KIUGRÓ ADATOK KISZŰRÉSE STATISZTIKAI TESZTEKKEL
Dixon Q-tesztje

Gyakori feladat az analitikai kémiában, hogy „kiugrónak tűnő” mérési adatokról kell eldöntenünk, hogy azokat elhagyjuk-e a további számításokból.

Erre vonatkozóan ne feledjük, hogy kisszámú mérési adat esetén (mondjuk néhány tíznél kevesebb) semmiképpen lehetünk benne biztosak, hogy a mért adatok normális (Gauss) eloszlásúak, mert nincs elegendő információnk ennek eldöntésére. Ne feledjük, ebből az is következik, hogy abban sem lehetünk biztosak, hogy az adatok átlaga a legjobb becslés a helyes értékre! (ilyenkor a medián számítása lehet a legjobb megoldás).

Ha azonban biztosak vagyunk a normális eloszlásban, akkor többféle statisztikai teszt közül választhatunk a kiugró pontok kiszűrésére. Ezek közül kettőt alkalmazunk most:

- Dixon-féle Q-teszt
- Grubb-féle teszt

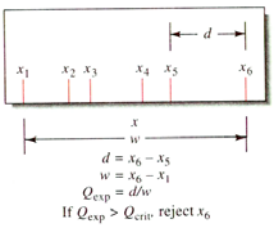
KIUGRÓ ADATOK KISZŰRÉSE STATISZTIKAI TESZTEKKEL
Dixon Q-tesztje

A Dixon-féle Q-teszt során a

$$Q = \frac{\text{köz}}{\text{terjedelem}}$$

mennyiséget kell kiszámolni és egy táblázatból vett Q_{krit} értékkel összehasonlítani; ha $Q > Q_{krit}$, akkor a gyanús érték elhagyható. A képletben a ***terjedelem*** a legkisebb és legnagyobb mért érték közötti távolság, a ***köz*** pedig a „gyanús" adatelem és legközelebbi szomszédja közötti távolság.

Number of Observations	90% Confidence	95% Confidence	99% Confidence
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568



1. FELADAT
Kiugró adat kiszűrése Q-teszttel

Egy kalcit minta százalékos CaO tartalmára a következő adatokat kaptuk:

55.95%, 56.00%, 56.04%, 56.08%, 56.23%

Az utolsó adat „gyanúsnak" tűnik. Kiugró értéknek számít-e 90%-os megbízhatósági szinten és így el kell-e hagynunk az adatok közül?

1. MEGOLDÁS
Kiugró adat kiszűrése Q-teszttel

A megadott adatokra:

55.95%, 56.00%, 56.04%, 56.08%, 56.23%

terjedelem= 56.23% - 55.95% = 0.28%

köz = 56.23% - 56.08% = 0.15%

$$Q = \frac{\text{köz}}{\text{terjedelem}} = \frac{0.15\%}{0.28\%} = 0.54$$

mivel 90%-os megbízhatósági szinten 5 adatra Q_{krit} = 0.64, aminél Q értéke kisebb, ezért ***az adatot bent kell hagynunk az adatsorban***.

2. FELADAT
Kiugró adat kiszűrése Q-teszttel

Fenol meghatározása HPLC módszerrel a következő adatokat szolgáltatta:

0.167, 0.177, 0.181, 0.181, 0.182, 0.183, 0.184, 0.186, 0.187, 0.189

Az első adat „gyanúsnak" tűnik. Kiugró értéknek számít-e 90%-os megbízhatósági szinten és így el kell-e hagynunk az adatok közül?

2. MEGOLDÁS
Kiugró adat kiszűrése Q-teszttel

A megadott adatokra:

$$0.167,\ 0.177,\ 0.181,\ 0.181,\ 0.182,\ 0.183,\ 0.184,\ 0.186,\ 0.187,\ 0.189$$

$$Q = \frac{\text{köz}}{\text{terjedelem}} = \frac{0.177 - 0.167}{0.189 - 0.167} = 0.455$$

mivel 90%-os megbízhatósági szinten 10 adatra Q_{krit} = 0.412, aminél Q értéke nagyobb, ezért ***az adatot elhagyhatjuk az adatsorból***.

