

Varianciaanálisis (ANOVA)

Csoportok összehasonlítása (ismétlés)

t-próba:

- ▶ Két **független** minta összehasonlítása az átlaguk és szórásuk alapján. Ha a csoportok varianciái azonosak: független mintás *t*-próba, ha nem, Welch-teszt. Példa: gyógyszer és placebo hatása a vérnyomásra.
- ▶ Egyazon alanytól két mintát veszünk, és az alanyokon **belüli** különbséget hasonlítjuk össze 0-val, azaz ha nincs különbség a két mérés között. Példa: gyógyszer hatása a kezelés elején és hat hét után.
- ▶ Egy minta összehasonlítása egy ismert átlagú és szórású mintával. Példa: egy adott csoport IQ-tesztje a populációhoz képest.

DE: a világ nem mindig egyszerűsíthető két kategóriára!

Varianciaanalízis (analysis of variance, ANOVA)

Kérdések: van-e különbség a csoportok között (t -próba általánosítása), és van-e hatása a vizsgált tényezőnek. Eljárás: regressziószámítás.

- ▶ **Egy- vs. több****tényezős**: ha egy független változó van, egytényezős, ha n , n -tényezős. Egytényezős: darazsak száma három helyszínen, kéttényezős: darazsak és szúnyogok száma két (vagy több) helyszínen.
- ▶ **Független mintás vs. ismételt mérések**: ha az adatok különböző elemeken végzett mérésekből származnak (pl. magyar, cseh és angol beszélők), független mintánk van. Ha egyazon adatközlőtől többféle adat származik, ismételt mérések dizájnunk van (pl. darazsak száma két helyszínen eső előtt és után).
- ▶ **Egy- vs. több****változós**: 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, placebo).
- ▶ Többféle módszer hatékonysága a nyelvtanulásban különböző időpontokban: ismételt méréses modell.
- ▶ Nominális, azaz kategoriális független változók által kiváltott hatás, pl. különböző szemantikai kategóriák hatása a reakcióidőre egy szófelismerési tesztben.

Feltételek

- ▶ Egyes csoportokon belül normális eloszlás és
- ▶ azonos szórás (varianciák homogenitása),
- ▶ egyénen belüli különbségek egymástól való függetlensége (szfericitás), vagyis az a feltétel, hogy egy első alacsony értékhez bármilyen második érték tartozhat.

Normális eloszlás feltétele parametrikus adatok esetén többnyire nem probléma, mert (1) 30 fölötti elemszám már természetesen normális eloszlású, (2) 10–20 elemnél nem nagy az eltérés, (3) 10-nél kisebb elemszám esetén nincs igazán értelme eloszlásról beszélni. Ellenőrizni persze mindenképpen kell.

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

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>.

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
→ véletlen hiba varianciabecslése = regressziószámítás
reziduális varianciája,
3. döntés: ha a **csoportok közötti** variancia nagyobb, mint a **csoportokon belüli** variancia, akkor a tényezőnek (független/magyarázó változónak) van hatása.

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* karakterváltozókat faktorrá alakítja.

Varianciatábla

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

SS_H = reziduális hiba a regressziószámítás alapján

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(függőváltozó,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 `↔ 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. Hiszen a $p = 0,05$ -ben meghatározott konfidenciahatár azt jelenti, hogy minden 20. összehasonlítás a 95%-os megbízhatósági határon **kívül** fog esni.

Módszerek:

- ▶ Párunkénti összehasonlítás t -próbákkal, majd a **Bonferroni-korrektúra** alkalmazása: szignifikancia-határ konfidenciaintervallum/összes lehetséges párosítás. Hátrány: nagy számú kombináció esetén nagyon nehéz szignifikáns különbséget kimutatni, mert a határként megjelölt p értéke 0,05-nél jóval kisebb lesz.
- ▶ **Tukey-féle** /tu:ki/ post-hoc teszt. Hátrány: csak a független mintás varianciaanalízisre alkalmazható, az ismételt mérésekre 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 szignifikáns.

Többtenyezős varianciaanalízis

Két vagy több független változó hatása a függő változóra.

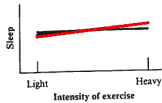
Nullhipotézisek: (1) Első tényező (független változó) nincs hatással a függő változóra. (2) Második tényező nincs hatással a függő változóra. (3) Két tényező nincs egymásra hatással, nincs közöttük interakció.

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

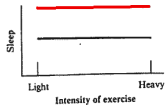
Interakció

— Morning
— Evening

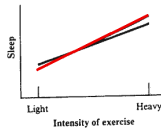
(a) No significant effects



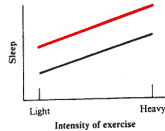
(b) Significant time of day effect;
no other effects



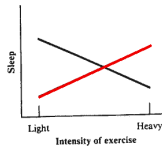
(c) Significant intensity of exercise
effect; no other effect



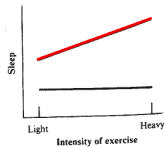
(d) Significant intensity of exercise
and time of day effects; no
interaction effect



(e) Significant interaction effect;
no other effects



(f) Significant time of day and
interaction effects; no other
effects



R-kód

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

Betöltés az R-be:

```
load(url("https://phon.nytud.hu/mady/courses/statistics/  
materials/novtap2.RData"))
```

Kiegyensúlyozott dizájn, vagyis a csoportok elemszáma azonos.

Először interakcióval együtt állítunk modellt:

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

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

```
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.

További feladat

ml_vow.RData alapján (letölthető innen:
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 megvizsgálni.