Korrelációszámítás alkalmazása az agrokémiában

A kutatómunka nagy részét képezik az olyan kérdések, amelyekben két vagy több tényező kölcsönhatását kell megvizsgálnunk. Meg ke mondanunk: valóban kapcsolatban állnak-e egymással azok a tényezők, és ha igen, a köztük fennálló összefüggés alakját is meg kell határoznunk.

Vizsgálat alapjául szolgáló megfigyelési eredményeket szemlélve, talán már egyszerű ránézéssel megállapíthatjuk például azt, hogyha az egyik tényezőt megnöveljük, akkor ez a másik tényező növekedését vonja maga után, de ennek pontos mértékét megállapítani nem tudjuk, mert a megfigyelési értékek erős ingadozást mutatnak. Jelen dolgozatban az ilyen természetű ingadozások tulajdonságaival foglalkozunk és a matematikai statisztika segítségével egy nagyon könnyen kezelhető módszert ismertetünk azok vizsgálatához.

Először vizsgáljuk meg azt a kérdést, mi okozza a mérési eredmények ingadozását. Mindenki előtt ismert az a tény, ha két mérést végzünk, ugyanarra a kérdésre vonatkozóan, a legritkább esetben kapunk pontosan egyforma eredményt. A két mérés során nem tudjuk pontosan ugyanazokat a körülményeket biztosítani. A két vizsgált tényező értékét viszonylag pontosan be tudjuk állítani, de az esetek legnagyobb részében még nem is tudjuk, hogy párhuzamosan milyen sokrétű folyamat zajlik le. A mezőgazdasági megfigyelésnél talán a legnagyobb mértékű ez az ingadozás. Hiszen a szántóföldön nincs két pontosan egyforma parcella, az időjárási tényezők különbözhetnek, a kártevők hatása nem egyforma, a megművelés, a learatás stb., stb. mind olyan hibaforrás, amelyet ugyan bizonyos határig csökkenteni lehet, és kell is, de teljesen megszüntetni nem lehet. Talán kisebb mértékben, de ugyanez áll a laboratóriumban végzett kísérletekre is.

Számos esetben végeztek már kísérleteket, amelyeket a korreláció számítás segítségével értékeltek ki. Pl. Botvay homokos talajokat vizsgálva megfigyelte a végleges és differenciális kapillárás vízemelési adatokat és a megfelelő leiszapolható alkatrészek adatainak összefüggését [1].

Keresztény cikkében a műtrágya adagok hatását vizsgálta a termés nagyságára vonatkozóan [5]. Jelen dolgozatban az említett cikkekben már ismertetett eljárásokat foglaljuk össze és egészítjük ki.

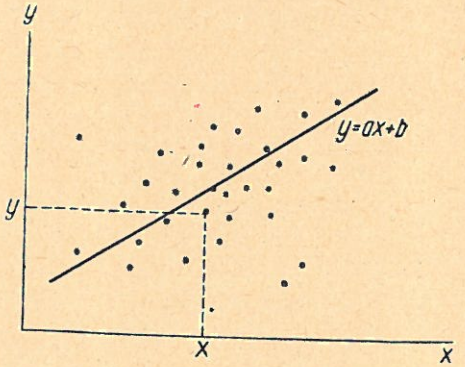
A már említett hibákat úgy küszöbölhetjük ki, hogy több mérést végzünk. Ebben az esetben feltételezhetjük, hogy az általunk nem ismert véletlen hatások kiegyenlítődnek. Feltételezhetjük ugyanis, hogy a véletlen hatások nem mindig ugyanabban az irányban érvényesülnek és így ha a megfigyelések átlagát számítjuk, akkor a vizsgált tényező hatására jó képet kaphatunk, mert az általunk beállított hatás érvényesült mindig ugyanabban az irányban.

