Varianciaanalízis

Varianciaanalízis

Szinomimák: varianciaelemzés, lineáris regresszió elemzés, analysis of variance, ANOVA, multiple regression analysis
Kérdések: (1) van-e különbség a csoportok között (t-próba általánosítása), (2) van-e hatása a vizsgált tényezőnek (regressziószámítás: magyarázó változók hatása a függő változóra).

- **Egy- vs. többtényezős**: ha egy független változó van, egytényezős, ha n, n-tényezős.
- **Független mintás vs. ismételt méréses**: ha az adatok különböző elemeken végzett mérésekből származnak (pl. magyar, cseh és angol beszélők), független mintánk van, ha ha egyazon adatközlőtől többféle adat származik, ismételt méréses dizájnunk van.
- **Egy- vs. többváltozós**: a függő változók száma. Az ANOVÁ-ban alapértelmezetten egy függő változó van, a MANOVÁ-ban (multivariate ANOVA) legalább kettő.

Alkalmazási területek

- Egy adott kezelés különböző változatainak hatása a kontrollcsoporthoz képest (pl. magasabb dózis, alacsonyabb dózis, placebó).
- Többféle módszer hatékonysága egymáshoz és a kontrollcsoporthoz képest.
- Nominális független változók által kiváltott hatás (pl. különböző szemantikai kategóriák hatása a reakcióidőre).

Feltételek

- Egyes csoportokon belül normális eloszlás és
- azonos szórás (varianciák homogenitása),
- megfigyelések egymástól való függetlensége (szfericitás).

Normális eloszlás feltételének megsértését nem szokás sarkalatos problémának tekinteni, mert (1) 30 fölötti elemszám már természetszerűleg normális eloszlású, (2) 10–20 elemnél nem nagy az eltérés, (3) 10-nél kisebb elem esetén nincs igazán értelme eloszlásról beszélni.

Varianciák homogenitása és a megfigyelések egymástól való függetlensége (szfericitás) viszont alapvető, különben az eredmények nem megbízhatóak.