KIUGRÓ ADATOK KISZŰRÉSE STATISZTIKAI TESZTEKKEL
Grubb-féle teszt

A Grubb-féle teszt során feltesszük, hogy normális eloszlásúak az adataink és azt vizsgáljuk, hogy a „gyanús" adat előfordulásának valószínűsége kisebb-e, mint a megbízhatósági szint. A következő egyszerű képlet értékét számoljuk ki a gyanús adatra:

$$Z = \frac{|\text{átlag} - x|}{s}$$

ahol az átlagot az összes mért adatra számoljuk, *s* az empirikus szórás, *x* pedig a „gyanús" mérési adat értéke. Ezek után a Z értékét egy táblázatból vett Z_{krit} értékkel kell összehasonlítani; ha Z> Z_{krit}, akkor a gyanús érték elhagyható.

N	90%	92.5%	95%	97.5%	99%
3	1.15	1.15	1.15	1.15	1.15
4	1.42	1.44	1.46	1.48	1.49
5	1.6	1.64	1.67	1.71	1.75
6	1.73	1.77	1.82	1.89	1.94
7	1.83	1.88	1.94	2.02	2.1
8	1.91	1.96	2.03	2.13	2.22
9	1.98	2.04	2.11	2.21	2.32
10	2.03	2.1	2.18	2.29	2.41
11	2.09	2.14	2.23	2.36	2.48
12	2.13	2.2	2.29	2.41	2.55
13	2.17	2.24	2.33	2.46	2.61
14	2.21	2.28	2.37	2.51	2.66
15	2.25	2.32	2.41	2.55	2.71
16	2.28	2.35	2.44	2.59	2.75
17	2.31	2.38	2.47	2.62	2.79
18	2.34	2.41	2.5	2.65	2.82
19	2.36	2.44	2.53	2.68	2.85
20	2.38	2.46	2.56	2.71	2.88

3. FELADAT
Kiugró adat kiszűrése Grubb-féle teszttel

Egy urán izotóp tömegspektroszkópiás meghatározása során a következő nyolc intenzitás adat született:

199.31; 199.53; 200.19; 200.82; 201.92; 201.95; 202.18; 245.57

Az utolsó adat „gyanúsnak" tűnik. Kiugró értéknek számít-e a 99%-os megbízhatósági szinten és így el kell-e hagynunk az adatok közül?

3. MEGOLDÁS
Kiugró adat kiszűrése Grubb-féle teszttel

A megadott adatokból:

199.31; 199.53; 200.19; 200.82; 201.92; 201.95; 202.18; 245.57

átlag= 206.433
empirikus szórás (standard deviáció)= 15.852

$$Z = \frac{|\text{átlag} - x|}{s} = \frac{|206.433 - 245.57|}{15.852} = 2.468$$

A táblázatból, 8 adatra a Z_{krit} értéke 99%-os szinten 2.22.
Ebből az következik, hogy ***igen, az adat elhagyható.***

A MEDIÁN ALKALMAZÁSA KIUGRÓ ADATOK ESETÉN
Kisszámú mérési adat esetén

Ha kisszámú mérési adatunk van, akkor nem lehetünk biztosak abban, hogy az adatok a normális eloszlást követik. Ekkor az átlagnál jobb becslés a valódi értékre a ***medián*** megadása, mivel ez sokkal robusztusabb a kiugró adatok kezelésében.

A medián definíció szerint az az érték egy adathalmazban, amelynél az adatoknak pontosan a fele nagyobb és pontosan a fele kisebb. Ha páratlan tagú adatsorról van szó, akkor a nagyság szerint rendezett adatok közül a középső értéke. Ha páros tagszámú az adatsor, akkor (egy „lepuhított definíció szerint") a két középső érték átlagát tekintjük mediánnak. Példa:

- **Páratlan elemszám esetén**:

  1 2 5 4 3 1 4 3 3 4 3 5 1

  A rendezett sokaság:

  1 1 1 2 3 3 3 3 4 4 4 5 5

  A medián a középső elem:

  1 1 1 2 3 3 **3** 3 4 4 4 5 5

- **Páros elemszám esetén**:

  1 4 2 4 2 3 5 3 1 1

  A rendezett sokaság:

  1 1 1 2 **2 3** 3 4 4 5

  A medián a középső elemek számtani közepe: 2,5.

