Kidolgozott feladatok az ANOVA témaköréből

A tájékozódást mindenféle színkódok segítik. A feladatok eredeti szövege zöld, a megoldások fekete, a figyelmeztető, magyarázó elemek piros színűek.

1. Nyuszika az erdő szélén új répatorta-gyárat szeretne üzembe helyezni. 3 különböző kiváló receptet is ismer, így kísérletesen kívánja megállapítani, hogy melyikkel legnagyobb a műszakonkénti termelékenysége. Mindhárom recepttel 3-3 napon át gyártja a tortákat, minden nap 3-3 műszakban dolgoznak szorgos munkásai, és feljegyzik a műszakonként előállított torták számát. Egy nap csak egyféle recept szerint megy a gyártás, és a 3 műszak 3 ismételt gyártási kísérletnek tekinthető.
Az alábbi ANOVA táblában Nyuszika kiszámította az eltérés-négyzetösszegeket:

hatás	SS	szabadsági fok	szórásnégyzet	F
recept	324,52	2	162,26	7,69
nap	126,67	6	21,11	2,86
error	132,67	18	7,37	

a) Írjuk föl a modellt, eldöntve, hogy mely faktor véletlen, melyik rögzített, és milyen a terv szerkezete! Milyen nullhipotéziseket vizsgálhatunk itt?

A faktorok (ahogy az ANOVA táblából is látszik) a recept és a nap, a műszakok ismétlésnek számítanak (egy műszakhoz egy adat tartozik). Mivel egy nap csak egy receptet használ, így a nap a receptbe van ágyazva.
El kell dönteni, hogy a faktorok rögzített vagy véletlen természetűek. A nap véletlen faktor, mivel 1) beágyazott faktor, 2) a nap (és a hozzá hasonló „természeti háttérre" vonatkozó idő-típusú faktorok) mindig véletlen faktor. A recept rögzített faktor, mivel ki akarja választani a legjobbat.
A modell:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

A nap véletlen faktor, tehát: $\beta_{j(i)} \sim N\left(0, \sigma_B^2\right)$
A nullhipotézisek:

$$\alpha_i = 0$$
$$\sigma_B^2 = 0$$

Az ANOVA táblából sejthető volt, hogy a modell hierarchikus, de csupán erre hivatkozni nem elegendő. Lehetett volna az is, hogy ez egy kereszt-osztályozás típusú modell, csak a kölcsönhatást bevonták az errorba.

b) Írjuk be a táblázatba a hiányzó értékeket! Hogy döntünk az egyes nullhipotézisekről?

A számokat fentebb lehet látni a táblázatban (feketével).
A recept szabadsági foka $3-1=2$.
A nap szabadsági foka $3\cdot(3-1)=6$
Az error szabadsági foka $3\cdot 3\cdot(3-1)=18$
Az összes faktor szabadsági fokait ha összeadjuk, eggyel kevesebbet kell kapnunk, mint az összes kísérlet száma. Ezt célszerű mindig ellenőrizni. Itt az összes kísérlet száma $3\cdot 3\cdot 3=27$ a számolt szabadsági fokok összege 26, tehát ilyen szempontból nyugodtak lehetünk.
A szórásnégyzetet (MS) úgy kapjuk, hogy az eltérés-négyzetösszeget (SS) elosztjuk a szabadsági fokkal.

Az F próbastatisztikák számlálójában mindig annak a faktornak a szórásnégyzete szerepel, amire a próbastatisztika vonatkozik. Egy ilyen hierarchikus modellben a beágyazott faktorhoz tartozó próbastatisztika nevezőjébe az error szórásnégyzet, a nem beágyazott faktorhoz tartozó próbastatisztika nevezőjébe pedig a beágyazott faktor szórásnégyzete kerül.

$$F^{A}=\frac{162{,}26}{21{,}11}=7{,}69$$

$$F^{B(A)}=\frac{21{,}11}{7{,}37}=2{,}86$$

A recepthez tartozó próbastatisztika szabadsági fokai 2 és 6, az $\alpha=0{,}05$-höz tartozó kritikus érték: 5,14, ennél nagyobb a próbastatisztika értéke, így az első nullhipotézist elutasítjuk.
A naphoz tartozó próbastatisztika szabadsági fokai 6 és 18, az $\alpha=0{,}05$-höz tartozó kritikus érték: 2,66, ennél nagyobb a próbastatisztika értéke, így a második nullhipotézist elutasítjuk.
Bár nem volt feladat, adhatunk szakmai választ is. Tehát szignifikáns különbséget sikerült kimutatni a receptek között, illetve a napok pluszingadozást okoznak a műszakokhoz képest.

c) Ha Nyuszika kiválasztja a legnagyobb termelékenységű receptet, mekkora ingadozásra számíthat a gyártás során?