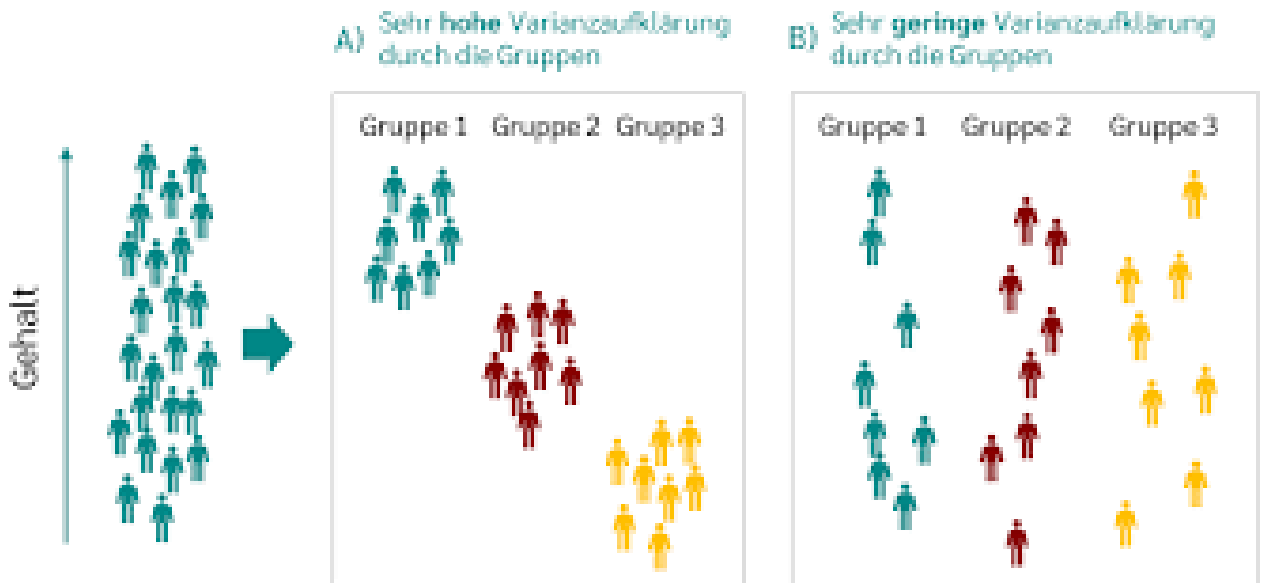
Egytényezős varianciaanalízis

Eljárás: az összes variancia felosztása a faktorok kombinációjából adódó csoportok **közötti** és a csoportokon **belüli** varianciára (innen az elnevezés).

1. csoporton belül: minden egyes csoport varianciája és átlaguk,
2. csoportok között: minden egyes csoport átlagának varianciája $\rightarrow$ véletlen hiba varianciabecslése = regressziószámítás reziduális varianciája,
3. döntés: ha a **csoportok közötti** variancia nagyobb, mint a **csoportokon belüli** variancia, akkor a tényezőnek (független változónak) van hatása.

Példa

Bal: az egyes csoportokon belül a variancia az egyes csoportok átlagához viszonyítva kicsi, a csoportátlagok egymás közötti varianciája viszont nagy. A csoporthoz tartozás magyarázó ereje tehát nagy.
Jobb: a csoportokon belüli variancia nagy, a csoportátlagok közötti variancia viszont kicsi. A csoporthoz tartozás magyarázó ereje kicsi.



Forrás: https://datatab.de/tutorial/varianzanalyse.

Varianciatábla

Variancia eredete *source*	Szabadsági fok *df*	Eltérés-négyzetösszeg *Sum Sq*	Átlagos eltérés-négyzetösszeg *Mean Sq*	F	p
Csoportok közötti *between*	$k-1$	SS_K	$MS_K = \frac{SS_K}{k-1}$	$F = \frac{MS_K}{MS_H}$	p
Csoporton belüli *within*	$k(n-1)$	SS_H	$MS_H = \frac{SS_H}{k(n-1)}$		
Teljes *total*	$nk-1$	SS_T	$MS_T = \frac{SS_T}{nk-1}$		

SS_H = reziduális hiba a regressziószámítás alapján.
Ha a kezelések **közötti** variancia, azaz az egyes csoportok (= kezelések, faktorszintek) átlagának varianciája szignifikánsan nagyobb, mint a csoportokon belüli, akkor a csoportok között különbséget tevő faktor hatása jelentős, azaz szignifikáns különbséget eredményez.

Példa

Reiczigel, Harnos & Solymosi, 316. o.: Tápoldat hatékonyságának tesztelése növények növekedésére. Eljárás: növények öntözése tömény, ill. híg tápoldattal, kontroll: víz. Kérdés: serkenthető-e a növények növekedése a tápoldat segítségével?

R-kód:

```
magassag = c(56,48,66,54,57,50,47,58,54,46,60,48)
tapoldat = rep(c("tomeny","hig","viz"),each=4)
novtap = data.frame(magassag,tapoldat)
```

`rep()`: tápoldat típusának ismétlése: opciók: times=4 (teljes sor ismétlése négyszer), each=4 (minden egyes elem ismétlése négyszer).
Fontos: az adatmátrixot a `data.frame()` paranccsal hozzuk létre, ami a *tapoldat* karaktervváltozókat faktorrá alakítja.

Varianciaelemzés az R-ben

Normális eloszlás tesztelése:
`tapply(novtap$magassag,novtap$tapoldat,shapiro.test)`

`tapply()`: függő változó kiszámítása független változó összes faktorszintjére a megadott függvény szerint, azaz `tapply(függővváltozó,függetlenváltozó(k),függvény)`.
Mindhárom csoport normális eloszlású.