Lehetséges egyidejűleg több tényező hatásának vizsgálata egy tényezőre vonatkozóan és a kapcsolat leírására a legkülönbözőbb alakú függvényeket feltételezhetjük, de most csak a legegyszerűbb esetet fogjuk vizsgálni: két tényező kapcsolatát, amikor köztük egyenesvonalú (lineáris) összefüggést feltételezhetünk.

Az egyszerűség kedvéért a továbbiakban azt a tényezőt, amelynek hatását vizsgáljuk „független változó"-nak (x), a másik tényezőt pedig, amelyet az előzőtől függőnek tekintünk „függő változó"-nak (y) fogjuk nevezni. A mérési eredményeket pedig mint pontokat fogjuk kezelni (x_i, y_i) koordinátákkal.

Számítások alapjai

Első lépésként megkeressük azt az egyenest, amelyik legjobban illeszkedik a mérési adatokat jellemző pontokra (1. ábra).



1. ábra

A mérési adatokra „legjobban illeszkedő" egyenes

Ez más szóval annyit jelent, hogy az egyenes általános egyenletében szereplő két konstans számot kiszámítjuk.

Az egyenes általános egyenlete

$$y = a\,x + b \qquad (1)$$

Az n darab mérési adatot jellemző (x_i, y_i) számpárokat egymásután behelyettesítjük (1) egyenletbe, s így n egyenletből álló rendszert kapunk, ahol az a és b számokat tekintjük ismeretleneknek.

$$\begin{aligned} y_1 &= a\,x_1 + b \\ y_2 &= a\,x_2 + b \\ &\vdots \\ y_n &= a\,x_n + b \end{aligned} \qquad (2)$$

Mivel n a bevezetésben elmondottak miatt jóval nagyobb minden esetben, mint kettő, ezért (2) lineáris egyenletrendszer általában nem oldható meg egyértelműen. Ez geometriailag annyit jelent, hogy kettőnél több ponton át általában nem húzható egy egyenes, ami világosan látható az első ábrából is.

Kézenfekvő azt az egyenest választani, amely talán nem megy át egy ponton sem, de eltérése a mérési adatokat jellemző pontoktól a legkisebb. Ezt egyszerűen megkapjuk, ha megkeressük ezt az a és b számot, amelynél a

$$\sum_{i=1}^{n} (y_i - a\,x_i - b)^2 = \text{minimum} \qquad (3)$$

Ebből egyszerű matematikai számítással, amelynek részletezése nem célja a dolgozatnak, kapjuk a és b értékét megadó képleteket

$$a = \frac{n\sum_{i=1}^{n} x_i\,y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (4)$$

$$b = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i\,y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (5)$$

Az így meghatározott egyenest az irodalomban „regressziós egyenes"-nek nevezik.

A regressziós egyenes iránytangense a határozza meg a legfontosabb tulajdonságát: a meredekségét és irányát. Ha pozitív előjelű, akkor az x növekedésének az y növekedése felel meg, ha negatív, akkor az x növekedésének az y csökkenése felel meg. Ha pedig zérus, akkor az y független az x értékeitől. A gyakorlatban a legritkább esetben kapunk az a értékére zérust, de az lehetséges, hogy nagyon kicsi lesz. A bevezetésben elmondottak miatt az alapadatok hibával terheltek, ezért az a értékének kiszámításánál is hibát követünk el. Meg kell vizsgálnunk azt, hogy a kapott a érték megbízható-e vagy a zérustól való eltérését a véletlen hiba okozta-e.

