EGER 1774
ESZTERHÁZY KÁROLY KATOLIKUS EGYETEM
Matematikai és Informatikai Intézet

Válogatott fejezetek valószínűségszámításból és statisztikából

A valószínűségszámítás törvényeinek modellezése és statisztikai számítások Excel segítségével

Összeállította
Dr. Tómács Tibor
egyetemi docens

Utolsó módosítás
2024. március 2.

Eger, 2024

Tartalomjegyzék

Bevezetés

Ez az összeállítás az Eszterházy Károly Katolikus Egyetem „*Válogatott fejezetek valószínűségszámításból és statisztikából*” című kurzusához készült a matematika szaktanári képzésen belül. Ahogy a tárgy címéből is kiderül két fejezetből áll a tananyag.

A valószínűségszámításon belül a klasszikus határérték tételekre helyezzük a hangsúlyt. Az elmélet tárgyalásán túl fő célunk ezeknek a tételeknek a számítógépes szimulációja.

A matematikai statisztikán belül a szórásanalízissel és a regressziószámítással foglalkozunk. A számításokat itt is számítógépen végezzük.

A feladatok megoldásait magyar nyelvű Excelen mutatjuk meg. Excel 2010 vagy ennél újabb verzióval dolgozzon, amely a gépre van telepítve, azaz nem online verzió. A jegyzetben használt jelölések megegyeznek az irodalomban megadott jegyzetek jelöléseivel. A jegyzet szabadon letölthető a következő címről:

https://tomacstibor.uni-eszterhazy.hu/tananyagok/Valoszinusegszamitas_szaktanar.pdf

1. fejezet

Valószínűségszámítás

1.1. A nagy számok Bernoulli-féle törvénye

1.1. feladat. Modellezzen 10 000 dobást egy szabályos dobókockával. Számolja ki minden dobás után a hatos dobás relatív gyakoriságát (azaz a hatos dobások és az összes dobás számának arányát), majd vizsgálja annak viselkedését.

Megoldás. Nyisson meg egy üres Excel lapot.

1. Jelölje ki az `A1:A10000` cellatartományt. Ehhez a Név mezőbe írja be, hogy `A1:A10000`, majd Enter.
2. Gépelje be a következőt: `=VÉLETLEN.KÖZÖTT(1;6)`, majd Ctrl + Enter. Ezzel az `A` oszlop első 10 000 sorába olyan pszeudo-véletlen számokat generál, melyek 1-től 6-ig minden egész számot azonos valószínűséggel eredményezhetnek.
3. A `B1` cellába gépelje be a következőt: `=DARABTELI(A$1:A1;6)/SOR(A1)`, majd kattintson kétszer a kitöltőjelre. (A kitöltőjel a kijelölt cella vagy cellatartomány jobb alsó sarkában található pici négyzet. A `DARABTELI` függvény 2023. márciusa utáni Excel-verziókban `DARABHA` néven érhető el!) Ezzel minden dobás után megadtuk az addigi dobásokban a hatosok relatív gyakoriságát.
4. A relatív gyakoriság vizsgálatához jelölje ki a `B` oszlopot, Beszúrás › Diagramok › Pont- (xy) vagy buborékdiagram beszúrása › Pont vonalakkal. A vízszintes tengely minimumát állítsa 0-ra, maximumát 10000-re a fő léptéket pedig 0,16667-re ($\frac{1}{6}$ kerekítése 5 tizedesjegyre). A függőleges tengely minimumát állítsa 0-ra, maximumát pedig 1-re. Ez megrajzolja a dobássorozathoz tartozó relatív gyakoriságot a dobások számának függvényében. Az F9 többszöri megnyomásával több dobássorozat esetén vizsgálhatja a viselkedést.

Tapasztalat. A dobások számának növelésével a hatos dobások relatív gyakorisága egyre kisebb mértékben ingadozik $\frac{1}{6}$ körül.

1.2. feladat. Más véletlen kísérletben is hasonló a tapasztalat? Például modellezzen 10 000 dobást egy szabályos pénzérmével. Számolja ki minden dobás után a fej dobás relatív gyakoriságát, majd vizsgálja annak viselkedését.

Megoldás. Az előző Excel lapon dolgozzon.

1. Javítsa ki az `A1` cella tartalmát a következőre: `=VÉLETLEN.KÖZÖTT(0;1)`, majd Enter. Ez azonos valószínűséggel generálhat 0-t és 1-et is. A 0 jelentse azt, hogy írást, az 1 pedig azt, hogy fejet dobunk. Ezt a fej dobás *indikátorának* is szoktuk nevezni.
2. Az `A1` cella kitöltő jelére kattintson kétszer. Ezzel az előző feladatban generált 10 000 kísérlet minden tagját megváltoztatja a fej dobás indikátorára.
3. A tapasztalat megfogalmazása előtt vegye észre, hogy most a relatív gyakoriság másképp is számolható mint az előbb. Ugyanis a 0 és 1 számokból álló sorozatban az 1 szám relatív gyakorisága pont ezen számok számtani közepe. Így a `B1` cellába egyszerűbb képlet is írható: `=ÁTLAG(A$1:A1)`. Ezt mindenhol kijavítjuk, melyhez a `B1` cella kitöltő jelére kattintson kétszer.
4. A grafikonon a fő léptéket javítsa ki 0,5-re.
5. Az F9 többszöri megnyomásával több dobássorozat esetén vizsgálhatja a viselkedést.

Tapasztalat. A dobások számának növelésével a fej dobások relatív gyakorisága egyre kisebb mértékben ingadozik $\frac{1}{2}$ körül.

Általánosságban az a tapasztalat, hogy *egy véletlen kimenetelű kísérletben egy adott esemény relatív gyakorisága a kísérletek számának növelésével egyre kisebb mértékben ingadozik egy adott érték körül. Ezt az értéket nevezzük az adott esemény valószínűségének.*

Erre a tapasztalatra alapozva a matematikában a valószínűség axiómáit a relatív gyakoriság legegyszerűbb tulajdonságai alapján határoztuk meg. Az így megalkotott elmélet akkor jó, ha az előbbi tapasztalatot igazolja. Erről szól a következő tétel.

Tétel (A nagy számok Bernoulli-féle törvénye)**.** *Legyen k_n egy p valószínűségű esemény bekövetkezéseinek a száma (gyakorisága) n kísérlet után. Ekkor minden $\varepsilon > 0$ esetén*

$$\lim_{n\to\infty} \mathrm{P}\left(\left|\frac{k_n}{n} - p\right| \geq \varepsilon\right) = 0.$$

Azaz akárhogyan is választunk egy pozitív ε számot, a kísérletek számának növelésével egyre kisebb eséllyel tér el az esemény relatív gyakorisága a valószínűségétől ε-nál nagyobb mértékben. Ez pontosan a gyakorlati tapasztalatot fejezi ki.

1.2. A nagy számok gyenge illetve Kolmogorov-féle erős törvénye

1.3. feladat. Az előző feladat megoldásában láttuk, hogy a fej dobás relatív gyakorisága számolható számtani középpel, amennyiben annak indikátorát figyeljük. Hogyan lehetne ezt megoldani a dobókocka esetén a hatos dobás relatív gyakoriságával?

Megoldás. Ehhez annyit kell tenni, hogy hatos dobásakor az 1 számot írja le, ellenkező esetben pedig a 0-t. Ezt nevezzük a hatos dobás indikátorának. Az előző Excel lapon ezt könnyen kipróbálhatja.

1. Javítsa ki az `A1` cella tartalmát erre: `=HA(VÉLETLEN.KÖZÖTT(1;6)=6;1;0)`, majd Enter.
2. Az `A1` cella kitöltő jelére kattintson kétszer. Ezzel az előző feladatban generált 10 000 kísérlet minden tagját megváltoztatja a hatos dobás indikátorára.
3. Ekkor a `B` oszlopban található `ÁTLAG` függvények már a hatos dobás relatív gyakoriságait adják. Így a grafikonon ugyanazt fogja tapasztalni, mint az 1.1. feladatban.

1.4. feladat. Csak egy esemény indikátorának számtani közepe mutat ilyen konvergens tulajdonságot, vagy más valószínűségi változónál is tapasztalható hasonló?

Megoldás. Dolgozzon az előző Excel lapon.

1. Javítsa ki az `A1` cella tartalmát erre: `=VÉLETLEN.KÖZÖTT(1;6)`, majd Enter.
2. Az `A1` cella kitöltő jelére kattintson kétszer. Ezzel az előző feladatban generált 10 000 kísérlet minden tagját megváltoztatja a dobókocka dobásának eredményére, azaz ezek már nem indikátorok. Ekkor a `B` oszlopban található `ÁTLAG` függvények már nem a hatos dobás relatív gyakoriságait adják meg, hanem az addigi dobások eredményeinek számtani közepét.
3. A függőleges tengely maximumát állítsa 5-re a fő léptéket pedig 3,5-re.
4. Az F9 többszöri megnyomásával több dobássorozat esetén vizsgálhatja a viselkedést.