4. FELADAT ÉS MEGOLDÁS
A medián érték robusztusságának szemléltetése

A következő mérési adatsorunk van:

10.1; 9.9; 9.7; 10.2; 10.5; 3.0; 10.3; 10.0; 10.0

ami nagyság szerint rendezve megfelel:

3.0; 9.7; 9.9; 10.0; 10.0; 10.1; 10.2; 10.3; 10.5

az átlag: 9.3
a szórás: 2.373
a medián: 10.0

ha elhagyjuk a kiugró adatot (3.0), akkor:

az átlag: 10.087
a szórás: 0.247
a medián: 10.05

Gondoljuk végig otthon azt is, hogyan alakulnak a számok extrém esetben, pl. amikor 10 mérési adatból kilenc értéke 1,000, míg egynek az értéke 1,000,000 !

MÉRÉSI ADATOK ÁTLAGÁNAK ÖSSZEVETÉSE A VALÓDI ÉRTÉKKEL
Null hipotézis

Egy másik gyakori, statisztikai jellegű feladat az analitikai kémiában, hogy (pl. egy validálás, QC mérés során) összevessük mérési adatainkat a helyes, valódi értékkel. Ez utóbbit valójában csak éppen ilyenkor ismerjük, mivel a QC mérések során egy referencia mintát mérünk meg, amelyet mi állítottunk elő (vagy kereskedelmi) és összetételét pontosan ismerjük.

Mivel a mérési adatok *pontosan* nagy valószínűséggel nem fognak egyezni a valódi értékkel, ezért statisztikai próbának kell alávetni az adatokat, hogy eldönthessük, a tapasztalt eltérés a mérési adatok átlaga és a valódi érték között még belefér-e a normális eloszlás szórásába, vagy nem. Az utóbbi esetben rendszeres hibával (negatív vagy pozitív) állunk szemben.

Az ellenőrzés az ún. null hipotézissel lehetséges. Feltevésünk az, hogy nincs rendszeres hiba (egy adott megbízhatósági szinten). Ez akkor teljesül, ha a

$$\bar{x} - \mu = \pm \frac{t \cdot s}{\sqrt{n}}$$

ahol μ a valódi érték, s az empirikus szórás, t a Student-féle érték, n pedig a mérési adatok száma.

5. FELADAT
Null hipotézis alkalmazása

Egy kénmeghatározási eljárás tesztelése során a 0.123% kéntartalmú referencia mintára a következő elemzési eredményeket kaptuk:

0.112%; 0.118%; 0.115%; 0.119%

Állapítsuk meg, hogy van-e rendszeres hiba a mérési adatokban, vagy nincs!
A megállapítást tegyük meg 95% és 99% megbízhatósági szintre is!

5. MEGOLDÁS
Null hipotézis alkalmazása

μ= 0.123%
n= 4
s= 0.0032
átlag= 0.116%
átlag - μ = 0.007

95% valószínűségi szinten $t_{95\%}$= 3.18 (mivel a szabadsági fokok száma= 3), így:

$$\frac{t \cdot s}{\sqrt{n}} = \frac{3.18 \cdot 0.0032}{\sqrt{4}} = 0.0051$$

az (átlag - μ) > 0.0051, tehát a null hipotézist el kell vetnünk.
Van rendszeres hiba ezen a valószínűségi szinten!

5. MEGOLDÁS
Null hipotézis alkalmazása

μ= 0.123%
n= 4
s= 0.0032
átlag= 0.116%
átlag - μ = 0.007

95% valószínűségi szinten $t_{95\%}$= 3.18 (mivel a szabadsági fokok száma= 3), így:

$$\frac{t \cdot s}{\sqrt{n}} = \frac{3.18 \cdot 0.0032}{\sqrt{4}} = 0.0051$$

az (átlag - μ) > 0.0051, tehát a null hipotézist el kell vetnünk.
Van rendszeres hiba ezen a valószínűségi szinten!

99% valószínűségi szinten $t_{99\%}$= 5.84, így ekkor

$$\frac{t \cdot s}{\sqrt{n}} = \frac{5.84 \cdot 0.0032}{\sqrt{4}} = 0.0093$$

az (átlag - μ) < 0.0093, tehát a null hipotézis megállja a helyét.
Nincs rendszeres hiba ezen a valószínűségi szinten!