Először kiszámítjuk a hibáját:

$$s_a = \sqrt{\frac{\sum_{i=1}^{n} (y_i - Y_i)^2}{(n-2)\sum_{i=1}^{n} (x_i - \bar{x})^2}} \qquad (6)$$

ahol x_i a független változó megfigyelési értékei,
y_i a függő változó megfigyelési értékei,
Y_i a függő változónak a regressziós egyenes alapján számított értéke,
$\bar{x}$ az x_i értékek átlaga,
n a megfigyelések száma.
a megbízhatóságát pedig a következő hányados alapján bíráljuk el

$$t_1 = \frac{|a|}{s_a} \qquad (7)$$

ahol t_1 a statisztikai számításoknál sokat használt Student-féle „t" eloszlást követ, $n-2$ szabadságfokkal. Ha $t_1 > t_1(p)$, akkor a értékét megbízhatónak mondjuk p valószínűségi szinten, ha $t_1 \leqq t_1(p)$, akkor nem tudjuk elbírálni, hogy a zérustól való eltérését a véletlen ingadozás okozta-e, vagy pedig az esetleg meglevő kapcsolat [2, 4, 6].

(A „p valószínűségi szinten megbízható, vagy szignifikáns" annyit jelent, hogy p a valószínűsége annak, hogy rosszul döntünk, és $1-p$ valószínűsége annak, hogy döntésünk helyes. A $t_1(p)$ pedig a Student eloszlás p-hez tartozó kritikus értéke. A gyakorlatban általában a $p = 0{,}05$, vagy $0{,}01$ valószínűségi szintet szokták felvenni.)

Ha azt az eredményt kapjuk, hogy a nem tér el szignifikánsan zérustól, akkor nincs értelme további vizsgálatokat végezni, mert nem bizonyult helyesnek a két változó kapcsolatára tett feltevés. Viszont, ha szignifikáns eredményt kaptunk, vizsgálatainkat tovább folytathatjuk.

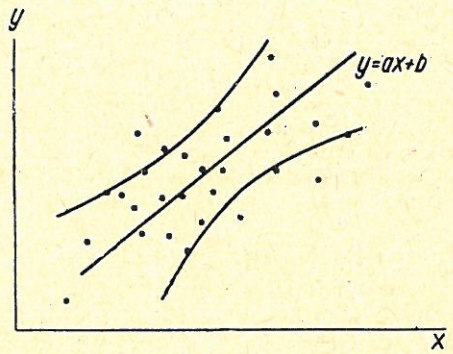
Ki kell számítanunk azt a hibát, amit akkor követünk el, ha egy későbbi kísérlet során a most meghatározott regressziós egyenes alapján akarjuk kiszámítani egy bizonyos x_i értékhez tartozó Y_i értéket. Y_i hibáját a következő képlet adja meg:

$$s_{y_i} = s\sqrt{1 + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \qquad (8)$$

ahol

$$s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - Y_i)^2}{n-2}} \qquad (9)$$

Megjegyezzük, hogy a számított értékek hibája függ attól a helytől, ahol az Y_i-t számítjuk. A (8)-ból látható, hogy a legkisebb hibát az $x_i = \bar{x}$ helyen kapjuk, mert ekkor $s_{yi} = s$, minden ettől különböző helyen már az $s_{yi} > s$ (2. ábra). Természetesen az így megadott hiba nem azt jelenti, hogy minden eredmény azon belül lesz, hanem csak azt, hogy az eredmények túlnyomó része.



2. ábra

A regressziós egyenes segítségével becsült értékek hibája

Hátra van még egy lépés : a korrelációs együttható kiszámítása. Az elmondottak alapján meghatároztuk a regressziós egyenest, de nem tudjuk megmondani, hogy a pontok milyen jól illeszkednek rá. Ugyan az s_{yi} kiszámítása bizonyos információt ad az illeszkedés pontosságára (minél kisebb, annál jobb), de ez nehezen kezelhető, mert függ az x változó értékétől. Ezért az összefüggés szorosságára a következő szám ad választ,

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(x_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \qquad (10)$$

amelyet korrelációs együtthatónak nevezünk. Értéke —1-től +1-ig változhat. ±1 az értéke, ha *pontosan* lineáris összefüggés áll fenn a két változó között. Másrészt viszont, minél kisebb, az abszolút értéke a kapcsolat annál kevesbé lineáris. Ha azt kapjuk, hogy $r = 0$, akkor semmilyen összefüggést nem szabad feltételeznünk a két változó között. Ha pozitív előjelű, akkor x növekedésének y növekedése felel meg, ha negatív, akkor x növekedésének y csökkenése felel meg. Figyelembe véve az előzőkben mondottakat, r előjelének meg kell egyeznie a előjelével, ami regressziós egyenes irányát határozza meg.