Tapasztalat. A dobások számának növelésével a dobások értékeinek számtani közepe egyre kisebb mértékben ingadozik 3,5 körül. Ezt az értéket a dobás *várható értékének* nevezzük.

Ha egy valószínűségi változóra vonatkozó mérési eredmények számtani közepe a valószínűségi változó várható értéke körül ingadozik, akkor speciálisan egy esemény indikátora esetén miért az esemény valószínűsége körül figyelhető meg egyre kisebb

mértékű ingadozás? Azért mert egy esemény indikátorának várható értéke megegyezik az esemény valószínűségével.

1.5. feladat. Ejtsünk le egy ceruzát az asztalra. A véletlen szám (valószínűségi változó) most legyen a ceruza és az asztallap egyik éle által bezárt szög tangense. Ezt a kísérletet modellezze 10 000-szer. Ekkor is hasonló a tapasztalat, mint az előző feladatban?

Megoldás. Dolgozzon az előző Excel lapon.

1. Javítsa ki az `A1` cella tartalmát erre: `=TAN(VÉL()*PI()/2)`, majd Enter. A `VÉL()` egy pszeudo-véletlen számot generál a $[0, 1]$ intervallumon egyenletes eloszlás szerint (azaz a generált szám a $[0, 1]$ intervallum bármely h hosszúságú részintervallumába h valószínűséggel eshet). Így a `VÉL()*PI()/2` a ceruza és az asztal éle által bezárt szöget modellezi (radiánban). A feladat értelmében ennek vettük a tangensét.
2. Az `A1` cella kitöltő jelére kattintson kétszer. Ezzel az előző feladatban generált 10 000 kísérlet minden tagját megváltoztatja a mostani feladatban leírtakra.
3. A függőleges tengely maximumát és a a fő léptéket is állítsa 100-ra.
4. Az F9 többszöri megnyomásával több dobássorozat esetén vizsgálhatja a viselkedést.

Tapasztalat. A dobások számának növelése ellenére is a legváratlanabb helyeken történhet a számtani közép értékében hatalmas változás, így az előbb tapasztalt konvergens viselkedés itt nem teljesül.

Ennek az az oka, hogy kicsi valószínűséggel ugyan, de előfordulhat, hogy a bezárt szög nagyon közel lesz a derékszöghöz, amelynek tangense nagyon nagy szám, így az addigi átlag értékét jelentősen megváltoztatja. Ez azt eredményezi, hogy ennek a valószínűségi változónak nem véges a várható értéke. Ilyen esetben nincs konvergens viselkedése a számtani középnek.

Tehát az a megfigyelés, hogy *egy valószínűségi változóra vonatkozó mérési eredmények számtani közepe a valószínűségi változó várható értéke körül ingadozik egyre kisebb mértékben a kísérletek számának növelésével*, csak bizonyos feltételek esetén teljesül.

Az erre vonatkozó feltételeket és állítást a következő tétel mondja ki:

Tétel (A nagy számok gyenge törvénye)**.** *Legyenek $\xi_1, \xi_2, \dots$ páronként független, azonos eloszlású, véges szórású valószínűségi változók. Ekkor minden $\varepsilon > 0$ esetén*

$$\lim_{n\to\infty} \mathrm{P}\left(\left|\frac{\xi_1 + \dots + \xi_n}{n} - \mathrm{E}\,\xi_1\right| \geq \varepsilon\right) = 0.$$

Azaz ilyen feltételekkel akárhogyan is választunk egy pozitív ε számot, a kísérletek számának növelésével egyre kisebb eséllyel tér el a számtani közép a várható értéktől ε-nál nagyobb mértékben.

Ugyanezt a tapasztalatot erősebb állítással fogalmazza meg a következő tétel:

Tétel (A nagy számok Kolmogorov-féle erős törvénye)**.** *Legyenek $\xi_1, \xi_2, \ldots$ független, azonos eloszlású, véges várható értékű valószínűségi változók. Ekkor*

$$\mathrm{P}\left(\lim_{n\to\infty} \frac{\xi_1 + \cdots + \xi_n}{n} = \mathrm{E}\,\xi_1\right) = 1.$$

Azaz ilyen feltételekkel a számtani közép 1 valószínűséggel (másképpen fogalmazva *majdnem biztosan*) a várható értékhez konvergál.

Ha a nagy számok Kolmogorov-féle erős törvényében a „függetlenség” feltételt kicseréljük a gyengébb „páronkénti függetlenség” feltételre, akkor a tétel így is igaz marad. Ez az ún. nagy számok Etemadi-féle erős törvénye.

1.3. Pénzérme dobásakor a fej és írás dobások számainak kiegyenlítődése

1.6. feladat. A szabályos pénzérme feldobásakor láttuk, hogy a fej dobásának a relatív gyakorisága $\frac{1}{2}$-hez konvergál majdnem biztosan a nagy számok Kolmogorov-féle erős törvénye alapján. Ez jelentheti-e azt, hogy a dobások számának növelésével a fej és írás dobások számai egyre közelebb kerülnek egymáshoz, azaz hosszútávon kiegyenlítődnek? Modellezze a problémát 10 000 dobás esetén.

Megoldás. Nyisson egy új Excel lapot.

1. Jelölje ki az `A1:A10000` cellatartományt. Ehhez a Név mezőbe írja be, hogy `A1:A10000`, majd Enter.
2. Gépelje be a következőt: `=VÉLETLEN.KÖZÖTT(0;1)`, majd Ctrl + Enter. Ezek a fej dobások indikátorai.
3. A `B1` cellába gépelje be, hogy `=ABS(DARABTELI(A$1:A1;1)-DARABTELI(A$1:A1;0))`, majd kattintson kétszer a kitöltőjelre. (A `DARABTELI` függvény 2023. márciusa utáni Excel-verziókban `DARABHA` néven érhető el!) Ezzel minden dobás után megadja az addigi dobásokban a fej és írás dobások számainak abszolút eltérését.
4. Az abszolút eltérés vizsgálatához jelölje ki a `B` oszlopot, Beszúrás › Diagramok › Pont- (xy) vagy buborékdiagram beszúrása › Pont vonalakkal. A vízszintes tengely minimumát állítsa 0-ra, maximumát 10000-re. A függőleges tengely minimumát állítsa 0-ra, maximumát pedig 200-ra. Ez megrajzolja a dobássorozathoz tartozó fej

és írások számának abszolút eltérését a dobások számának függvényében. Az F9 többszöri megnyomásával több dobássorozat esetén vizsgálhatja a viselkedést.

Tapasztalat. A függvény nem mutat semmilyen konvergens viselkedést. Ugyanakkor, ha a fej és írás dobások számai egyre közelebb kerülnének egymáshoz, azaz hosszútávon kiegyenlítődnének, akkor az előző függvénynek 0-hoz kellene konvergálnia. Tehát a tapasztalat nem igazolja a fej és írás dobások kiegyenlítődését. Sőt sok esetben az tapasztalható, hogy a dobások számának növelésével egyre nagyobb az eltérés.

Ezt a tapasztalatot a következő tétel magyarázza:

Tétel. *Jelölje F_n egy szabályos pénzérme n-szer történő feldobása után a fej dobások számát, míg I_n az írás dobások számát. Ekkor minden $K > 0$ esetén*

$$\lim_{n\to\infty} \mathrm{P}(|F_n - I_n| \geq K) = 1.$$

Azaz akármilyen nagynak is választjuk a K értékét, a dobások számának növelésével egyre nagyobb a valószínűsége (határértékben 1), hogy a fej és írás dobások számának abszolút eltérése nagyobb K-nál.

De akkor hogyan lehetséges, hogy a fej dobás relatív gyakorisága, azaz az előző tétel jelölésével $\frac{F_n}{n}$ majdnem biztosan $\frac{1}{2}$-hez konvergál? A két tény nem zárja ki egymást, ugyanis

$$\frac{F_n}{I_n} = \frac{F_n}{n - F_n} = \frac{\frac{F_n}{n}}{1 - \frac{F_n}{n}} \to \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1$$

teljesül majdnem biztosan, azaz nem az abszolút eltérés konvergál 0-hoz, hanem a hányadosuk konvergál 1-hez. Ha két számsorozat hányadosa konvergál 1-hez, abból nem következik, hogy az abszolút eltérésük 0-hoz tart. Például

$$\frac{n^2 + n}{n^2} \to 1 \quad \text{de} \quad |n^2 + n - n^2| = n \to \infty.$$

1.4. A nagy számok Marcinkiewicz – Zygmund-féle erős törvénye

1.7. feladat. Láttuk, hogy bizonyos feltételekkel a $\frac{\xi_1 + \cdots + \xi_n}{n}$ majdnem biztosan konvergens. Mi történik, ha a nevezőben n helyett n^q áll? Először vizsgálja meg a $q > 1$ esetet 10 000 dobás után egy szabályos dobókockával.

Megoldás. Nyisson egy új Excel lapot.