Nyuszika $\sigma_B^2+\sigma_e^2$ nagyságú ingadozásra számíthat. Ezek értékét csak becsülni tudjuk, hiszen ezek a sokaság paraméterei.
A σ_e^2 becslése az error szórásnégyzet: $\hat{\sigma}_e^2=7{,}37$.
A σ_B^2 becslését onnan kaphatjuk meg, hogy $EMS(B)=3\sigma_B^2+\sigma_e^2$.
Ez általában úgy néz ki, hogy $EMS(B)=p\sigma_B^2+\sigma_e^2$, ahol p a cellánkénti ismétlések száma, egy beágyazott faktor esetén. Mindezt általánosságban, bármilyen modellre a csillagos-négyszöges táblázatból kaphatjuk meg.
Innen $\hat{\sigma}_B^2=\frac{s_B^2-s_R^2}{3}=\frac{21{,}11-7{,}37}{3}=4{,}58$.
Így a teljes ingadozás becslése: $\hat{\sigma}_B^2+\hat{\sigma}_e^2=4{,}58+7{,}37=11{,}95$

2. Bergengóciában egy ásatás során egy kőtáblát találnak, amin egy ANOVA tábla látható. Sajnos a szöveg és a számok java része nem olvasható (Ahol üres helyet láthatunk az alábbi táblázatban, ott olvashatatlan a szöveg).

hatás	SS	szabadsági fok	szórásnégyzet	F	p
A faktor	420	3	140	3,50	0,031
B faktor	86	2	43	1,075	0,356
AB kölcsönhat.	333	6	55,5	1,388	0,259
error	960	24	40		

a) Ezek alapján el tudjuk-e dönteni, hogy a vizsgált faktorok véletlen vagy rögzített faktorok voltak-e? Miért?

Azt kell megvizsgálni, hogy az A faktor próbastatisztikáját hogyan kapták meg. Ugyanis a nevezőben csak akkor szerepelhet az error szórásnégyzet, ha mindkettő faktor rögzített. Ha a két faktor közül bármelyik (tehát akár az A, akár a B, de lehet mindkettő is) véletlen faktor, akkor mindkettőhöz tartozó próbastatisztika nevezőjébe a kölcsönhatás szórásnégyzete kell, hogy kerüljön (tehát a rögzített faktoréba is!).

$\frac{140}{55{,}5} = 2{,}52$; $\frac{140}{40} = 3{,}5$, tehát mindkét faktor rögzített.

b) Írjuk föl a modellt és a nullhipotéziseket!

Két faktor szerinti keresztosztályozásról van szó.

A modell:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

A nullhipotézisek:

$$\alpha_i = 0$$
$$\beta_j = 0$$
$$\alpha\beta_{ij} = 0$$

Mindig annyi nullhipotézisnek kell lennie, ahány statisztikai próbát végzünk (azaz ahány F, illetve p-érték van az ANOVA-táblában).

c) Határozzuk meg, hogy a hiányzó helyeken milyen értékek állhattak egykoron!

Az error szabadsági fokát az eltérés-négyzetösszeg (SS) és a szórásnégyzet (MS) hányadosaként kapjuk: $\frac{960}{40} = 24$

A kölcsönhatás szabadsági fokát a két főhatás szabadsági fokának szorzataként kapjuk.
Két SS hiányzik, ezeket a megfelelő MS és szabadsági fok szorzataként kapjuk meg.
A B faktor szórásnégyzetét az SS és szabadsági fok hányadosaként kapjuk.
Mint azt már korábban megállapítottuk, minden F-próbastatisztika nevezőjében az error szórásnégyzet szerepel, így ezzel osztjuk le a szórásnégyzeteket:

$\frac{43}{40} = 1{,}075$; $\frac{55{,}5}{40} = 1{,}388$

d) Ezek alapján állíthatjuk-e, hogy az A faktornak nincs hatása?

Az A faktorhoz tartozó p-érték 0,031 < 0,05, azaz a nullhipotézist elutasítjuk, tehát ezek alapján az A faktornak van hatása, így nyilván nem állíthatjuk, hogy nincs.

Ez a feladat véletlenül lett ilyen „duplán trükkös”, eredetileg olyan faktort szerettem volna választani, ahol a nullhipotézist elfogadjuk. Ilyen majd még lesz a második feladatsorban.