Varianciák homogenitásának ellenőrzése:
`bartlett.test(novtap$magassag,novtap$tapoldat)`:
varianciák azonosak.

NB: Bartlett-próba kettőnél több próba összehasonlítására is alkalmazható, de csak normális eloszlás esetén $\leftrightarrow$ `var.test()` (F-próba) csak két mintát tud összehasonlítani.

Varianciaanalízis két függvény alapján:
`aov()`
`lm()`
Különbség: `aov()` csak azonos elemszámú cellák (kiegyensúlyozott elrendezés) esetén alkalmazható. Eltérő csoportelemszámok esetén `lm()` (indoklás Reiczigel et al., 375ff.).

`h = aov(novtap$magassag~novtap$tapoldat)`, vagy
`h = aov(magassag~tapoldat,data=novtap)`
`summary(h)`.
Táblázat elrendezése megegyezik a 6. diával.

Kapott F-érték az adott szabadságfokokra nem mutat szignifikáns eltérést a kezelések közötti és kezeléseken belüli átlagos eltérés-négyzetösszegek között $\Rightarrow$ tápoldat alkalmazása nincs hatással a növekedésre.

Igaz ez a víz és a tömény oldat összehasonlítására is?

Post hoc-tesztek

Probléma: az összehasonlítások nagy számával nő az α-hiba lehetősége, azaz annak a valószínűsége, hogy hibás szignifikáns p-értéket kapunk.

Módszerek:

- Páronkénti összehasonlítás t-próbákkal, majd a **Bonferroni-korrektúra** alkalmazása: szignifikancia-határ konfidenciaintervallum/összes lehetséges párosítás. Hátrány: nagy számú kombináció esetén nagyon nehéz szignifikáns különbséget kimutatni.
- **Tukey-féle** /tuːki/ post-hoc teszt: csak a független mintás varianciaanalízisre alkalmazható, az ismételt mérésesre nem.
- **Dunnett-próba**: általánosabb alkalmazhatóság.

Post hoc-tesztek

1. Tukey-féle post hoc-teszt bemenete az `aov()` kimeneteként kapott objektum:

```
h = aov(novtap$magassag~novtap$tapoldat)
TukeyHSD(h)
```

Egyik párosítás sem különbözik szignifikánsan.

2. *t*-próba Bonferroni-korrektúrával
Pl. víz és tömény oldat összehasonlítása. Lehetséges kombinációk száma 3, tehát a konfidencia-intervallum határa Bonferroni-korrektúra után $0,005/3 = 0,0167$.

```
hig = novtap$tapoldat == "hig"
t.test(novtap$magassag[!hig]~novtap$tapoldat[!hig]
```

p = 0.4462, azaz a különbség messze nem szigifikáns.