1. Jelölje ki az `A1:A10000` cellatartományt. Ehhez a Név mezőbe írja be, hogy `A1:A10000`, majd Enter.

2. Gépelje be a következőt: `=VÉLETLEN.KÖZÖTT(1;6)`, majd Ctrl + Enter.
3. A `C1` cellába gépelje be, hogy `q=` a `D1` cellába pedig, hogy `2`. Ezzel most a $q = 2$ esetet fogjuk vizsgálni.
4. A `B1` cellába gépelje be, hogy `=SZUM(A$1:A1)/(SOR(A1)^D$1)`, majd kattintson kétszer a kitöltőjelre. Ezzel minden dobás után megadja a $\frac{\xi_1+\cdots+\xi_n}{n^q}$ értékét, ahol n az addigi dobások száma, $\xi_1 + \cdots + \xi_n$ pedig az addigi dobások értékeinek összege.
5. Ennek a sorozatnak a vizsgálatához jelölje ki a `B` oszlopot, Beszúrás › Diagramok › Pont- (xy) vagy buborékdiagram beszúrása › Pont vonalakkal. A vízszintes tengely minimumát állítsa 0-ra, maximumát 10000-re. A függőleges tengely minimumát állítsa 0-ra, maximumát pedig 2-re. Az F9 többszöri megnyomásával több dobássorozat esetén vizsgálhatja a viselkedést.
6. A `D1` cella javításával vizsgáljon több $q > 1$ esetet.

Tapasztalat. Ahogy egyre nagyobb a q értéke, a sorozat egyre gyorsabban tart 0-hoz. Ez a viselkedés még $q = 1{,}5$ esetén is elég jól látható, de innen ahogy közeledünk 1-hez a q értékével, már 10 000 kísérlet során nem dönthető el tapasztalati úton, hogy mi történik valójában.

Ehhez vegyük észre, hogy

$$\frac{\xi_1 + \cdots + \xi_n}{n^q} = \frac{\xi_1 + \cdots + \xi_n}{n} \cdot \frac{1}{n^{q-1}}.$$

Mivel $q > 1$ miatt $\frac{1}{n^{q-1}} \to 0$, így $\frac{\xi_1+\cdots+\xi_n}{n^q}$ egy konvergens és egy nullsorozat szorzata, vagyis maga is nullsorozat. Tehát a q-nak 1-hez közeli értékeinél csak azért nem tudtuk tapasztalat alapján megfigyelni a 0-hoz történő konvergenciát, mert a konvergencia nagyon lassú ebben az esetben, így 10 000 kísérlet után még nem volt érzékelhető.

Az erre vonatkozó állítást a következő tétel mondja ki:

Tétel (A nagy számok Marcinkiewicz – Zygmund-féle erős törvénye $q > 1$ esetén)**.** *Ha* $\xi_1, \xi_2, \ldots$ *független, azonos eloszlású valószínűségi változók,* $q > 1$ *és* $\mathrm{E}|\xi_1|^{1/q} < \infty$*, akkor majdnem biztosan teljesül, hogy*

$$\lim_{n\to\infty} \frac{\xi_1 + \cdots + \xi_n}{n^q} = 0.$$

1.8. feladat. Az előző feladatot vizsgálja meg a $q \leq 1$ esetben is.

Megoldás. Könnyen látható, hogy $q \leq 0$ esetben általános konvergenciatulajdonság nem fogalmazható meg, mert az ekkor alapvetően a $\xi_1 + \cdots + \xi_n$ összeg tulajdonságaitól függ. A $q = 1$ eset sem érdekes, mert ezt már korábban tárgyaltuk. Így csak a $0 < q < 1$ esetet vizsgáljuk. Dolgozzon az előző Excel lapon.

1. A függőleges tengely maximumát állítsa 100-ra.
2. A `D1` cellába írja be, hogy `0,9`, amivel a $q = 0{,}9$ esetet vizsgálhatja. Az F9 többszöri megnyomásával több dobássorozat esetén is megfigyelheti a viselkedést.
3. Próbálja ki a $q = 0{,}8\ 0{,}7 \ldots 0{,}1$ eseteket is.

Tapasztalat. Egyetlen esetben sem figyelhető meg konvergens viselkedés. Vagyis úgy tűnik, hogy ekkor nem teljesül a Marcinkiewicz – Zygmund-féle törvény. Kérdés, hogy valamilyen korlátozással erre az esetre is lehet-e valamilyen törvényt találni?

1.9. feladat. Módosítsa az előző feladatot úgy, hogy most minden dobás értékéből levonja a dobás várható értékét (azaz 3,5-et). Ezzel egy olyan valószínűségi változóra vonatkozó mintarealizációt kap, melynek 0 a várható értéke. Az ilyen valószínűségi változót *centráltnak* nevezzük.

Megoldás. Dolgozzon az előző Excel lapon.

1. Javítsa ki az `A1` cella tartalmát erre: `=VÉLETLEN.KÖZÖTT(1;6)-3,5`, majd kattintson kétszer a kitöltőjelre.
2. A függőleges tengely minimumát állítsa -2-re, maximumát pedig 2-re.
3. A `D1` cellába írja be, hogy `0,9`, amivel a $q = 0{,}9$ esetet vizsgálhatja. Az F9 többszöri megnyomásával több dobássorozat esetén is megfigyelheti a viselkedést.
4. Próbálja ki a $q = 0{,}8\ 0{,}7 \ldots 0{,}1$ eseteket is.

Tapasztalat. Ha $0{,}5 < q \leq 1$, akkor a sorozat 0-hoz konvergál. Ha $0 < q \leq 0{,}5$, akkor nem figyelhető meg ilyen viselkedés.

Ezt a következő tétel magyarázza:

Tétel (A nagy számok Marcinkiewicz – Zygmund-féle erős törvénye $0{,}5 < q \leq 1$ esetén)**.** *Ha $\xi_1, \xi_2, \ldots$ független, azonos eloszlású centrált valószínűségi változók, $0{,}5 < q \leq 1$ és $\mathrm{E}|\xi_1|^{1/q} < \infty$, akkor majdnem biztosan teljesül, hogy*

$$\lim_{n\to\infty} \frac{\xi_1 + \cdots + \xi_n}{n^q} = 0.$$

1.5. Centrális határeloszlási tétel

Láttuk, hogy $q \leq 0{,}5$ esetén nem működik a nagy számok törvénye. De a határon, azaz $q = 0{,}5$ esetén nagyon érdekes dolog történik, ami az ún. normális eloszlással kapcsolatos.

Definíció. Legyen $m \in \mathbb{R}$ és $\sigma > 0$. A ξ valószínűségi változót m és σ paraméterű normális eloszlásúnak nevezzük, ha az eloszlásfüggvénye minden $x \in \mathbb{R}$ esetén

$$\mathrm{P}(\xi < x) = \Phi\left(\frac{x - m}{\sigma}\right),$$

ahol

$$\Phi\colon \mathbb{R} \to \mathbb{R}, \quad \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-y^2/2}\,\mathrm{d}y.$$

Az m és σ paraméterű normális eloszlású valószínűségi változó várható értéke m, szórása σ és sűrűségfüggvénye $x \in \mathbb{R}$ helyen

$$\frac{1}{\sigma}\varphi\left(\frac{x-m}{\sigma}\right),$$

ahol

$$\varphi\colon \mathbb{R} \to \mathbb{R}, \quad \varphi(t) = \Phi'(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Tétel (Centrális határeloszlási tétel)**.** *Ha $\xi_1, \xi_2, \ldots$ független, azonos eloszlású, pozitív véges szórású centrált valószínűségi változók, akkor*

$$\frac{\xi_1 + \cdots + \xi_n}{\sqrt{n}}$$

határeloszlása normális $m = 0$ és $\sigma = \mathrm{D}\,\xi_1$ paraméterekkel, azaz minden $x \in \mathbb{R}$ esetén

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{\xi_1 + \cdots + \xi_n}{\sqrt{n}} < x\right) = \Phi\left(\frac{x}{\mathrm{D}\,\xi_1}\right).$$

1.10. feladat. Modellezze a centrális határeloszlási tételt abban az esetben, amikor valószínűségi változó egy szabályos dobókocka esetén a dobott szám értéke 3,5-del csökkentve. (Azért kell kivonni a dobott szám értékéből 3,5-et, mert így 0 lesz a várható érték, azaz centrált lesz a valószínűségi változó.)

Megoldás. A határeloszlás vizsgálatához nem elég egy kísérletsorozat generálása, mint az eddigiekben. Most sok hosszú kísérletsorozatra lesz szükségünk. Az első kísérletsorozat eredményeit generáljuk az `A1:ZZ1` cellatartományban, ami $26 + 26^2 = 26 \cdot 27 = 702$ kísérletet jelent. Ezt a kísérletsorozatot fogjuk 10 000-szer megismételni. Nyisson egy új Excel lapot.

