűség, ami megadja, azon esetek arányát, ahányszor az intervallumbecslés valójában tartalmazza a populáció paraméter érétkét, ha a becslést sokszor megismételjük. (pl: 5%)

konf. szint	α	$z_{\frac{\alpha}{2}}$
90%	0,1	1,645
95%	0,05	1,96
99%	0,01	2,575

  - **a p becslés hibája:** $E = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$
  - **intervallum becslése így:** $\hat{p} - E < p < \hat{p} + E$

- **Átlagbecslés**
  - ***σ ismeretében (X)***
    - **feltételek**
      - ♦ minden ugyan olyan hosszúságú minta kiválasztásának egyenlő az esélye
      - ♦ a <u>populáció</u> σ szórása ismert
      - ♦ a populáció normális eloszlású és/vagy $n > 30$
    - a minta átlaga $(\bar{x})$ a populáció átlag (μ) legjobb becslése
    - **hiba:** $E = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
    - **intervallum becslése így:** $\bar{x} - E < \mu < \bar{x} + E$
  - ***σ nem ismert (t)***
    - **feltételek**
      - ♦ minta véletlenszerű
      - ♦ normális eloszlásból származik és/vagy $n > 30$
    - **jelölések:** $\boldsymbol{t_{\frac{\alpha}{2}}}$ kritikus érték: $t = \frac{(\bar{x}-\mu)\sqrt{n}}{s}$; **s** a minta szórása
    - **szabadsági fok:** $n - 1$
    - **hiba:** $E = t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$
    - **intervallum becslése így:** $\bar{x} - E < \mu < \bar{x} + E$
- **Varianciabecslés *(X^2)***
  - ***feltételek***
    - a minta legyen egyszerű véletlenszerű
    - a populációnak normális eloszlásúnak kell lennie (a nagy minta nem elégséges)
  - a minta variancia (s^2) a legjobb pontbecslése a populáció varianciájának (σ^2)
  - ***képlet*:** $X^2 = \frac{(n-1)s^2}{\sigma^2}$
  - ***szabadsági fok*:** $df = n - 1$
  - ***tulajdonságai***
    - a szabadsági fokok növekedésével egyre szimmetrikusabb lesz
    - értékei nem negatív számok
  - ***intervallumbecslése így*:** $\sqrt{\frac{(n-1)s^2}{X_R^2}} < \sigma < \sqrt{\frac{(n-1)s^2}{X_L^2}}$

❖ **<u>Hipotézis vizsgálat</u>**

- **Definíciók**
  - ***hipotézis*:** a populáció valamilyen tulajdonságára vonatkozó állítás, vagy kijelentés.
  - ***ritka esemény szabály*:** ha adott feltevések mellett egy bizonyos esemény valószínűsége kicsi, de mi mégis megfigyeljük egy ilyen esemény bekövetkezését, akkor arra a konklúzióra jutunk, hogy a feltevés nem igaz.
  - ***null hipotézis*:** egy állítás a populáció valamilyen paraméter értékéről miszerint az <u>egyelő</u> valamilyen feltételezett értékkel.
  - ***alternatív hipotézis:*** egy állítás, ami szerint a paraméter értéke valamilyen módon különbözik a nulla hipotézistől
  - ***teszt statisztika:*** egy számérték, aminek segítségével döntést hozunk a null hipotézisről
    - **arányra:** $z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$
    - **átlagra:** $z = \frac{(\bar{x}-\mu_{\bar{x}})\sqrt{n}}{\sigma}$

- **varianciára:** $X^2 = \frac{(n-1)s^2}{\sigma^2}$

▪ ***kritikus tartomány:*** (elutasítási tartomány) a teszt statisztika értékeinek az a tartománya ami arra vezet, hogy a null hipotézist elutasítsuk
- elvetjük a H_0-t ha a teszt statisztika a kritikus tartományba esik
- nem tudjuk elvetni a H_0-t ha a teszt statisztika nem a kritikus tartományba esik

▪ ***szignifikancia szint (α):*** az a valószínűség, amivel a test statisztika kritikus tartományba esik amikor a null hipotézis valójában igaz

