Varianciaanalízis

Varianciaanalízis (analysis of variance, ANOVA)

Kérdések: (1) van-e különbség **kettőnél több** csoport között (t-próba általánosítása), (2) van-e hatása a vizsgált tényezőnek: független változó(k) hatása a függő változóra.

- **Egy- vs. többtényezős**: ha egy független változó van, egytényezős (pl. három korosztály összehasonlítása), ha kettő, kéttényezős (pl. két korosztály teljesítménye két különböző iskolában), stb.
- **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 (pl. pulzus reggel, délben és este), 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 jelentéskategóriák hatása a szavak felismerésének reakcióidejére).

Feltételek

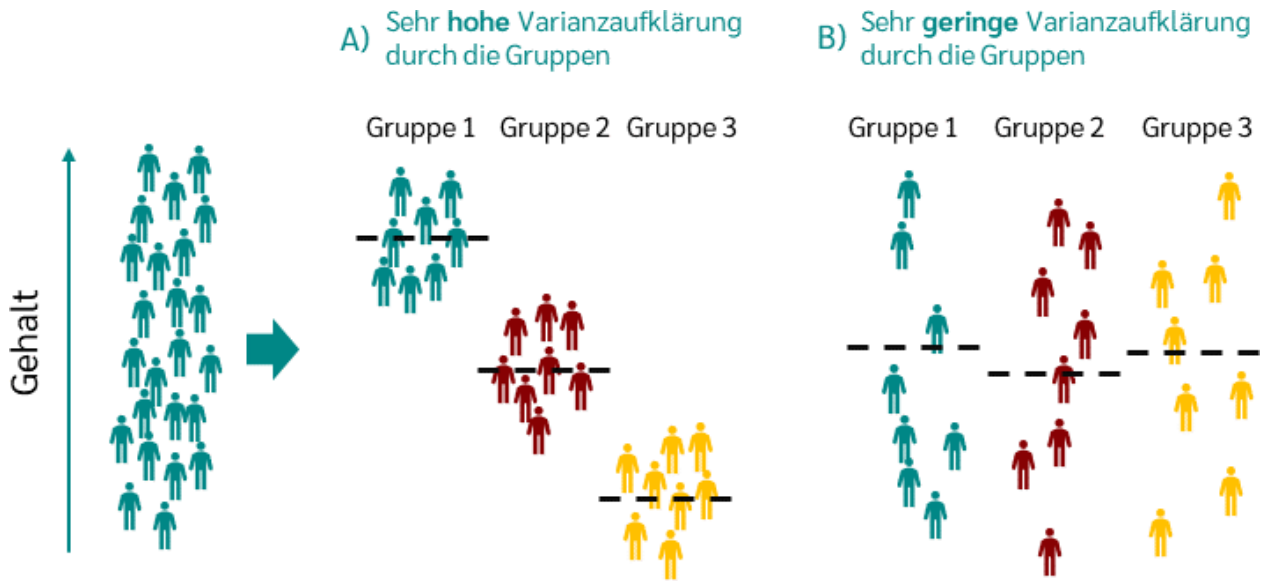
- Egyes csoportokon belül normális eloszlás és
- azonos szórás (varianciák homogenitása),
- megfigyelések egymástól való függetlensége (szfericitás, pl. egy egyénen belül a reggeli és esti teljesítmény között nem várunk összefüggést).

A normális eloszlás feltételének megsértését nem szokás sarkalatos problémának tekinteni, mert (1) 30 fölötti elemszám már gyakran normális eloszlású, (2) 10–20 elemnél nem nagy az eltérés, (3) 10-nél kisebb elem esetén nincs igazán értelme eloszlásról beszélni. A normális eloszlástól való erős eltérés esetén jobb a nemparametrikus Kruskal-Wallis próbát alkalmazni.

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

Példa

Három különböző végzettségű csoport jövedelmének összehasonlítása.
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

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.