1. Jelölje ki az `A1:ZZ10000` cellatartományt. Ehhez a Név mezőbe írja be, hogy `A1:ZZ10000`, majd Enter.
2. Gépelje be a következőt: `=VÉLETLEN.KÖZÖTT(1;6)-3,5`, majd Ctrl + Enter.
3. A Név mezőbe írja be, hogy `AAA1`, majd Enter. Az `AAA1` cellába gépelje be, hogy `=SZUM(A1:ZZ1)/(GYÖK(OSZLOP(ZZ1)))`. Ezzel az első kísérletsorozat után megadja a $\frac{\xi_1+\cdots+\xi_n}{\sqrt{n}}$ értékét $n = 702$ esetén. A többi megadásához kattintson kétszer az `AAA1` cella kitöltőjelére.
4. Ennek a sorozatnak az eloszlás-vizsgálatához jelölje ki az `AAA` oszlopot, Beszúrás » Diagramok » Statisztikai diagram beszúrása » Hisztogram. Ezzel megkapja a gyakorisági

hisztogramot, amelynek az alakja megfelelő normálás után a sűrűségfüggvényt közelíti. (A megfelelő normálás azt jelenti, hogy a gyakoriság értékeket még osztani kell a részintervallum hosszával és a kísérletsorozatok számával, ami most 10 000.) Az F9 többszöri megnyomásával több dobássorozat esetén vizsgálhatja a viselkedést.

Tapasztalat. Minden esetben a normális eloszlás sűrűségfüggvényére jellemző harang alakú gyakorisági hisztogramot kapjuk.

Ez a harang alakú sűrűségfüggvény nem csak a normális eloszlást jellemzi. Ehhez hasonló például az ún. Cauchy-eloszlás sűrűségfüggvénye is: $\frac{1}{\pi(1+x^2)}$. Hogyan lehetne kideríteni, hogy a tapasztalt harang alakú gyakorisági hisztogram a normális eloszlás miatt van-e? Ennek megállapítására két mérőszámot szoktak használni, a ferdeséget és a lapultságot.

Definíció. A ξ valószínűségi változó eloszlásának *ferdesége* illetve *lapultsága*

$$\frac{\mathrm{E}(\xi - \mathrm{E}\,\xi)^3}{\mathrm{D}^3\,\xi} \quad \text{illetve} \quad \frac{\mathrm{E}(\xi - \mathrm{E}\,\xi)^4}{\mathrm{D}^4\,\xi} - 3,$$

feltéve, hogy ezek a kifejezések léteznek.

Tétel. *Ha ξ normális eloszlású valószínűségi változó, akkor az eloszlásának ferdesége és lapultsága is* 0.

Definíció. Ha $\xi_1, \ldots, \xi_n$ a ξ valószínűségi változóra vonatkozó n darab mérési eredmény (más néven minta), azaz $\xi_1, \ldots, \xi_n$ a ξ-vel azonos eloszlású független valószínűségi változók, akkor ξ *korrigált tapasztalati ferdesége*

$$\frac{n \sum_{i=1}^{n} (\xi_i - \overline{\xi})^3}{(n-1)(n-2)S_n^{*3}}$$

illetve *korrigált tapasztalati lapultsága*

$$\frac{n(n+1) \sum_{i=1}^{n} (\xi_i - \overline{\xi})^4}{(n-1)(n-2)(n-3)S_n^{*4}} - \frac{3(n-1)^2}{(n-2)(n-3)},$$

ahol

$$\overline{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i \quad \text{illetve} \quad S_n^* = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\xi_i - \overline{\xi})^2}$$

a *mintaátlag* illetve *korrigált tapasztalati szórás.*

Egy valószínűségi változó ferdeségét és lapultságát a minta alapján a korrigált tapasztalati ferdeséggel és lapultsággal lehet becsülni.

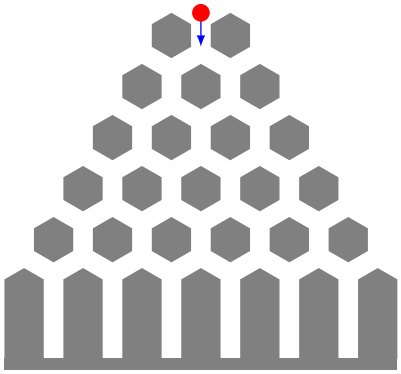
1.11. feladat. Az előző feladatban számolja ki az `AAA` oszlopban található minta korrigált tapasztalati ferdeségét és korrigált tapasztalati lapultságát. Ez alapján lehet-e normális eloszlású?

Megoldás. A korrigált tapasztalati ferdeség a `=FERDESÉG(AAA:AAA)`, illetve a korrigált tapasztalati lapultság a `=CSÚCSOSSÁG(AAA:AAA)` függvényekkel számolható (Excelben a lapultság helyett a csúcsosság kifejezést használják).

Tapasztalat. Mindkét érték minden esetben nagyon közel van 0-hoz, így a minta jól közelíti a normális eloszlást.

1.6. Galton-deszka, bolyongás

A következő sematikus ábrán egy ún. *Galton-deszkát* láthatunk.



A piros golyót a nyíl irányában leejtve nekiütközik egy szürkével jelölt éknek. Ezután 0,5 valószínűséggel jobbra illetve 0,5 valószínűséggel balra esik. Ezután megint egy éknek ütközik stb., majd kiköt valamelyik folyosón. Sok golyót leejtve, azok milyen eloszlásban helyezkednek el a folyosókon?

Ezt a folyamatot a következőképpen is át lehet fogalmazni: Egy szabályos pénzérmét feldobunk. Ha a fej oldalára esik, akkor 1 forintot nyerünk (a golyó jobbra esik), ellenkező esetben pedig 1 forintot veszítünk (a golyó balra esik). Ezt a játékot többször megismételve a nyeremény azt jelképezi, hogy a Galton-deszka melyik folyosóján landolt a golyó. Ezt a játékot leíró folyamatot *bolyongásnak* nevezzük.

Tétel. *Jelentse B_n a bolyongásban a nyeremény értékét n darab játék után. Ekkor B_n minden lehetséges k értéke esetén*

$$\mathrm{P}(B_n = k) = \binom{n}{\frac{n+k}{2}} 2^{-n}.$$

Ha n nagy, akkor a centrális határeloszlási tétel miatt

$$\mathrm{P}(B_n = k) \approx \frac{1}{\sqrt{n}} \varphi\left(\frac{k}{\sqrt{n}}\right).$$

1.12. feladat. Szemléltesse a bolyongást $n = 702$ játék után 10 000 kísérletsorozatban.

Megoldás. Nyisson egy új Excel lapot.

1. Jelölje ki az `A1:ZZ10000` cellatartományt. Ehhez a Név mezőbe írja be, hogy `A1:ZZ10000`, majd Enter.
2. Gépelje be a következőt: `=HA(VÉL()<0,5;1;-1)`, majd Ctrl+Enter. A `VÉL()` értéke a $[0, 1]$ intervallumon egyenletes eloszlású, ezért ennek értéke 0,5 valószínűséggel kisebb mint 0,5. Így a `VÉL()<0,5` esemény modellezheti a fej dobást.
3. A Név mezőbe írja be, hogy `AAA1`, majd Enter. Az `AAA1` cellába gépelje be, hogy `=SZUM(A1:ZZ1)`. Ezzel az első kísérletsorozat után megadja a B_n értékét $n = 702$ esetén. A többi megadásához kattintson kétszer az `AAA1` cella kitöltőjelére.
4. Ennek a sorozatnak az eloszlás-vizsgálatához jelölje ki az `AAA` oszlopot, Beszúrás › Diagramok › Statisztikai diagram beszúrása › Hisztogram. Ezzel megkapja a gyakorisági hisztogramot, amelynek az alakja megfelelő normálás után a $k \mapsto \mathrm{P}(B_n = k)$ függvényt közelíti. (A megfelelő normálás azt jelenti, hogy a gyakoriság értékeket még osztani kell a a kísérletsorozatok számával, ami most 10 000.) Az F9 többszöri megnyomásával több dobássorozat esetén vizsgálhatja a viselkedést.

Tapasztalat. Minden esetben a φ függvényre jellemző harang alakú gyakorisági hisztogramot kapjuk.

1.7. Iterált-logaritmus tétel

A nagy számok Marcinkiewicz – Zygmund-féle erős törvénye szerint, ha $\xi_1, \xi_2, \dots$ független, azonos eloszlású centrált valószínűségi változók és $\mathrm{E}\,|\xi_1|^{1/q} \in \mathbb{R}$, akkor $q > 0{,}5$ esetén az $(\xi_1 + \dots + \xi_n)n^{-q}$ majdnem biztosan 0-hoz tart. Ugyanakkor $q = 0{,}5$ esetén a centrális határeloszlási tétel szerint $(\xi_1 + \dots + \xi_n)n^{-1/2}$ bár nagy valószínűséggel a 0-tól nem távolodik el nagyon, de pozitív valószínűséggel az abszolút értéke tetszőlegesen nagy lehet. Ezen két tétel „között van” az iterált-logaritmus tétel, amely szerint a $\xi_1 + \dots + \xi_n$ összeget alkalmas normáló tényezővel osztva, a sorozat realizációi „nem húzódnak össze 0-ra”, de nem is „kenődnek szét” a számegyenesen.