3. Az öreg király közhírré tétette, annak adja a lánya kezét, aki olyan abrakot tud készíteni varázslatos telivérének, amitől az gyorsabban fog futni, mint bármely ló a vidéken. A leleményes királyfi 5 féle abrakot készített, és ezek közül ki akarta választani, hogy melyikkel futnak a lovak a leggyorsabban. Ehhez kísérleteket végzett. Elment a helyi ménesbe, és itt a lovakat az abrakokkal etette, és mérte, hogy milyen gyorsan futnak.

a) Milyen modellt és nullhipotézist használt a királyfi?

Alapvetően egyetlen faktor van biztosan a modellben, mégpedig az abrak. Ha csak ezt a faktort vesszük be a modellbe, akkor ez egy egy faktor szerinti ANOVA.

A modell:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

A nullhipotézis:

$$\alpha_i = 0$$

A faktor rögzített, mert az abrakok közül a legjobbat szerette volna kiválasztani.

Mivel nem volt teljesen egyértelműsítve, hogy a királyfi a kísérleteket hogyan végezte, így több másik modell is elképzelhető.

Például ha egy abrakkal több lovat is etetett, de ezeknek több futását is megmérte, akkor például egy hierarchikus modell írható fel, ahol az abrak faktorba van a lovak faktor beágyazva (ekkor ez véletlen faktor), és a futások adják az ismétlést.
Ha viszont a rendelkezésre álló lovakat mind az 5 abrakkal kipróbálta, akkor ez a lovak és az abrakok közötti keresztosztályozás lesz. (Az abrakok még mindig rögzített faktor, a lovak pedig véletlen.)
Természetesen bármilyen mesét és hozzá passzoló modellt+nullhipotéziseket maximális pontszámmal fogadtam el.

b) Mikor követ el másodfajú hibát? Milyen következményei lehetnek, ha másodfajú hibát vét?

Másodfajú hibát akkor követ el, ha elfogadja a nullhipotézist, ami viszont hamis. Azaz jelen helyzetben úgy találja, hogy az abrakok között nincs különbség, viszont valójában lenne közöttük egy legjobb (vagy több jobb, de mindenféleképp lenne reményteli abrak).
Emiatt valószínűleg nem próbálna benevezni a versenybe, pedig lehetséges, hogy nála lett volna a nyerő abrak. Azaz lehet, hogy lett volna esélye a királylány kezére, de nem is próbálkozott.
A fordított válasz az elsőfajú hibához tartozik, és így természetesen nem jó válasz a kérdésre. Ez az lenne, hogy azt hiszi, hogy megtalálja a legjobb abrakot, miközben az egy teljesen hétköznapi abrak. Ekkor benevez a versenyre, és nagy valószínűséggel elbukja azt, és így nem nyeri el a királylány kezét.
Ennek a következménye, hogy önmagában azt a következtetést, hogy „nem nyeri el a királylány kezét" nem tudtam elfogadni, hiszen az elsőfajú hiba is ugyanerre a végkövetkeztetésre jutna. (Mellesleg: A királyfi a királylány kezére pályázik, de hibát követ el. Ebből mindenféle statisztikai megfontolás nélkül is meg lehet sejteni, hogy nem fogja elnyerni a királylány kezét. Tehát ezt nem tudtam értékelni.)

c) A királyfi tudta előre, hogy a két legjobban teljesítő abrakot külön össze fogja hasonlítani t-próbával. Tervezett vagy post-hoc összehasonlítást végez-e ilyenkor?

Tervezett összehasonlítást akkor végzünk, ha már az eredmények megszerzése előtt is tudjuk, hogy melyik kettő csoportot fogjuk összehasonlítani. Azt nem tudjuk az eredmények megszerzése előtt, hogy melyik az a két csoport (abrak), ami a két legjobb eredményt adja. Így ez post-hoc összehasonlítás.
Sajnos kimaradt a feladatból, hogy megkérdezzem, hogy miért az a jó válasz. Így sokan egyáltalán nem írtak indoklást, amit viszont így maximum ponttal kellett elfogadnom. Nem meglepő módon a második feladatsorban visszatér ez a kérdés, és ott már nem követtem el ezt a hibát.

4. Okoska Törp tudja, hogy az ANOVA feltétele a hibák függetlensége, és ezt sokszor hangoztatja is. Nem tudja viszont a további kérdésekre a választ. Segítsünk neki!

a) Mitől kell a hibáknak függetlennek lenniük?

A hibáknak egymástól kell függetlennek lenniük. (Csoporton belül és csoportok között is.)

b) Hogyan vizsgáljuk, hogy ez ténylegesen teljesül-e?

A reziduumokat a mérések sorrendjében kell ábrázolni. Ha ebben bármilyen tendenciát vélünk felfedezni, akkor abból arra következtethetünk, hogy nem teljesül a függetlenség.
Nem jó válasz az, ha valaki csak annyit írt, hogy „reziduumábrával", mert vannak más típusú reziduumábrák, amiken ezt nem lehet vizsgálni.

c) Mit tehetünk, ha az adódik, hogy ez mégsem teljesül?