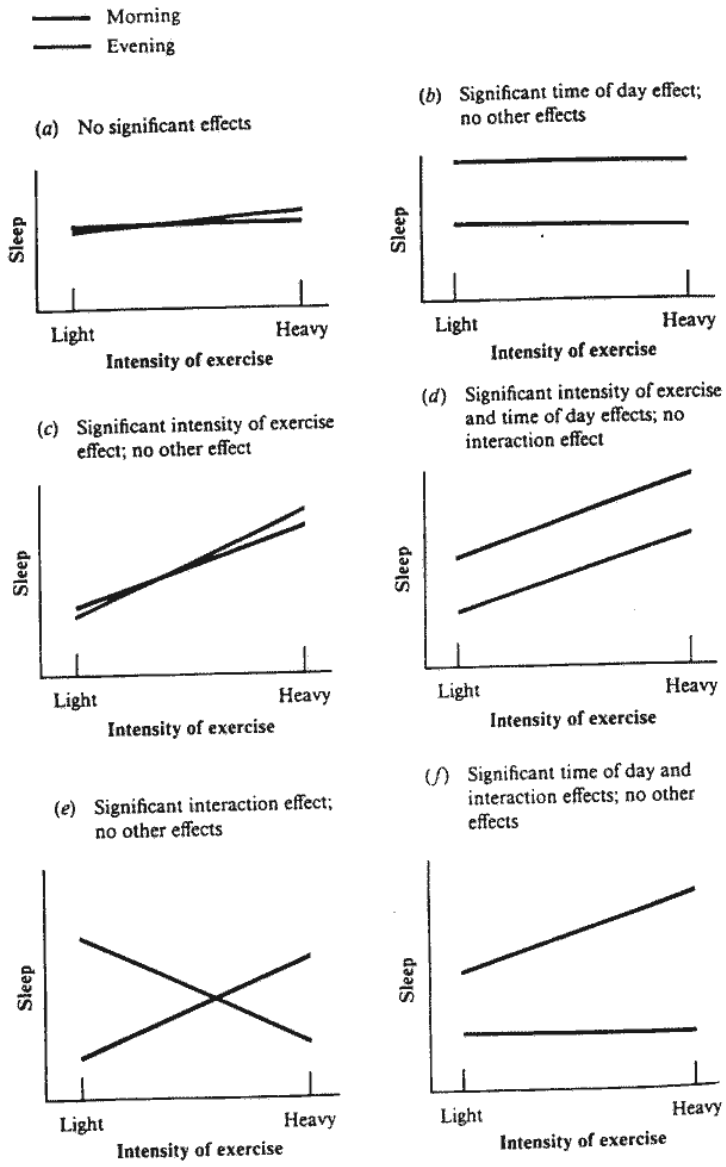
Többtényezős varianciaanalízis

Két vagy több független változó hatása a függő változóra.

Nullhipotézisek: (1) Első tényező (független változó) nincs hatással a függő változóra. (2) Második tényező nincs hatással a függő változóra. (3) Két tényező nincs egymásra hatással, nincs közöttük interakció.

Eljárás: először a két független változó közötti interakciót teszteljük, majd ezek hatását külön-külön.

Interakció


Morning
Evening
(a) No significant effects
Sleep
Light
Heavy
Intensity of exercise
(b) Significant time of day effect; no other effects
Sleep
Light
Heavy
Intensity of exercise
(c) Significant intensity of exercise effect; no other effect
Sleep
Light
Heavy
Intensity of exercise
(d) Significant intensity of exercise and time of day effects; no interaction effect
Sleep
Light
Heavy
Intensity of exercise
(e) Significant interaction effect; no other effects
Sleep
Light
Heavy
Intensity of exercise
(f) Significant time of day and interaction effects; no other effects
Sleep
Light
Heavy
Intensity of exercise

R-kód

Újabb növényeket öntözünk meg tápoldattal és vízzel, de most növényenként két eltérő fajtát tesztelünk, a *hagyományos* és a *rendkívüli* fajtát.

Kód közvetlenül letölthető innen, így:
load(url("https://phon.nytud.hu/mady/courses/statistics/materials/novtap2.RData"))

Az internetről közvetlenü betöltött adatmátrix neve novtap2, 24 sorból és három oszlopból áll.

```
h = aov(magassag~tapoldat*fajta,data=novtap2)
summary(h)
```

Tápoldat típusa és fajta nincs hatással egymásra, tehát nincs interakció a két független változó között.

```
h = aov(magassag~tapoldat+fajta,data=novtap2)
summary(h)
```

Egyes *p*-értékek így még kisebbek.

Értékelés

Döntés H_1 javára: az alkalmazott tápoldat mindkét növényfajta esetében szignifikánsan nagyobb növekedést okoz.
Kérdés: elég-e a két fajta esetében híg tápoldatot alkalmazni a szignifikáns növekedés kiváltásához?