Tétel (Iterált-logaritmus tétel)**.** *Legyenek $\xi_1, \xi_2, \dots$ független, azonos eloszlású, véges pozitív szórású centrált valószínűségi változók. Ekkor majdnem biztosan teljesül, hogy*

$$\limsup_{n\to\infty} \frac{\xi_1 + \dots + \xi_n}{\sqrt{n \ln(\ln n)}} = \sqrt{2}\,\mathrm{D}\,\xi_1 \quad \textit{és} \quad \liminf_{n\to\infty} \frac{\xi_1 + \dots + \xi_n}{\sqrt{n \ln(\ln n)}} = -\sqrt{2}\,\mathrm{D}\,\xi_1.$$

Ezen tétel alapján a

$$\frac{\xi_1 + \dots + \xi_n}{\sqrt{n \ln(\ln n)}}$$

sorozatnak végtelen sokszor kell $\sqrt{2}\,\mathrm{D}\,\xi_1$ közeléből $-\sqrt{2}\,\mathrm{D}\,\xi_1$ közelébe jutnia és viszont. Azonban a valóságban ezek az átjutások nagyon lassan történnek.

Emlékeztetőként, $\limsup_{n\to\infty} a_n$ (ejtsd: limesz szuperior) az a_n számsorozat legnagyobb, míg $\liminf_{n\to\infty} a_n$ (ejtsd: limesz inferior) a legkisebb torlódási pontját jelenti. Egy sorozatnak egy szám akkor torlódási pontja, ha bármely környezetében végtelen sok tagja van a sorozatnak.

1.13. feladat. Modellezze az iterált-logaritmus tételt abban az esetben, amikor valószínűségi változó egy szabályos dobókocka esetén a dobott szám értéke 3,5-del csökkentve. (Azért kell kivonni a dobott szám értékéből 3,5-et, mert így 0 lesz a várható érték, azaz centrált lesz a valószínűségi változó.)

Megoldás. Nyisson egy új Excel lapot.

1. Jelölje ki az `A1:A10000` cellatartományt. Ehhez a Név mezőbe írja be, hogy `A1:A10000`, majd Enter.
2. Gépelje be a következőt: `=VÉLETLEN.KÖZÖTT(1;6)-3,5`, majd Ctrl + Enter.
3. A `B1` cellába gépelje be, hogy `=SZUM(A$1:A1)/GYÖK(SOR(A1)*LN(LN(SOR(A1))))`, majd kattintson kétszer a `B1` cella kitöltőjelére.
4. Ennek a sorozatnak a vizsgálatához jelölje ki a `B` oszlopot, Beszúrás › Diagramok › Pont- (xy) vagy buborékdiagram beszúrása › Pont vonalakkal. A vízszintes tengely minimumát állítsa 0-ra, maximumát 10000-re. A függőleges tengely minimumát állítsa -4-re, maximumát pedig 4-re. Az F9 többszöri megnyomásával több dobássorozat esetén vizsgálhatja a viselkedést.

2. fejezet

Matematikai statisztika

2.1. Szórásanalízis

2.1.1. Egyszeres osztályozás

Vizsgáljuk a ξ valószínűségi változót egyetlen tényező r darab különböző szintjén. Az i. szinthez tartozó valószínűségi változót jelölje ξ_i. Feltesszük, hogy ξ_i normális eloszlású m_i várható értékkel és σ szórással, ahol az m_i, σ paraméterek ismeretlenek. Feltesszük még, hogy $\xi_1, \dots, \xi_r$ függetlenek. A ξ_i-re vonatkozó minta legyen $\xi_{i1}, \xi_{i2}, \dots, \xi_{in_i}$.

Arról szeretnénk dönteni, hogy az egyetlen tényező különböző szintjei hatással vannak-e a mért értékekre? A nullhipotézisünk az lesz, hogy nemleges a válasz, vagyis a különböző szintek nincsenek hatással a mért értékekre. Ez azzal ekvivalens, hogy $m_1 = m_2 = \dots = m_r$.

Legyen α annak a valószínűsége, hogy amennyiben igaz a nullhipotézis, a minta alapján mi mégis elutasítjuk azt. Ezt *első fajú hibának* szoktuk nevezni. Vezessük be még a következő jelöléseket is:

$$n := n_1 + n_2 + \dots + n_r, \quad \overline{\xi}_{..} := \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} \xi_{ij}, \quad \overline{\xi}_{i.} := \frac{1}{n_i} \sum_{j=1}^{n_i} \xi_{ij},$$

$$Q_1 := \sum_{i=1}^{r} n_i \big(\overline{\xi}_{i.} - \overline{\xi}_{..} \big)^2 \text{ a szintek közötti eltérések négyzetösszege,}$$

$$Q_2 := \sum_{i=1}^{r} \sum_{j=1}^{n_i} \big(\xi_{ij} - \overline{\xi}_{i.} \big)^2 \text{ a szinteken belüli eltérések (véletlen hibák) négyzetösszege,}$$

$$\mathsf{F} := \frac{n-r}{r-1} \cdot \frac{Q_1}{Q_2},$$

továbbá F az eloszlásfüggvénye az $r-1$ és $n-r$ szabadsági fokú F-eloszlásnak (lásd például [3, 36. oldal]). Ekkor $1 - F(\mathsf{F}) \geq \alpha$ esetén elfogadjuk a nullhipotézist, ellenkező esetben pedig elutasítjuk. Ezt a statisztikai próbát nevezzük *egyszeres osztályozású*

szórásanalízisnek.

Excelben ennek kiszámolásához az Adatok » Adatelemzés menüpontot fogjuk használni. Ehhez először aktiválni kell az *Analysis ToolPak* bővítményt a következő módon: Fájl » Beállítások » Bővítmények majd Ugrás gomb. Pipálja ki az *Analysis ToolPak* sort majd OK.

2.1. feladat. Egy gazdaságban a búza terméshozamára vagyunk kíváncsiak. Azt vizsgáljuk, hogy a búza fajtája milyen hatással van a terméshozamra (most ez az egyetlen tényező). A gazdaság 3 különböző búzafajtát termeszt, azaz a vizsgált tényezőnek most $r = 3$ különböző szintje van. Az 1. fajtát 4, a 2. fajtát 3, végül a 3. fajtát 5 különböző parcellán termesztik ($n_1 = 4, n_2 = 3, n_3 = 5$). A ξ_{ij} jelentse az i. fajta j. parcellán kapott terméshozamát tonna/hektár-ban mérve. A kapott mintarealizációk a következők:

$$\begin{array}{llll} \xi_{11} = 5{,}24 & \xi_{12} = 4{,}17 & \xi_{13} = 4{,}35 & \xi_{14} = 4{,}77 \\ \xi_{21} = 5{,}09 & \xi_{22} = 6{,}05 & \xi_{23} = 5{,}89 & \\ \xi_{31} = 4{,}18 & \xi_{32} = 4{,}10 & \xi_{33} = 4{,}17 & \xi_{34} = 3{,}98 \quad \xi_{35} = 3{,}60 \end{array}$$

Döntsön $\alpha = 0{,}05$ esetén arról a nullhipotézisről, hogy a különböző búzafajták nincsenek hatással a terméshozamra.

Megoldás. Nyisson egy új Excel lapot. Gépelje be az adatokat a következő ábra szerint:

	A	B	C	D	E
1	5,24	4,17	4,35	4,77	
2	5,09	6,05	5,89		
3	4,18	4,1	4,17	3,98	3,6

A feladatot az Adatok » Adatelemzés menüponttal fogjuk megoldani. A legördülő listában válassza az *Egytényezős varianciaanalízis* sort, majd OK.

Bemeneti tartomány: `$A$1:$E$3`
Csoportosítási alap: Sorok
☐ *Feliratok az első oszlopban*
Alfa: `0,05`
Kimeneti tartomány: `$A$4`
OK

Ekkor a következő táblázatot kapjuk:

Tényezők	*SS*	*df*	*MS*	*F*	*p-érték*	*F krit.*
Csoportok között	5,2334	2	2,6167	16,3286	0,00101	4,256495
Csoporton belül	1,4423	9	0,1603			

A táblázatban $Q_1 = 5{,}2334$, $Q_2 = 1{,}4423$, $\mathsf{F} = 16{,}3286$ és $1 - F(\mathsf{F}) = 0{,}0010$. Mivel $1 - F(\mathsf{F}) = 0{,}0010 < 0{,}05 = \alpha$, ezért elutasítjuk a nullhipotézist, tehát a búza fajtája hatással van a terméshozamra. (Természetesen minden $\alpha > 0{,}001$ választás esetén ugyanez lett volna a döntés.)