Sok mindent nem lehet tenni. Talán a legjobb válasz az, hogy újra elvégezzük a méréseket úgy, hogy ügyelünk arra, hogy teljesüljön a függetlenség. Szintén elfogadható válasz az, hogy megpróbáljuk kompenzálni a hatást, vagy a kiértékelés során figyelmeztetjük magunkat, illetve a megbízónkat, hogy az eredmények nem feltétlenül megfelelőek, mert valószínűleg nem teljesül a függetenség.

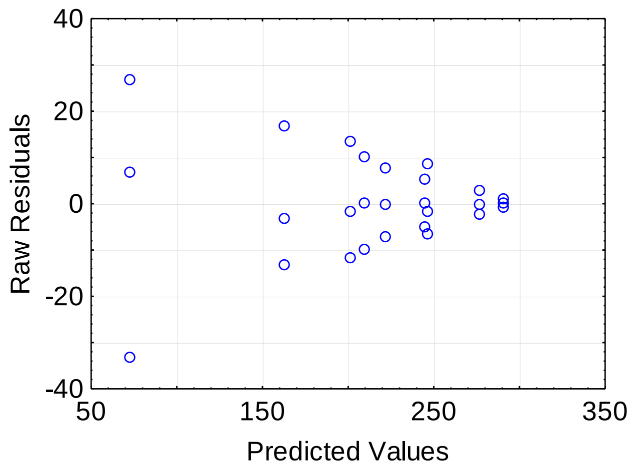
Rossz válasz az, hogy randomizáljuk az adatokat. Ennek nyilvánvalóan semmi értelme, ugyanis a randomizálásnak a mérések elvégzése előtt van értelme.

d) Mit és mikor tehetünk annak érdekében, hogy ez teljesüljön?

Randomizálni kell a méréseket. (Azaz a méréseket véletlenszerű sorrendben kell elvégezni.) Az előbbi mondatban már benne van, hogy mindezt a mérések elvégzése előtt lehet megtenni.

Voltam annyira kedves, hogy ha valaki a c) feladathoz leírta azt, hogy a méréseket véletlenszerű sorrendben kell elvégezni (azaz a mondatból egyértelműen kiderül az, hogy a randomizálást a mérések előtt lehet megtenni), a d)-hez viszont nem írt semmit, akkor a c)-re természetesen nem adtam pontot, de a d) feladatra megadtam a pontot.

5. A Bölcs Bagoly ANOVA vizsgálatot végzett, a reziduumokat a mért érték függvényében ábrázolta, így kapta az alábbi ábrát.



a) Milyen következtetést vonhatunk le ebből az ábrából?

Ha a reziduumokat a becsült érték függvényében ábrázoljuk, akkor mindig a hiba varianciájának konstans voltát vizsgáljuk.

Ezen az ábrán azt láthatjuk, hogy a hiba varianciája monoton változik a becsült értékkel, nagyobb becsült értékeknél kisebb a variancia. (Tehát nem konstans.)

b) Mit tanácsoljunk neki, hogyan folytassa tovább az elemzést?

Ha a variancia úgy nem konstans, hogy a becsült értékkel (várható értékkel) monoton változik, akkor van remény arra, hogy Box-Cox transzformációval a variancia konstanssá tehető. Tehát a Bölcs Bagolynak Box-Cox transzformációt kellene végeznie az adataira.

Fontos, hogy általában ha nem konstans a hiba varianciája, a Box-Cox transzformáció nem biztos, hogy segít. Ehhez az kell, hogy a variancia a várható értékkel monoton változzon. (Mint ahogy itt most van.) Természetesen erre a jelenségre is lesz egy feladat, ami felhívja a figyelmet a második feladatsorban.

2. feladatsor

1. Egy vállalatnak nagy mennyiségű nyúlszőrre van szüksége, a nyulak szőrét pedig a növekedésük során ápolni kell. A vállalat 3 potenciális készítmény közül szeretné kiválasztani, hogy melyik az, amivel a legjobban nő a nyulak szőre. A 3 készítményhez kisorsolnak 4-4 nyulat, amelyek szőrét a növekedésük során végig az adott készítménnyel kezelik. A kísérletet fél éven keresztül folytatták, és mérték, hogy mennyi szőrt lehet az állatokról lenyírni. Az adatokat alább láthatjuk szabványos nyúlszőr-egységekben.