▪ ***kritikus érték:*** amik elválasztják a tartományt azoktól az értékektől, ahol nem utasítjuk el.
- nagyban függ a null hipotézis fajtájától, szignifikancia szinttől

▪ ***P-érték:*** (valószínűségérték) annak a valószínűsége, hogy a teszt statisztika olyan értéket adjon, ami legalább annyira szélsőséges, mint az az érték amit a mintákból kaptunk, azzal a felvetéssel hogy a null hipotézis igaz.
- $\boldsymbol{a}$ null hipotézist elvetjük, ha a P-érték nagyon kicsi $\leq 0{,}05$
- nem tudjuk elvetni H_0-t ha a P-érték $> \alpha$

▪ ***elsőfajú hiba*** $(P(E_1) = \alpha)$***:*** amikor hibás módon elutasítjuk a null hipotézist amikor az igaz

▪ ***másodfajú hiba (***$P(E_2) = \beta$***):*** amikor nem utasítjuk el a null hipotézist akkor amikor nem igaz.
- minden adott α esetén a minta elemszám n növelése a β csökkenését okozza
- minden fix minta elemszám n esetén α csökkenése β növekedését okozza
- ha α és β együttes csökkenését akarjuk elérni akkor a minta elemszámot kell növelnünk.

▪ ***hipotézis teszt erőssége:*** az $1 - \beta$ valószínűségi érték, ami a helytelen null hipotézis elutasításának valószínűsége. A hipotézis teszt erőssége egy igaz alternatív hipotézis támogatásának valószínűsége.

❖ **Korreláció és regresszió**

➢ **Korreláció:** két változó kötött lép fel, ha az egyik a másikkal valamilyen módon kapcsolatban van

▪ ***lineáris korrelációs együttható*****:** r méri a lineáris kapcsolat erősségét egy x és y párokból álló minta értékei között (ρ a populáció lineáris korrelációs együtthatója)

▪ ***feltételek:***
- az (x, y) párokból álló adatok véletlen független minta adatai legyenek
- vizuálisan (ábrázolva) körülbelül stimmel az egyenes alak
- az outliereket el kell távolítani, amennyiben hibás adatok voltak

▪ ***képlete:*** $r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$ vagy $r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$

▪ ***tulajdonságai***
- $-1 \leq r \leq 1$
- az r értéke nem változik, ha bármelyik változónak megváltoztatjuk a mértékegységét
- az r értékét nem befolyásolja az x és y felcserélése
- nem oksági kapcsolat a korreláció!

▪ ***teszt statisztika:*** közelíthető a hipotézis student t statisztikával $df = n - 2$ esetén $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$

➢ **Regresszió**

❖ **Kombinatorika**

➢ **Permutáció:** n elem lehetséges sorrendjének száma

▪ ***ismétléses:*** $\mathrm{P}_{\mathrm{n}}^{\mathrm{k}_{1\ldots\mathrm{l}}} = \frac{\mathrm{n!}}{\mathrm{k_1!\cdot k_2!\ldots k_l!}}$

▪ ***ismétlés nélküli:*** $\mathrm{P_n} = \mathrm{n!}$

➢ **Kombináció:** n elemből k elemet kiválasztunk, úgy hogy a sorrend nem számít.

- ***ismétléses:*** $C_n^{k,i} = \binom{n+k-1}{k}$
- ***ismétlés nélküli:*** $C_n^k = \frac{n!}{k!(n-k)!} = \binom{n}{k}$

➢ **Variáció:** n elemből kiválasztva k darabot, úgy hogy a sorrend számít.

- ***ismétléses:*** $V_n^{k,i} = n^k$
- ***ismétlés nélküli:*** $V_n^k = \frac{n!}{(n-k)!}$

Ezt a jegyzetet a Statisztikai módszerek a fizikában című elsőéves fizika bsc-s tárgyhoz készítettem. A definíciókat és egyéb képleteket ebből a tárgyból kiadott jegyzetből írtam ki. Ha bármilyen elírás belekerült azért elnézést kérek. Bárminemű megjegyzést a jegyzettel kapcsolatban (amely építő) szívesen fogadok: xantoxlilian[kukac]gmail.com címen.