2.1.2. Kétszeres osztályozás interakció nélkül

Vizsgáljuk két tényező hatását egy ξ valószínűségi változóra. Legyen az 1. tényezőnek r_1, míg a 2. tényezőnek r_2 különböző szintje. Jelölje ξ_{ij} az 1. tényező i. szintjéhez és a 2. tényező j. szintjéhez tartozó valószínűségi változót. Feltesszük, hogy ξ_{ij} független valószínűségi változók, melyek normális eloszlásúak m_{ij} várható értékkel és σ szórással, ahol minden paraméter ismeretlen.

Két nullhipotézist fogunk vizsgálni. Az első szerint az 1. tényező különböző szintjei nincsenek hatással ξ-re, a második szerint pedig a 2. tényező különböző szintjei nincsenek hatással ξ-re.

Legyen α az első fajú hiba valószínűsége. Vezessük be a következő jelöléseket:

$$\overline{\xi}_{..} := \frac{1}{r_1 r_2} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \xi_{ij},$$

$$\overline{\xi}_{i.} := \frac{1}{r_2} \sum_{j=1}^{r_2} \xi_{ij} \quad (i = 1, 2, \ldots, r_1),$$

$$\overline{\xi}_{.j} := \frac{1}{r_1} \sum_{i=1}^{r_1} \xi_{ij} \quad (j = 1, 2, \ldots, r_2),$$

$$Q_1 := r_2 \sum_{i=1}^{r_1} \big(\overline{\xi}_{i.} - \overline{\xi}_{..}\big)^2 \text{ az 1. tényező szintjei közötti eltérések négyzetösszege,}$$

$$Q_2 := r_1 \sum_{j=1}^{r_2} \big(\overline{\xi}_{.j} - \overline{\xi}_{..}\big)^2 \text{ a 2. tényező szintjei közötti eltérések négyzetösszege,}$$

$$Q_3 := \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \big(\xi_{ij} - \overline{\xi}_{i.} - \overline{\xi}_{.j} + \overline{\xi}_{..}\big)^2 \text{ hibatag,}$$

$$\mathsf{F}_1 := (r_2 - 1) \cdot \frac{Q_1}{Q_3}, \quad \mathsf{F}_2 := (r_1 - 1) \cdot \frac{Q_2}{Q_3},$$

továbbá F_1 az eloszlásfüggvénye az $r_1 - 1$ és $(r_1 - 1)(r_2 - 1)$ szabadsági fokú F-eloszlásnak, illetve F_2 az eloszlásfüggvénye az $r_2 - 1$ és $(r_1 - 1)(r_2 - 1)$ szabadsági fokú F-eloszlásnak.

Ekkor $1 - F_i(\mathsf{F}_i) \geq \alpha$ esetén elfogadjuk az i-edik nullhipotézist, ellenkező esetben pedig elutasítjuk $(i = 1, 2)$. Ezt a statisztikai próbát nevezzük *kétszeres osztályozású interakció nélküli szórásanalízisnek*.

2.2. feladat. Az előző feladatot tovább gondolva tegyük fel, hogy nem csak a búza fajtáját, hanem a parcella talajtípusát is vizsgálni szeretnénk a terméshozamot illetően, vagyis nem egy, hanem két tényező hatását figyeljük. Tegyük fel, hogy 3 fajta búzát 4 típusú talajba vetettek ($r_1 = 3, r_2 = 4$). Azaz az 1. tényezőnek 3, a 2. tényezőnek pedig 4 szintje van. A ξ_{ij} jelentse az i. búzafajta j. talajtípuson vett terméshozamát. A kapott mintarealizációk a következők:

$$\begin{array}{llll} \xi_{11} = 7{,}51 & \xi_{12} = 6{,}34 & \xi_{13} = 5{,}07 & \xi_{14} = 6{,}17 \\ \xi_{21} = 5{,}43 & \xi_{22} = 4{,}81 & \xi_{23} = 3{,}42 & \xi_{14} = 4{,}00 \\ \xi_{31} = 5{,}76 & \xi_{32} = 4{,}71 & \xi_{33} = 4{,}45 & \xi_{34} = 4{,}33 \end{array}$$

Döntsön $\alpha = 0{,}05$ esetén a következő nullhipotézisekről:

1. A különböző búzafajták nincsenek hatással a terméshozamra.
2. A különböző talajtípusok nincsenek hatással a terméshozamra.

Megoldás. Nyisson egy új Excel lapot. Gépelje be az adatokat a következő ábra szerint:

	A	B	C	D
1	7,51	6,34	5,07	6,17
2	5,43	4,81	3,42	4
3	5,76	4,71	4,45	4,33

A feladatot az Adatok » Adatelemzés menüponttal fogjuk megoldani. A legördülő listában válassza az *Kéttényezős varianciaanalízis ismétlések nélkül* sort, majd OK.

Bemeneti tartomány: `$A$1:$D$3`
☐ *Feliratok*
Alfa: `0,05`
Kimeneti tartomány: `$A$4`
OK

Ekkor a következő táblázatot kapjuk:

Tényezők	*SS*	*df*	*MS*	*F*	*p-érték*	*F krit.*
Sorok	7,6532	2	3,8266	35,9278	0,0005	5,1433
Oszlopok	5,9744	3	1,9915	18,6978	0,0019	4,7571
Hiba	0,6391	6	0,1065			

A táblázatban $Q_1 = 7{,}6532$, $Q_2 = 5{,}9744$, $Q_3 = 0{,}6391$, $\mathsf{F}_1 = 35{,}9278$, $\mathsf{F}_2 = 18{,}6978$, $1 - F_1(\mathsf{F}_1) = 0{,}0005$ és $1 - F_2(\mathsf{F}_2) = 0{,}0019$.

Mivel $1 - F_1(\mathsf{F}_1) = 0{,}0005 < 0{,}05 = \alpha$, ezért elutasítjuk az első nullhipotézist, azaz a különböző búzafajták hatással vannak a terméshozamra.

Másrészt $1 - F_2(\mathsf{F}_2) = 0{,}0019 < 0{,}05 = \alpha$, ezért elutasítjuk a második nullhipotézist is, azaz a különböző talajtípusok hatással vannak a terméshozamra.

Ebben az esetben azt nem tudjuk vizsgálni, hogy a két tényező milyen hatással van egymásra, azaz, hogy egy konkrét búzafajta különbözőképpen terem-e a különböző talajtípusokon, vagy hogy egy konkrét talajtípuson különbözőképpen teremnek-e a különböző búzafajták, mert minden búzafajta–talajtípus kombinációból csak egy mintaelemünk van. A két tényező egymásra hatásának vizsgálatához több mintaelemre van szükség minden kombináció esetén. Ezzel foglalkozik a következő szórásanalízis típus.

2.1.3. Kétszeres osztályozás interakcióval

Vizsgáljuk két tényező hatását egy ξ valószínűségi változóra. Legyen az 1. tényezőnek r_1, míg a 2. tényezőnek r_2 különböző szintje. Jelölje ξ_{ij} az 1. tényező i. szintjéhez és a 2. tényező j. szintjéhez tartozó valószínűségi változót. Feltesszük, hogy ξ_{ij} független valószínűségi változók, melyek normális eloszlásúak m_{ij} várható értékkel és σ szórással, ahol minden paraméter ismeretlen. Minden ξ_{ij}-hez készítsünk egy s elemű mintát: $\xi_{ij1}, \xi_{ij2}, \ldots, \xi_{ijs}$. Három nullhipotézist fogunk vizsgálni:

1. az 1. tényező különböző szintjei nincsenek hatással ξ-re,
2. a 2. tényező különböző szintjei nincsenek hatással ξ-re,
3. a két tényező együttes hatása nem befolyásolja a ξ értékét.