	1. készítmény	2. készítmény	3. készítmény
	77	91	68
	89	83	73
	72	96	82
	83	90	77
eltérés-négyzetösszeg (SS):	162,75	86,00	106,00

a) Miért fontos, hogy véletlenszerűen sorsolják ki, hogy melyik nyulat melyik készítménnyel kezelik?

Ez a kérdés a randomizálás fontosságára, céljára kérdez rá.

Azért fontos, hogy biztosítsuk az adatok függetlenségét.

Ha nem randomizálunk, és például az 1. készítményhez olyan nyulakat választunk ki, amiknek a szőre alapvetően nagyobb hozammal növekszik, akkor nem fogjuk tudni megállapítani, hogy a nyulak szőre a kezelés hatására növekszik jobban (azaz az a kezelés jobb, mint a többi), vagy csak a nyulak jobbak önmagukban. Másik oldalról megközelítve a dolgot, így „csalni" tudunk a kísérletekkel, és olyan készítményt is hatásosnak tudunk beállítani, ami valójában nem is az.

b) Írjuk fel a modellt és a nullhipotézist!

Ez egy egyszerű egyfaktoros ANOVA, a nyulak ismétléseknek tekinthetők. A kezelés faktor rögzített, hiszen ki akarjuk választani közülük a legjobbat.

A modell:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

A nullhipotézis:

$$\alpha_i = 0$$

c) Végezzük el az elemzést!

Az ANOVA táblát kell felírni.

hatás	SS	szabadsági fok	szórásnégyzet	F
készítmény	463,50	2	231,75	5,879
error	354,75	9	39,42	

Az errorhoz tartozó SS-t a megadott csoportokon belüli SS-ek összeadásával kapjuk meg.

A csoportokon belüli SS (ezt egyébként a megadott adatokból ki lehetett volna számolni, csak a számítás megkönnyítése végett adtam meg):

$$\sum_j \left(y_{ij} - y_{i\bullet}\right)^2$$

Az error SS:

$$\sum_i \sum_j \left(y_{ij} - y_{i\bullet}\right)^2$$

A hatás SS-ének számításához ki kell számolni a csoportátlagokat és az összes adat átlagát:

$$y_{1\bullet} = 80{,}25 \qquad y_{2\bullet} = 90{,}00 \qquad y_{3\bullet} = 75{,}00 \qquad y_{\bullet\bullet} = 81{,}75$$

A hatás SS-e:

$$SS_A = 4\left(y_{1\bullet} - y_{\bullet\bullet}\right)^2 + 4\left(y_{2\bullet} - y_{\bullet\bullet}\right)^2 + 4\left(y_{3\bullet} - y_{\bullet\bullet}\right)^2 = 4 \cdot 2{,}25 + 4 \cdot 68{,}06 + 4 \cdot 45{,}56 = 463{,}50$$

A készítmény szabadsági foka 2, mert 3 készítmény volt. Az error szabadsági foka $3 \cdot (4-1) = 9$, mert 3 csoportban 4-4 nyulat kezeltek.

Érdemes ellenőrizni a szabadsági fokokat, az összes szabadsági fok összegének 1-gyel kevesebbnek kell lennie, mint ahány kísérletet végeztek (jelenleg a nyulak száma). 2 + 9 = 11, 12 nyúl volt, tehát ez így jó.

A szórásnégyzeteket osztással kapjuk, az F próbastatisztikát pedig az MS-ek hányadosaként.
A próbastatisztika szabadsági fokai 2 és 9, az $\alpha = 0{,}05$-höz tartozó kritikus érték: 4,26, ennél nagyobb a próbastatisztika értéke (5,879), így a nullhipotézist elutasítjuk.
Így a készítményekkel elérhető várható nyúlszőr-hozamok különböznek.

d) Végezzünk páronkénti összehasonlítást az 1. és 2. készítményre!

Kétmintás t-próbát kell végezni úgy, hogy a nevezőbe az error MS négyzetgyökét kell írni. (σ_e^2 becslése).

$$H_0: \quad \mu_1 = \mu_2$$

$$t_0 = \frac{80{,}25 - 90{,}00}{\sqrt{39{,}42}\sqrt{\frac{1}{4} + \frac{1}{4}}} = \frac{-9{,}75}{6{,}28 \cdot 0{,}71} = -2{,}196$$

Arra kell figyelni, hogy a gyök alá a nevezőbe mindig a csoportok elemszámát kell írni, akármiből is számítjuk a nevezőben lévő szórást.

Mivel az error MS szabadsági foka 9, így a használatos t-próba szabadsági foka is 9.
Az $\alpha = 0{,}05$-höz tartozó kétoldali kritikus érték: ±2,262. A kapott próbastatisztika értéke ezek között van, tehát a nullhipotézist elfogadjuk.
Tehát az adatok nem mondanak ellent annak, hogy az 1. és 2. készítménnyel kezelt nyulak esetén a várható szőrhozam megegyezik.