Eljárás: 1-es és 2-es fajtára a víz és híg oldat p-értékének összehasonlítása Tukey-féle post hoc-teszttel (összes kombinációt interakciót feltételező modellel kapjuk csak meg).

```
h = aov(magassag~tapoldat*fajta,data=novtap2)
TukeyHSD(h)
```

	p adj
viz:1-hig:1	0.0181639
viz:2-hig:2	0.0005648

A híg oldat szignifikánsan nagyobb növekedést eredményez mindkét fajta esetében, a tömény és a híg oldat között viszont nem szignifikáns a különbség.

Gyakorló feladatok

ml_vow.RData alapján (letölthető: phon.nytud.hu/mady/courses/statistics/materials).

Igaz-e az, hogy a felső nyelvállású magánhangzók rövidebbek, mint a középső és alsó nyelvállásúak? (Szükséges oszlopok: dur, hgt.)

Hatással van-e a tartamra a környező mássalhangzó zöngéssége (voi), a magánhangzó-hosszúság (quan), és a magánhangzó minősége (qual)? Melyik tulajdonságok vannak interakcióban egymással?

Az adatok elemzése előtt érdemes a viszonyokat boxplotokon is megszemlélni.

Ismételt méréses módszerek

Humán tudományok örök problémája: egy személytől általában nem egy, hanem többféle adatot gyűjtünk. Ennek elemzésére az egyszerű varianciaanalízis NEM alkalmas, mert ott alapfeltétel a minták függetlensége (mint a független mintás t-próba esetén).

A varianciaanalízis függő mintás megfelelője az **ismételt méréses** varianciaanalízis, angolul *repeated measures ANOVA*.

Fontos: az ismételt mérés nem arra vonatkozik, hogy egyazon beszélőtől többször vesszük fel ugyanazt az adatot (pl. mondatokat öt ismétléssel olvasnak fel), hanem hogy **egyazon személlyel** ismételt méréseket végzünk.

Például orvostudományban: egy bizonyos gyógyszer hatása kezelés előtt, a kezelés megkezdése után két héttel, egy hónappal stb.

Eljárás

Egy függő és egy vagy több független változó tesztelése, ahol az ismétlés **belső** tényezői (személy vagy tárgy, akiken/amiken az ismételt méréseket végeztük) közötti különbséget **véletlen** hatásnak tekintjük (*within-subjects factor*).
Az alanyok tartozhatnak két különböző csoporthoz, (különböző nyelvek beszélői, egy növényfaj különböző fajtái stb.), ez a **köztes** tényező (*between-subjects factor*).

Feltételek:

- legalább öt alany (személy, növény, tárgy, bármi, amin több mérést végzünk),
- faktorkombinációnként egyetlen adat - azaz ha egyazon faktort több ismétléssel mértünk, ezeket a teszt futtatása előtt átlagolni kell minden egyes alanyra és cellára,
- kiegyensúlyozott dizájn, azaz ha az egyik faktor két szintjéhez két további faktor tartozik, akkor a másik faktornál is vizsgálni kell ugyanezt a két szintet akkor is, ha nem releváns.

Hátulütők

- R-ben nincs több faktor kombinációjára átlagoló beépített függvény,
- mivel átlagokkal számolunk, az egyes cellákon belüli varianciát nem tudjuk figyelembe venni,
- nem tudunk több *within-subject* tényezőt kombinálni (pl. résztvevő és többféle mondat egyazon kategóriából),
- csak akkor alkalmazható, ha a páronként összehasonlított faktorok különbségeinek varianciái azonosak (szfericitási feltétel),
- nincs post-hoc tesztje, csak t-próbák Bonferroni-korrektúrával (konfidenciaszint/összes lehetséges kombináció száma).

Ezekre a kevert modellek jelentenek majd kiutat, amik cserébe számos új problémát libbentenek fel.