Varianciatábla

Variancia eredete *source*	Szabadsági fok *df*	Eltérés-négyzetösszeg *Sum Sq*	Átlagos eltérés-négyzetösszeg *Mean Sq*	F	p
Kezelések közötti *between*	$k-1$	SS_K	$MS_K = \frac{SS_K}{k-1}$	$F = \frac{MS_K}{MS_H}$	p
Kezelésen 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

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 (egy elemen egyetlen mérést végzünk), az ismételt mérésesre (egy elemen kettőnél több mérést végzünk) 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 határát alacsonyabb *p*-értékben határozzuk meg: Bonferroni-korrektúra $0,05/3 = 0,0167$.

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

p = 0.4462, azaz a különbség messze nem szigifikáns.

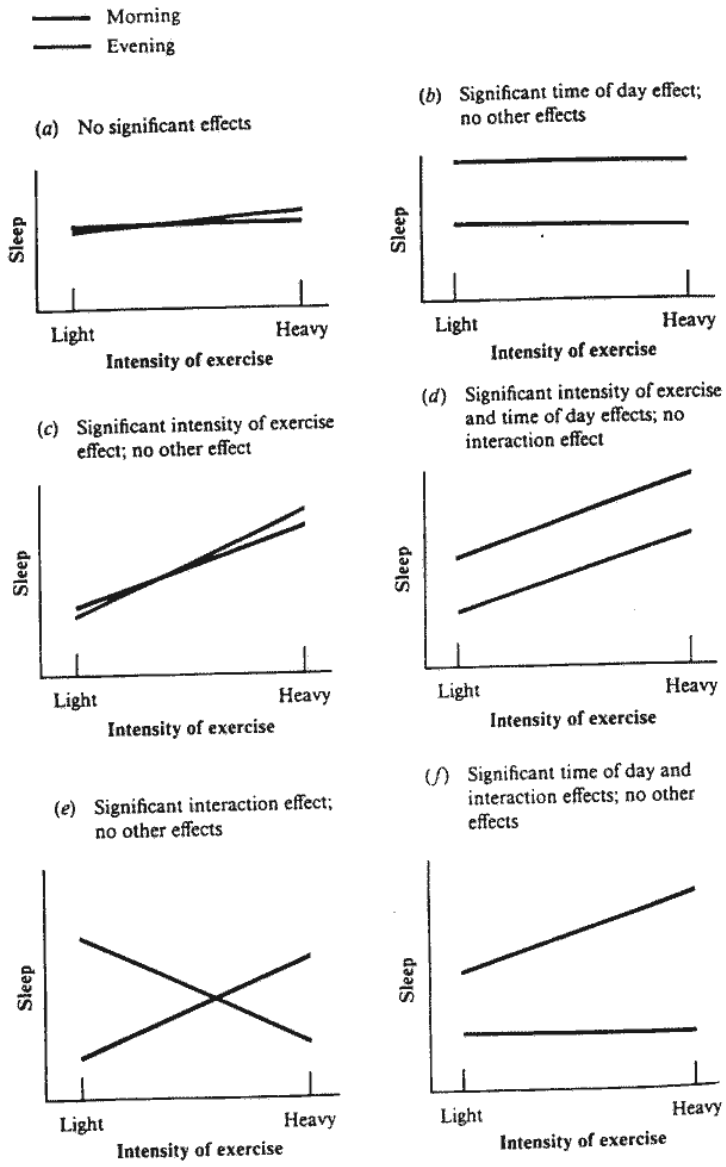
Többtényezős varianciaanalízis

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.

novtap2.RData letöltése innen:
https://phon.nytud.hu/mady/courses/statistics/materials/

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

A fajta és a tápoldat szignifikáns hatással van a növény méretére. A tápoldat típusa és a fajta viszont nincs hatással egymásra, tehát nincs interakció a két független változó között. Ezért az interakciót elhagyjuk a képletből.

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

Egyes p-értékek így még kisebbek, az oldat és a fajta hatása (magyarázó ereje) erős.

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 vízhez képest, a tömény és a híg oldat között viszont nem szignifikáns a különbség. Tehát lehet spórolni a hatóanyaggal.