2. Az adóhatóság 3 kisvállalkozás bevételeit nézi heti bontásban egy vizsgálat során. Összesen 12 hetet vizsgáltak, 3 hónapon keresztül (annak ellenére, hogy a hetek átívelhetnek a hónapokon, itt az első 4 hetet tekintjük az első hónapnak, a második 4-et a másodiknak, a harmadik 4-et pedig a harmadiknak). Azt kívánták megállapítani, hogy a 3 vállalkozás közül melyiknek a legnagyobb a heti várható bevétele.
Az alábbi ANOVA táblában az eltérés-négyzetösszegek már megjelentek:

hatás	SS	szabadsági fok	szórásnégyzet	F
vállalkozás	168	2	84	3,5
hónap	348	2	174	7,25
vállalk.*hónap	96	4	24	2
error	324	27	12	

a) Írjuk fel a modellt és a nullhipotéziseket!

Két faktort vizsgálunk a feladatban, a hónapokat és a vállalkozásokat, a hetek ismétléseknek tekinthetők, hiszen egy héthez egyetlen adat tartozik.
A vállalkozás rögzített faktornak tekinthető, hiszen ki akarták választani, hogy melyiknek a legnagyobb a várható bevétele. A hónap időjellegű faktor, így véletlen faktor.
A vállalkozás és a hónap keresztosztályozás típusú viszonyban vannak, mert mindegyik vállalkozás esetén ugyanazt a 3-3 hónapot vizsgáljuk.

A modell:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

A hónap véletlen faktor, tehát: $\beta_j \sim N\left(0, \sigma_B^2\right)$ és $\alpha\beta_{ij} \sim N\left(0, \sigma_{AB}^2\right)$

A nullhipotézisek:

$$\alpha_i = 0$$

$$\sigma_B^2 = 0$$

$$\sigma_{AB}^2 = 0$$

b) Töltsük ki a hiányzó cellákat, és döntsünk a nullhipotézisekről!

A vállalkozás szabadsági foka 2, mert 3 vállalkozás volt. A hónap szabadsági foka 2, mert 3 hónapot vizsgáltak. A kölcsönhatás szabadsági foka $2 \cdot 2 = 4$. Az error szabadsági foka $3 \cdot 3 \cdot (4-1) = 27$, mert a 3 vállalkozás 3 hónapjában 4-4 hét volt.

Érdemes ellenőrizni a szabadsági fokokat, az összes szabadsági fok összegének 1-gyel kevesebbnek kell lennie, mint ahány kísérletet végeztek. 2 + 2 + 4 + 27 = 35, 36 adatunk van, tehát ez így jó.

A szórásnégyzetek osztással számolandók ki.
Mivel van véletlen faktor, így mindkét főhatásnál a próbastatisztika nevezőjébe a kölcsönhatás szórásnégyzetét kell írni. A kölcsönhatásba természetesen az errort.

$$F^A = \frac{84}{24} = 3{,}5$$

$$F^B = \frac{174}{24} = 7{,}25$$

$$F^{AB} = \frac{24}{12} = 2$$

A vállalkozáshoz tartozó próbastatisztika szabadsági fokai 2 és 4, az $\alpha = 0{,}05$-höz tartozó kritikus érték: 6,94, ennél kisebb a próbastatisztika értéke, így az első nullhipotézist elfogadjuk.
A hónaphoz tartozó próbastatisztika szabadsági fokai 2 és 4, az $\alpha = 0{,}05$-höz tartozó kritikus érték: 6,94, ennél nagyobb a próbastatisztika értéke, így a második nullhipotézist elutasítjuk.
A kölcsönhatáshoz tartozó próbastatisztika szabadsági fokai 4 és 27, az $\alpha = 0{,}05$-höz tartozó kritikus érték: 2,73, ennél kisebb a próbastatisztika értéke, így a harmadik nullhipotézist elfogadjuk.
Bár nem volt feladat, adhatunk szakmai választ is. Az adatok alapján nem sikerült különbséget kimutatni a vállalatok heti várható bevétele között, a hónapok pluszingadozást okoznak, a kölcsönhatásról pedig nem sikerült kimutatni, hogy pluszingadozást okozna.

c) A hatóság előre tudta, hogy végül a két legnagyobb átlagos bevételt hozó vállalkozást külön össze fogja hasonlítani. Milyen típusú összehasonlítás ez? Miért?

Mivel előre nem lehet tudni, hogy melyik két vállalkozás lesz az, amelyiknek a legnagyobb lesz az átlagos bevétele, így ez post-hoc összehasonlítás.
Megjegyzendő, hogy mivel nem sikerült különbséget kimutatni, így nem túlságosan ésszerű páronkénti összehasonlítást végezni itt.