Cellánkénti átlagok számítása

`anova.mean.r` nevű R-függvény letöltése innen:
phon.nytud.hu/mady/courses/statistics/materials

Szkript és függvény közötti különbség: függvényben létrehozott változók (R-objektumok) nem jelennek meg a munkamemóriában. Szkript és függvény egyaránt betölthető a `source("eleresiutvonal")` paranccsal, a szkriptet közvetlenül be is lehet másolni egy szövegszerkesztőből az R-be (copy-paste).

Ha a függvényben szintaktikai hiba van, betöltés helyett hibajelzést kapunk.

Függvény első sora:
`fuggvenynev = function(kotelezoargumentum1, kotelezoargumentum2, ...)`, ahol három pont további opcionális számú opcionális argumentumot jelöl.

Példa

Mondatvégi kétszótagú, /s/-re és /z/-re végződő szavakban megmértük a frikatíván belüli zöngés tartomány hosszát. Zöngésebbek-e a mondatvégi /z/-k, mint az /s/-ek?

zfin.RData, letöltés innen:
phon.nytud.hu/mady/courses/statistics/materials

```
zmean = anova.mean(zfin$cvoice,zfin$subj,zfin$voiced)
```

Kapott adatmátrix oszlopainak elnevezése:

```
names(zmean) = c("cvoice","subj","voiced")
```

Ismételt méréses varianciaanalízis függvénye

- Függő változó: mássalhangzó zöngésségének tartama (cvoice).
- Független változó: zöngésség (voiced).
- Within-subjects factor: beszélő (subj).
- Between-subjects factor: nincs.

```
summary(aov(cvoice ~ voiced + Error(subj/voiced),
data=zmean))
```

Releváns *p*-érték: `Error: subj:voiced` sor alatt (ez jelzi az alanyok szerinti interakciót).

Ábrázolás:

```
interaction.plot(x-tengely, ismételt_mérés_alanya,
paraméter)
interaction.plot(zmean$voiced,zmean$subj,zmean$cvoice)
```

Több tényező

Többtényezős varianciaanalízis képlete, pl. ha megelőző mássalhangzóra is kíváncsiak vagyunk:

```
summary(aov(cvoice ~ voiced*c1 +
Error(subj/(voiced*c1)), data=zmean))
```

Ehhez a cellánkénti átlagokat újra kell számolni:

```
zmean = anova.mean(zfin$cvoice, zfin$subj,
zfin$voiced, zfin$c1)
```

Eredmények

Értelmezés:

`Error:  subj:voiced` zöngésségi tartamok beszélőnként, zöngésség függvényében (a p-érték változott, mert az átlagokat újraszámoltuk).

`Error:  subj:c1` zöngésségi tartamok beszélőnként, a megelőző mássalhangzó függvényében.

`Error:  subj:voiced:c1` zöngésségi tartamok beszélőnként, zöngésség és megelőző mássalhangzó interakciója, azaz befolyásolja-e a megelőző mássalhangzó a zöngésség hatását.

Köztes tényező (between-subjects factor)

Bodo Winter példája: alapfrekvencia az udvariasság függvényében (inf=informal, pol=polite), férfiaknál és nőknél.

Letölthető innen:
https://bodo-winter.net/tutorials.html
dataset for tutorial 2

Betöltés legegyszerűbb `read.csv` függvénnyel, mert ott a vessző az alapértelmezett cellaelválasztó jel. Mentés a `pol` objektumba.

Figyelem! A nem Mac-felhasználóknak gondjai lehetnek a betöltéssel a Mac-specifikus sortörésjelek miatt (ez a Windows és a Linux txt-formátuma között is jelenthet nehézséget).

Megoldás: megnyitás szövegmegjelenítővel (notepad++, gedit vagy más), és mentés a ti OP-rendszeretek kódolása szerinti formátumban. Vagy letöltés innen:
https://phon.nytud.hu/mady/courses/statistics/materials/politeness.RDa

Értelmezzük a kapott eredményeket!