Legyen α az első fajú hiba valószínűsége. Vezessük be a következő jelöléseket:

$$\overline{\xi}_{...} := \frac{1}{r_1 r_2 s} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{k=1}^{s} \xi_{ijk},$$

$$\overline{\xi}_{i..} := \frac{1}{r_2 s} \sum_{j=1}^{r_2} \sum_{k=1}^{s} \xi_{ijk} \quad (i = 1, 2, \ldots, r_1),$$

$$\overline{\xi}_{.j.} := \frac{1}{r_1 s} \sum_{i=1}^{r_1} \sum_{k=1}^{s} \xi_{ijk} \quad (j = 1, 2, \ldots, r_2),$$

$$\overline{\xi}_{ij.} := \frac{1}{s} \sum_{k=1}^{s} \xi_{ijk} \quad (i = 1, 2, \ldots, r_1;\ j = 1, 2, \ldots, r_2),$$

$$Q_1 := r_2 s \sum_{i=1}^{r_1} \big(\overline{\xi}_{i..} - \overline{\xi}_{...} \big)^2 \text{ az 1. tényező szintjei közötti eltérések négyzetösszege,}$$

$$Q_2 := r_1 s \sum_{j=1}^{r_2} \big(\overline{\xi}_{.j.} - \overline{\xi}_{...} \big)^2 \text{ a 2. tényező szintjei közötti eltérések négyzetösszege,}$$

$$Q_3 := s \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \big(\overline{\xi}_{ij.} - \overline{\xi}_{i..} - \overline{\xi}_{.j.} + \overline{\xi}_{...} \big)^2 \text{ a két tényező együttes hatásaiból adódó eltérések négyzetösszege,}$$

$$Q_4 := \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{k=1}^{s} \big(\xi_{ijk} - \overline{\xi}_{ij.} \big)^2 \text{ hibatag,}$$

$$\mathsf{F}_1 := \frac{r_1 r_2 (s-1)}{r_1 - 1} \cdot \frac{Q_1}{Q_4}, \quad \mathsf{F}_2 := \frac{r_1 r_2 (s-1)}{r_2 - 1} \cdot \frac{Q_2}{Q_4}, \quad \mathsf{F}_3 := \frac{r_1 r_2 (s-1)}{(r_1 - 1)(r_2 - 1)} \cdot \frac{Q_3}{Q_4},$$

továbbá F_1 az eloszlásfüggvénye az $r_1 - 1$ és $r_1 r_2(s-1)$ szabadsági fokú F-eloszlásnak, F_2 az eloszlásfüggvénye az $r_2 - 1$ és $r_1 r_2(s-1)$ szabadsági fokú F-eloszlásnak, továbbá F_3 az eloszlásfüggvénye az $(r_1-1)(r_2-1)$ és $r_1 r_2(s-1)$ szabadsági fokú F-eloszlásnak.

Ekkor $1 - F_i(\mathsf{F}_i) \geq \alpha$ esetén elfogadjuk az i-edik nullhipotézist, ellenkező esetben pedig elutasítjuk $(i = 1, 2, 3)$. Ezt a statisztikai próbát nevezzük *kétszeres osztályozású szórásanalízisnek interakcióval.*

2.3. feladat. Folytatva az előző feladatot, ha a két tényező közötti kapcsolatot is vizsgálni szeretnénk, akkor minden búzafajta–talajtípus kombinációból több mérést kell végeznünk. Legyen az előző feladatban minden kombinációra 3 mérésünk ($s = 3$). Jelentse ξ_{ijk} az i. búzafajta j. talajtípuson vett terméshozamára vonatkozó k. mérési eredményt. A kapott mintarealizációk a következők:

$$\begin{array}{llll}
\xi_{111} = 7{,}51 & \xi_{121} = 6{,}34 & \xi_{131} = 5{,}07 & \xi_{141} = 6{,}17 \\
\xi_{112} = 7{,}03 & \xi_{122} = 5{,}81 & \xi_{132} = 4{,}19 & \xi_{142} = 5{,}90 \\
\xi_{113} = 6{,}91 & \xi_{123} = 6{,}61 & \xi_{133} = 5{,}27 & \xi_{143} = 6{,}28 \\
\xi_{211} = 5{,}43 & \xi_{221} = 4{,}81 & \xi_{231} = 3{,}42 & \xi_{141} = 4{,}00 \\
\xi_{212} = 4{,}95 & \xi_{222} = 3{,}82 & \xi_{232} = 3{,}19 & \xi_{142} = 3{,}80 \\
\xi_{213} = 5{,}48 & \xi_{223} = 4{,}18 & \xi_{233} = 2{,}02 & \xi_{143} = 3{,}94 \\
\xi_{311} = 5{,}76 & \xi_{321} = 4{,}71 & \xi_{331} = 4{,}45 & \xi_{341} = 4{,}33 \\
\xi_{312} = 5{,}90 & \xi_{322} = 5{,}24 & \xi_{332} = 4{,}65 & \xi_{342} = 5{,}41 \\
\xi_{313} = 6{,}01 & \xi_{323} = 4{,}07 & \xi_{333} = 4{,}59 & \xi_{343} = 5{,}70
\end{array}$$

Döntsön a következő nullhipotézisekről:

1. A különböző búzafajták nincsenek hatással a terméshozamra.
2. A különböző talajtípusok nincsenek hatással a terméshozamra.
3. A búzafajta és a talajtípus között nincs kapcsolat a terméshozamot illetően.

Megoldás. Nyisson egy új Excel lapot. Gépelje be az adatokat a következő ábra szerint:

	A	B	C	D	E
1					
2		7,51	6,34	5,07	6,17
3		7,03	5,81	4,19	5,9
4		6,91	6,61	5,27	6,28
5		5,43	4,81	3,42	4
6		4,95	3,82	3,19	3,8
7		5,48	4,18	2,02	3,94
8		5,76	4,71	4,45	4,33
9		5,9	5,24	4,65	5,41
10		6,01	4,07	4,59	5,7

Az első sorba illetve oszlopba ne vigyen adatokat, oda most csak feliratok kerülhetnek.

A feladatot az Adatok » Adatelemzés menüponttal fogjuk megoldani. A legördülő listában válassza a *Kéttényezős varianciaanalízis ismétlésekkel* sort, majd OK.

Bemeneti tartomány: `$A$1:$E$10`
Mintánként hány sor: `3`
Alfa: `0,05`
Kimeneti tartomány: `$A$11`
OK

Ekkor a következő táblázatot kapjuk:

Tényezők	*SS*	*df*	*MS*	*F*	*p-érték*	*F krit.*
Minta	24,1034	2	12,0517	59,3583	5E-10	3,4028
Oszlopok	18,2751	3	6,0917	30,0035	3E-08	3,0088
Kölcsönhatás	2,0206	6	0,3368	1,6587	0,1746	2,5082
Belül	4,8728	24	0,2030			

A táblázatban $Q_1 = 24{,}1034$, $Q_2 = 18{,}2751$, $Q_3 = 2{,}0206$, $Q_4 = 4{,}8728$, $\mathsf{F}_1 = 59{,}3583$, $\mathsf{F}_2 = 30{,}0035$, $\mathsf{F}_3 = 1{,}6587$, $1 - F_1(\mathsf{F}_1) = 5 \cdot 10^{-10}$, $1 - F_2(\mathsf{F}_2) = 3 \cdot 10^{-8}$ és $1 - F_3(\mathsf{F}_3) = 0{,}1746$.

$1 - F_1(\mathsf{F}_1) = 5 \cdot 10^{-10} < 0{,}05 = \alpha$, ezért elutasítjuk az 1. nullhipotézist, azaz a különböző búzafajták hatással vannak a terméshozamra.

$1 - F_2(\mathsf{F}_2) = 3 \cdot 10^{-8} < 0{,}05 = \alpha$, ezért elutasítjuk a 2. nullhipotézist, azaz a különböző talajtípusok hatással vannak a terméshozamra.

$1 - F_3(\mathsf{F}_3) = 0{,}1746 > 0{,}05 = \alpha$, ezért elfogadjuk a 3. nullhipotézist, azaz a búzafajta és a talajtípus között nincs kapcsolat a terméshozamot illetően.

2.2. Lineáris regresszió

Az $\eta, \xi_1, \ldots, \xi_k$ valószínűségi változók esetén adjuk meg a legjobb

$$\eta \simeq g(\xi_1, \ldots, \xi_k) \qquad (*)$$

közelítést adó g függvényt. Ezt úgy értjük, hogy az

$$\mathrm{E}\big(\eta - g(\xi_1, \ldots, \xi_k)\big)^2$$

értékét kell minimalizálni. Ez az úgynevezett *legkisebb négyzetek elve*. Az így kapott g függvényt *regressziós felületnek* nevezzük. Ha g lineáris, akkor $k = 1$ illetve $k = 2$ esetén a g függvényt *elsőfajú regressziós egyenesnek* illetve *elsőfajú regressziós síknak* nevezzük. A regressziós felület továbbá $\xi_1, \ldots, \xi_k$ ismeretében η megbecsülhető $(*)$ alapján.

Sok esetben a regressziós felület meghatározása bonyolult feladat. Ilyenkor azzal egyszerűsíthetjük a problémát, hogy $\mathrm{E}\big(\eta - g(\xi_1, \dots, \xi_k)\big)^2$ minimumát csak a

$$g(x_1, \dots, x_k) = a_0 + a_1 x_1 + \dots + a_k x_k \quad (a_0, a_1, \dots, a_k \in \mathbb{R})$$

alakú – azaz lineáris – függvények között keressük. Ezt a problémát *lineáris regressziónak* illetve a minimumhoz tartozó függvényben szereplő $a_0, \dots, a_k$ konstansokat a *lineáris regresszió együtthatóinak* nevezzük. Az így kapott g függvényt $k = 1$ illetve $k = 2$ esetén *másodfajú regressziós egyenesnek* illetve *másodfajú regressziós síknak* nevezzük.