3. Az ANOVA feltétele a normális eloszlás. Ez a megfogalmazás nem elég pontos. A következő ezzel kapcsolatos állításokról döntsük el, hogy igazak-e vagy hamisak-e! Minden esetben indokoljuk a döntésünket!

a) Az ANOVA elvégzése során feltételezzük, hogy az adatok egyazon normális eloszlásból származnak.

Hamis. A hibákról feltételezzük a normális eloszlást. Ha például egy rögzített faktornak hatása van, akkor ott a csoportok várható értéke különböző lesz, így az adatok máris nem egyazon normális eloszlásból fognak származni.

b) Ha a reziduumok normális eloszlást követnek, akkor az adatok biztosan egyazon normális eloszlásból származnak.

Hamis. Az indok ugyanaz, mint az előző kérdésnél.

c) Nem tudjuk elvégezni az ANOVA vizsgálatot, azaz a számítás nem kivitelezhető, ha a hibák nem normális eloszlásúak.

Hamis. A számítás elvégzésénél egyszerűen az alapműveleteket végezzük el, semmilyen eloszlással kapcsolatos feltételezés nem befolyásolja ezeknek az elvégezhetőségét.

d) ANOVA vizsgálat során elképzelhető, hogy az adatok egyazon normális eloszlásból származnak, viszont a hibák nem.

Hamis. Ahhoz, hogy az összes adat normális eloszlásból származzon, az kell, hogy egyetlen rögzített faktornak se legyen hatása, és a véletlen faktorokra és a hibákra is teljesüljön, hogy normális eloszlásból származnak.

e) A normalitás vizsgálatakor a reziduumokat ún. Gauss-hálós elrendezésben ábrázoljuk.

Igaz. A normalitás vizsgálatakor a reziduumokat és nem a hibákat ábrázoljuk, hiszen azok értékét nem ismerhetjük.
Ha valaki indoklás nélkül „tippelt”, akkor is kapott pontot. Ennek a következménye, hogy ha valaki eltalálta, hogy igaz vagy hamis az állítás, de rossz indoklást írt, akkor is megkapta ezt a pontot. Mivel a d) és e) kérdésnél a válasz úgymond triviális, így itt egyrészt bármilyen nem értelmetlen indoklást elfogadtam, másrészt itt a jó tipp 2-2 pontot ért, a másik 3 kérdésnél csak 1-1-et a 4-ből.

4. Egy földmintában szeretnék meghatározni a nehézfémtartalmat. A nyomozó hatóság 3 különböző laboratóriumba is elküldte a földminta 1-1 részét. Mindhárom laborban 4-4 párhuzamos mintaelőkészítést végeztek, majd minden mintából 2-2 elemzést végeztek.

   a) Írjuk föl a modellt és a nullhipotéziseket!

Két faktorunk van, a labor és a mintaelőkészítés. Mivel a laborokban külön-külön mintaelőkészítéseket végeztek, ezért a mintaelőkészítés szintjei csak a laborokon belül értelmezhetők, azaz a mintaelőkészítés faktor a labor faktorba van ágyazva.
A mintaelőkészítés véletlen faktor egyrészt mert beágyazott faktor, másrészt nyilván nincs értelme azt kérdezni, hogy melyik mintaelőkészítés a legjobb.
A labor faktorra nem dönthető el teljesen egyértelműen, hogy rögzített vagy véletlen.
Véletlen faktornak tekinthető akkor, ha a hatóság azért küldte el több laborba a mintát, mert több szakértő véleményét is ki szerette volna kérni, hogy pontosabb eredményt kapjon.
Rögzített faktornak tekinthető, ha azért küldték több helyre, mert a laborokat szerették volna összehasonlítani, majd ezek közül kiválasztani azt, hogy a vizsgálatokat később melyik laborral fogják elvégeztetni.
Ez a második igazából végtelenül erőltetett elképzelés, és mindenkinek éreznie kellene, hogy az első megoldás a jó, de ha valaki ezt így megindokolta, akkor azt is elfogadtam.
A modell:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

Ha a labor rögzített faktor, a nullhipotézis:

$$\alpha_i = 0$$

Ha véletlen faktor, $\alpha_i \sim N(0, \sigma_A^2)$, a nullhipotézis:

$$\sigma_A^2 = 0$$

A mintaelőkészítés véletlen faktor, tehát: $\beta_{j(i)} \sim N(0, \sigma_B^2)$, a nullhipotézis:

$$\sigma_B^2 = 0$$

   b) Adjuk meg, hogyan számoljuk ki az egyes próbastatisztikákat! Milyen szabadsági fokú F-próbát végzünk itt? (Az F eloszlásnak 2 szabadsági foka van!)