Ha r értéke közel van zérushoz, ugyanúgy, mint az a kiszámításánál tettük, meg kell vizsgálnunk, megbízhatónak tekinthető-e a kapott eredmény, vagy pedig a véletlen ingadozásának tulajdonítható-e. Ennek elbírálása a következő képlet alapján történik :

$$t_2 = \frac{|r|}{\sqrt{1 - r^2}} \sqrt{n - 2} \qquad (11)$$

ahol t_2 $n - 2$ szabadságfokú Student-féle „t" eloszlást követ. Ha $t_2 > t_2(p)$, akkor r eltér zérustól, ha $t_2 \leqq t_2(p)$, akkor r nem tér el zérustól p valószínűségi szinten.

Sok esetben a vizsgálatot a korrelációs együttható kiszámításával szokták kezdeni, mégpedig azért, mert csak egy bizonyos számnál nagyobb korreláció esetén tartják érdemesnek a regressziós egyenes kiszámítását. Egyes eseteknél már ± 0,5 körüli r értéket jelentős eredménynek tekinthetünk, de általában 0,6—0,8 abszolút értékű korrelációs együtthatót tekinthetünk csak jelentősnek. Ezt mindig az adott esetben kell elbírálni, a vizsgálat metódusától függően.

Mindezeket a vizsgálatokat hasonlóan elvégezhetjük abban az esetben is, ha több változó kapcsolatát kell megfigyelnünk. Ebben az esetben „parciális korrelációs együttható"-ról beszélünk és „regressziós sík"-ról. Ha pedig a két változót kell csak vizsgálni, de köztük nem lineáris kapcsolat áll fenn, akkor a „korrelációs hányados"-t kell kiszámítanunk az összefüggés alakjára tett feltétel mellett. Ez utóbbi eljárás még egyszerűsíthető úgy, hogy ha például logaritmikus összefüggést tételezünk fel, vagyis

$$z = a \log u + b$$

akkor először kiszámítjuk a $v = \log u$ számokat és a $z = av + b$ lineáris összefüggést kapjuk, amit az itt leírt módon értékelhetünk.

Felmerülhet olyan probléma is, hogy ugyanazokra a tényezőkre két egymástól független kísérletsorozatot végeztünk, s így kapott korrelációs együtthatókat átlagolni szeretnénk. Nyomatékosan hangsúlyozni akarjuk, hogy korrelációs együtthatókkal közvetlenül semmi további műveletet végezni nem szabad. R. A. Fischer [3] által bevezetett z transzformáció segítségével lehet csak két korrelációs együtthatót összevonni, de csak akkor, ha ugyanarra a kérdésre vonatkoznak. Ezekkel a kérdésekkel jelen dolgozatban bővebben foglalkozni nem célunk. A megadott irodalomban ezekre a kérdésekre is kimerítő választ kaphatunk. Ha pedig különböző vizsgálatok eredményeiből akarunk egy közös korrelációs együtthatót kiszámítani, akkor minden esetben külön meg kell vizsgálnunk, miként lehetséges azt

kiszámítani, sőt az is kérdéses, hogy egyáltalán megoldható-e a probléma.

Mielőtt az ismertetett módszer egy konkrét alkalmazására térnénk át, a szükséges képleteket a számítás egyszerűsítése céljából más alakban írjuk fel.