A lineáris regresszió együtthatóinak értékét a gyakorlatban kellő információ hiányában nem tudjuk kiszámolni, így az $(\eta, \xi_1, \dots, \xi_k)$-ra vonatkozó minta alapján kell ezeket megbecsülni. Legyen ez a minta $(\eta_i, \xi_{i1}, \dots, \xi_{ik})$ $i = 1, \dots, n$. Bizonyítható, hogy ekkor az

$$(X^\top X)^{-1} X^\top Y$$

oszlopvektor j-edik komponense (amit a továbbiakban $\widehat{a}_j$ módon fogunk jelölni) az a_j jó becslése, ahol

$$Y = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix} \quad \text{és} \quad X = \begin{pmatrix} 1 & \xi_{11} & \dots & \xi_{1k} \\ 1 & \xi_{21} & \dots & \xi_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \xi_{n1} & \dots & \xi_{nk} \end{pmatrix}.$$

Ezután az $\eta \simeq \widehat{a}_0 + \widehat{a}_1 \xi_1 + \dots + \widehat{a}_k \xi_k$ közelítést fogjuk használni az η becslésére.

Ha $k = 1$ akkor $\xi := \xi_1$ jelöléssel az η és ξ közötti lineáris kapcsolat feltételezésének jogosságát az ún. *determinációs együtthatóval* (jele R^2) szokták mérni, amely a tapasztalati korrelációs együtthatójuk négyzete (lásd [3, 55. oldal]). Ennek értéke minél közelebb van 1-hez, annál jobban feltételezhető a lineáris kapcsolat. Ha 0-hoz van közel, akkor a két valószínűségi változó között nem célszerű függvénykapcsolatot keresni, mert feltehetőleg függetlenek (pontosabban korrelálatlanok).

2.4. feladat. Szeretnénk előrejelezni a talajvízszint mértékét az őszi csapadék mennyisége alapján. Jelentse η a talajvízszintet mm-ben és ξ az őszi csapadék mennyiségét cm-ben. Az (η, ξ)-re vonatkozó elmúlt 18 évi mérésből származó mintarealizációt a 2.1. táblázat tartalmazza.

a. Ez alapján becsülje meg a lineáris regresszió együtthatóit.

b. Számolja ki a determinációs együtthatót. Ez alapján jónak tekinthető a lineáris közelítés?

c. A becsült másodfajú regressziós egyenest ábrázolja a mintarealizációval együtt.

d. Becsülje meg a talajvízszintet, ha az őszi csapadék 29,6 cm.

Megoldás. Jelölje ki a mintát (mindkét oszlopot) a pdf-ben. Ctrl + C segítségével tegye a vágólapra. Nyisson meg egy üres munkalapot Excelben, lépjen az `A1` cellára és Ctrl + V segítségével illessze be. Most még a két adatsor egy oszlopban van. Ezek szétválasztásához tegye a következőket: Adatok » Szövegből oszlopok, majd ◉ *Fix széles* Tovább Tovább Befejezés

a. Az $\widehat{a}_1, \widehat{a}_0$ együtthatók kiszámolásához jelölje ki a `C1:D1` cellatartományt, majd gépelje be a következő tömbképletet: `=LIN.ILL(A1:A18;B1:B18)`, majd Ctrl + Shift + Enter. Ennek hatására `C1` fogja $\widehat{a}_1$ értékét, illetve `D1` fogja $\widehat{a}_0$ értékét tartalmazni.

b. A determinációs együtthatót a következő módon számolhatja ki:

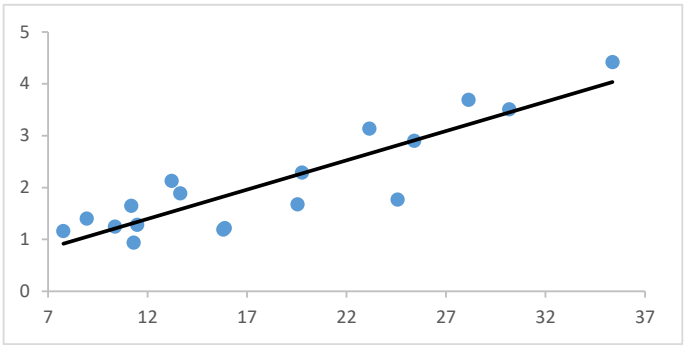
`=RNÉGYZET(A1:A18;B1:B18)`.

Ennek értéke négy tizedesjegyre kerekítve 0,7931. Ez azt jelenti, hogy a lineáris közelítés jónak mondható.

c. Jelölje ki az η-ra vonatkozó `A1:A18` cellatartományt, majd Beszúrás » Diagramok » Pont- (xy) vagy buborékdiagram beszúrása » Pont

Lépjen a diagramterületre, majd helyi menüből (jobb egérgomb) válassza ki az Adatok kijelölése pontot, majd Szerkesztés *Adatsor X értékei:* `=Munka1!$B$1:$B$18` OK OK. Ezzel megjelennek a mintarealizáció pontjai.

Következzen a másodfajú regressziós egyenes becslésének a meghúzása (az Excel ezt *trendvonalnak* nevezi). Lépjen rá valamelyik kék jelölő pontra. Helyi menüből válassza a Trendvonal felvétele pontot és válassza ki a *Lineáris* típust.



d. A talajvízszint becslése, ha az őszi csapadék 29,6 cm, ezek alapján $\widehat{a}_0 + 29{,}6\widehat{a}_1$, amit a következő módon is kiszámolhatunk: `=TREND(A1:A18;B1:B18;29,6)`. A kapott érték 3,38 két tizedesjegyre kerekítve. Tehát 29,6 cm csapadék lehullása után az adatok alapján 3,38 mm-re becsüljük a talajvízszintet.

2.5. feladat. Jelentse η a Duna egy árhullámának tetőző vízállását Budapesten cm-ben, ξ_1 az árhullámot kiváltó csapadék mennyiségét mm-ben és ξ_2 a Duna vízállását Budapestnél az esőzés kezdetekor cm-ben. Szeretnénk előrejelezni η mértékét ξ_1 és ξ_2 mennyisége alapján. Az (η, ξ_1, ξ_2)-re vonatkozó elmúlt 26 évi mérésből származó

mintarealizációt a 2.2. táblázat tartalmazza.

a. Ez alapján becsülje meg a lineáris regresszió együtthatóit.

b. Az idén az árhullámot kiváltó csapadék 102 mm volt, illetve a Duna vízállása Budapestnél az esőzés kezdetekor 648 cm volt. Ezekből az adatokból becsülje meg, hogy a Duna árhullámának tetőző vízállása Budapesten hány cm lesz.

Megoldás. Jelölje ki a mintát (mindhárom oszlopot) a pdf-ben. Ctrl + C segítségével tegye a vágólapra. Nyisson meg egy üres munkalapot Excelben, lépjen az `A1` cellára és Ctrl + V segítségével illessze be. Most még a három adatsor egy oszlopban van. Ezek szétválasztásához tegye a következőket: Adatok » Szövegből oszlopok, majd ◉ *Fix széles* Tovább Tovább Befejezés

a. Az $\widehat{a}_2, \widehat{a}_1, \widehat{a}_0$ együtthatók kiszámolásához jelölje ki a `D1:F1` cellatartományt, majd gépelje be a következőt: `=LIN.ILL(A1:A26;B1:C26)`, majd Ctrl + Shift + Enter. Ennek hatására az $\widehat{a}_2, \widehat{a}_1, \widehat{a}_0$ értékek rendre megjelennek a `D1`, `E1`, `F1` cellákban.

b. A `D2` cellába gépelje be a 102 értéket, az `E2` cellába a 648 értéket. Ekkor a tetőző vízállás becslése ezzel számolható ki: `=TREND(A1:A26;B1:C26;D2:E2)`. A kapott érték egészre kerekítve 800 cm.

2.1. táblázat

η	ξ
1,25	10,36
1,40	8,94
2,13	13,21
1,19	15,80
1,65	11,18
1,89	13,64
1,68	19,53
1,77	24,56
1,28	11,48
1,16	7,77
0,94	11,30
3,69	28,13
3,51	30,18
3,14	23,14
1,22	15,88
2,29	19,76
4,42	35,36
2,90	25,40

2.2. táblázat

η	ξ_1	ξ_2
590	58	405
660	52	450
780	133	350
770	179	285
710	96	330
640	72	400
670	72	550
520	43	480
660	62	450
690	67	610
500	64	380
460	33	460
610	57	425
710	62	560
620	54	420
660	48	620
620	86	390
590	74	350
740	95	570
730	44	710
720	53	700
720	77	580
640	46	700
805	123	560
510	26	370
673	62	430

Irodalomjegyzék

[1] Tómács Tibor: *Valószínűségszámítás*
https://tomacstibor.uni-eszterhazy.hu/tananyagok/Valoszinusegszamitas.pdf

[2] Tómács Tibor: *Matematikai statisztika gyakorlatok*
https://tomacstibor.uni-eszterhazy.hu/tananyagok/Matematikai_statisztika_gyakorlatok.pdf

[3] Tómács Tibor: *Matematikai statisztika*
https://tomacstibor.uni-eszterhazy.hu/tananyagok/Matematikai_statisztika.pdf