Szerencsére a kérdésre adott választ nem befolyásolja, hogy a labor rögzített vagy véletlen faktor.
A labor szabadsági foka 2, mert 3 labor volt. A mintaelőkészítés szabadsági foka $3 \cdot (4-1) = 9$, mert 3 laborban 4-4 mintaelőkészítést végeztek. Az error szabadsági foka $3 \cdot 4 \cdot (2-1) = 12$, mert az összesen $3 \cdot 4 = 12$ mintából 2-2 ismételt mérést végeztek.
A laborhoz tartozó próba számlálójában a laborhoz tartozó szórásnégyzet, a nevezőjében a beágyazott faktor, azaz a mintaelőkészítés szórásnégyzete van. Az F-próba szabadsági fokai így 2 és 9.

$$F^{\text{labor}}(2;9) = \frac{MS_{\text{labor}}}{MS_{\text{minta}}}$$

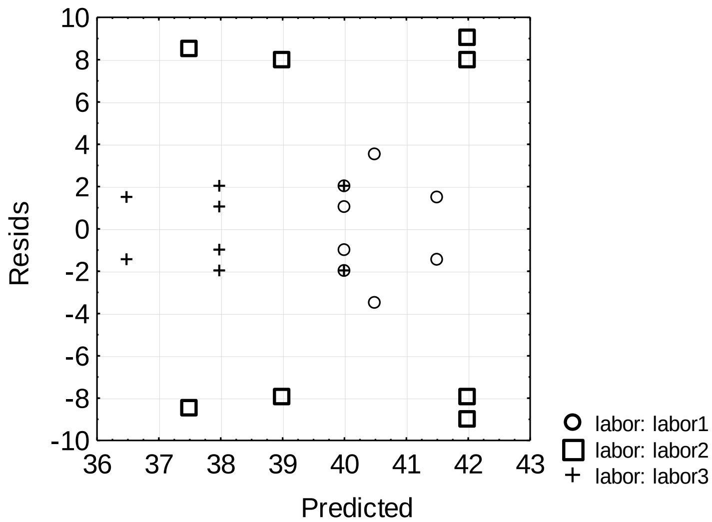
A mintaelőkészítéshez tartozó próba számlálójában a mintaelőkészítéshez tartozó szórásnégyzet, a nevezőjében az error szórásnégyzet van. Az F-próba szabadsági fokai így 9 és 12.

$$F^{\text{minta}}(9;12) = \frac{MS_{\text{minta}}}{MS_{\text{error}}}$$

   c) Az elemzés során azt találják, hogy a laborokra vonatkozó nullhipotézisre p = 0,131. Mondhatjuk-e ez alapján, hogy a labornak nincs hatása? Miért?

Ha p = 0,131, akkor a nullhipotézist elfogadjuk. Ekkor nem állíthatjuk azt, hogy a labornak nincs hatása, csupán azt, hogy az adott szignifikanciaszinten nem tudtunk hatást kimutatni.
Itt arra kell gondolni, hogy csak a nullhipotézis elutasítása bizonyító erejű.
Itt az előző feladatsor 2 d) kérdésével ellentétben sikerült olyan p értéket megadnom, amire a nullhipotézist elfogadjuk.

Az alábbi ábrán a reziduumokat a becsült értékek függvényében, a laborok szerint csoportosítva ábrázoltuk.



d) Mire következtethetünk ebből?

Amikor a reziduumokat a becsült érték függvényében ábrázoljuk, akkor mindig a hiba varianciájáról tudunk következtetést levonni.
Látható, hogy a 2. labor esetén sokkal jobban szórnak a pontok a 0 körül, azaz ebben a laborban nagyobb a hiba varianciája, mint a másik kettőben. Azaz az ANOVA-nak az a feltétele, hogy a hibák σ_e^2 varianciája konstans nem teljesül.

e) Milyen lépéseket kellene tenni ennek tudatában?

Látható, hogy nem az a helyzet, hogy a variancia a várható értékkel függ össze, hiszen mindhárom labor ugyanolyan értéktartományban mért, tehát egyáltalán nem reménykedhetünk, hogy az adatok Box-Cox transzformációjával a variancia konstanssá tehető.
Ki kell hagyni az elemzésből a 2. labor eredményeit, hiszen sokkal kevésbé precízen (nagyobb ismételhetőségi varianciával) tudják a nehézfémtartalmat mérni.
Mellesleg egyáltalán nem irreális ez a helyzet, hogy bizonyos laborok kevésbé precízen, mások meg precízebben tudnak mérni.