Jelöljük :

$$\sum_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} x\right)^2}{n}$$

$$\sum_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$\sum_{xy} = \sum_{i=1}^{n} (x_i - x)\,(y_i - y) - \left|\sum_{i=1}^{n} x^1\right| \left(\sum_{i=1}^{n} x_i\, y_i\right)$$

Ezek segítségével fogjuk kifejezni a többi képletet. (4) és (5) a következő alakú lesz :

$$a = \frac{\sum xy}{\sum xx} \qquad (4^*)$$

$$b = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i\, y_i\right)}{n \cdot \sum xx} \qquad (5^*)$$

(6) egyszerűsítésére felhasználjuk a következő egyenlőséget

$$\sum_{i=1}^{n} (y_i - Y_i)^2 = (1 - r^2) \sum_{yy} \qquad (6^*)$$

akkor

$$s_a = \sqrt{\frac{(1 - r^2) \sum_{yy}}{(n - 2) \sum_{xx}}}$$

ahol r a korrelációs együttható.
(8)-ból kapjuk

$$s_{Y_i} = s \sqrt{1 + \frac{x_i - \bar{x})^2}{\sum_{xx}}} \qquad (8^*)$$

ahol

$$s = \sqrt{\frac{(1 - r)^2 \sum_{yy}}{n - 2}} \qquad (9^*)$$

Végül (10) a következő alakban írható

$$r = \frac{\sum_{xy}}{\sqrt{\sum_{xx} \sum_{yy}}} \qquad (10^*)$$

Példa.

Sarkadi János számos vizsgálatot végzett a talaj szervesanyaga és termőképessége közötti kapcsolat felderítésére. Szíves engedélyével ezek közül egy méréssorozatot részletesen fogok ismertetni, amelyben Westsik tavaszi vetésforgó kísérletében a talaj összhumusztartalma és a burgonyatermés nagysága közötti összefüggés jellemzésére szolgáltatott adatokat. (Az összes humusztartalom mg %; a burgonyatermés pedig q/kh értendő.)

A megfigyelések eredményeit az 1. táblázat ismerteti.

1. táblázat

A megfigyelés eredményei

Humusz mg%	Termés q/kh
523	27,3
594	62,8
517	52,0
593	49,6
696	61,5
845	71,8
617	51,0
780	79,6
634	53,9
671	53,6
732	63,1
724	71,9
739	78,9
732	72,2
682	63,0

A megfelelő adatpárokat ábrázolva kapjuk a 3. ábrát.

A független változó (x) a humusztartalom lesz, a függő változó (y) pedig a burgonya terméseredménye. A szükséges számítást elvégezve kapjuk

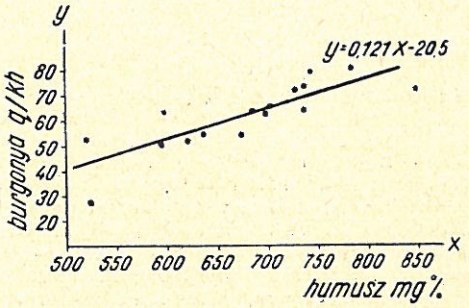
$$\sum_{xx} = 119\,683{,}0$$
$$\sum_{xy} = 14\,476{,}8$$
$$\sum_{yy} = 2\,590{,}5$$

Először kiszámítjuk a korrelációs együtthatót (10*) alapján

$$r = 0{,}82$$

Ez az r érték már olyan jelentős összefüggést

jelent, amely zérustól minden bizonnyal eltér, ezért a gyakorlatban ilyen esetben a következő lépést nyugodtan elhagyhatjuk. Itt azonban a



3. ábra

A talaj humusztartalma és a burgonyatermés közti összefüggés

teljesség kedvéért kiszámítjuk a megbízhatóságát elbíráló t_2 értéket.

Kapjuk

$$t_2 = 5{,}18$$

amely 13 szabadságfok mellett még a $t\,(0{,}001) = 4{,}221$ kritikus értéknél is jóval nagyobb, tehát megállapíthatjuk, hogy igen erősen szignifikáns. Ezután (4*) és (5*) felhasználásával kiszámítjuk a regressziós egyenes együtthatóit.

$$a = 0{,}121$$
$$b = -20{,}46$$

Az a megbízhatóságát is el kell bírálnunk. (6*) alapján

$$s_a = 0{,}0234$$

majd pedig (7) alapján

$$t_1 = 5{,}17$$

amely szintén jelentősen nagyobb a t_1 (0,001) kritikus értéknél. Ennek alapján a kiszámított regressziós egyenest megbízhatónak tekinthetjük és az egyenlete a következő lesz

$$y = 0{,}121\,x - 20{,}46 \qquad (12)$$

(lásd 3. ábra). Ezzel tehát meghatároztuk azt a függvényt, amellyel nagyon jól jellemezhetjük, hogyan függ az adott esetben a burgonyatermés nagysága a talaj humusztartalmától.

Hátra van még annak a hibának a kiszámítása amit akkor követünk el, ha (12) egyenletet a különböző x értékekhez tartozó Y értékek kiszámításához, illetve becsléséhez akarjuk felhasználni. (Az előzőkben megjegyeztük, hogy ha a regressziós egyenes segítségével számítani akarjuk a függő változó értékeit, akkor Y-al jelöljük.)

Számítsuk ki például azt, hogy a vizsgált talajtípushoz hasonló 700 mg%-os humusztartalmú parcellán mekkora burgonyatermésre számíthatunk.

Behelyettesítve (12)-be:

$Y\,(700) = 0{,}121 \cdot 700 - 20{,}46 = 64{,}24 \pm s_{y(700)}$

$Y\,(700)$ hibáját pedig kiszámíthatjuk (8*) alapján.

$$s = 8{,}11$$

Általában Y_i hibáját a következő képlet adja meg:

$$s_{Yi} = 8{,}11 \sqrt{1 + \frac{x_i - 672}{119\,683}}$$

x_i helyére behelyettesítve a példánkban szereplő 700-at, megkapjuk az Y_{700} hibáját:

$$s_{y\,(700)} = 8{,}11$$

mert a négyzetgyök alatt levő szám csak a harmadik tizedesjegyben tér el 1-től, s ezért az általunk számított pontosságon belül nem okoz eltérést. (Természetesen ez egy másik x helyen már nem biztos, hogy ismét elhanyagolható lesz!)

Összefoglalás

Talajtani vizsgálatoknál gyakran előforduló feladatot vizsgáltunk: két változó kapcsolatát olyan esetben, amikor nem tudjuk elválasztani sok különböző zavaró hatástól. A matematikai statisztika segítségével a mérések, eredmények ingadozásából vontunk le következtetéseket és elbíráltuk, vajon a feltételezett kapcsolat milyen jó képet ad a vizsgált összefüggésről. Most csak a legegyszerűbb esettel foglalkoztunk, amikor lineáris kapcsolatot feltételezhettünk. Kiszámítottuk a „regressziós egyenes"-t és a „korrelációs együttható"-t, amely a regresszió egyenes illeszkedésének, vagy pontosságának a mértéke. Végül az ismertetett módszert egy konkrét méréssorozat esetében alkalmaztuk.

MARTON ÁDÁM

Érkezett: 1958. március 28.

Irodalom

[1] *Bolvai, K.:* Adatok az alföldi homoktalajaink kapillárás vízemelőképességének értékeléséhez. Agrokémia és Talajtan 4. 119—132. 1955.

[2] *Cramer, H.:* Mathematical Methods of Statistics. Univ. Press. Princeton. 1951.

[3] *Fischer, R. A.:* Statistical Methods for Research Workers. VIII. Ed. Oliver & Boyd. Edinburgh. 1941.

[4] *Kendall, M. G.:* The Advanced Theory of Statistics. II. Ed. Charles Griffin & Co. London. 1948.

[5] *Keresztény, B.:* Törvényszerűségek a műtrágyázásokban. Agrokémia és Talajtan. 4. 365—384. 1955.

[6] *Rényi, A.:* Valószínűségszámítás. Tankönyv kiadó. Budapest. 1954.