Cuprins

Capitolul 1

ANOVA - Analiza de varianţă

1.1 ANOVA uni-factorială. Concepte de bază.

Analiza de varianţă continuă seria testelor de semnificaţie şi întăreşte înţelegerea analizei cauzale, facând trecerea spre analiza de regresie.

După cum ne reamintim, `t` este un test de semnificaţie care testează diferenţa dintre două medii ale unor variabile metrice. ANOVA este o generalizare a testului `t` pentru mai mult de două medii, pentru că deseori în cercetarea socială studiem mai mult de două grupuri în acelaşi timp. De subliniat este faptul că ANOVA poate să fie folosită cu rezultate foarte bune şi la comparaţia dintre două medii, însă testul îşi arată adevărata valoare la trei sau mai multe medii.

Exemple de acest tip pot fi multiple:

- putem testa dacă mai multe campanii pro-nataliste diferă unele de altele sub aspectul efectelor acestora (sau în sens contrar, să vedem dacă mai multe campanii diferite de prevenire a sarcinilor nedorite diferă semnificativ unele de altele); unele campanii se pot orienta cu preponderenţă către spoturi TV, altele către distribuirea de materiale informative tipărite, altele pe consiliere directă ş.a.m.d.
- putem deasemenea să testăm dacă mai multe acţiuni de creştere a participării civice diferă sau nu în ce priveşte efectele
- putem testa dacă mai multe strategii locale de combatere a sărăciei diferă sau nu între ele, etc.

ANOVA combină şi extinde testele `t` şi χ^2, prin testarea egalităţii dintre trei sau mai multe medii (pentru trei sau mai multe grupuri). Se testează aşadar legătura dintre o variabilă metrică (pentru care se calculează media) şi o variabilă calitativă (a cărei valori sau categorii sunt considerate grupuri independente). De asemenea, ANOVA face o introducere clară în analiza cauzală: variabila cauză (independentă) este cea calitativă iar variabila efect (dependentă) este cea metrică.

În exemplul pe care îl vom expune, avem următoarele două variabile:

VÂRSTĂ	**vârsta** la angajare a unei persoane (variablă raport, metrică, continuă)
STRATIN	**str**ategia de **a**tragere a **tin**erilor (variabilă calitativă, nominală, cu mai multe categorii - pot exista mai multe strategii posibile)

În această carte, noi luăm în considerare doar varianta de *ANOVA uni-factorială*, cu un singur factor (în engl. one-way ANOVA), deoarece folosim o singură variabilă categorială (denumită în limba engleză "factor") pentru a testa diferenţele între mediile grupurilor definite de categoriile acesteia (denumite în engleză "levels" - niveluri sau "treatments" - tratamente).

În cazul nostru, ne raportăm *doar* la strategia de atragere a tinerilor; este important însă de ştiut că vârsta la angajare a tinerilor poate fi influenţată de diverşi factori şi că există variante ale analizei de varianţă care iau în calcul mai mulţi asemenea factori. De pildă, varianta care ia în calcul doi factori se numeşte *ANOVA bi-factorială* (în engl. two-way ANOVA)iar varianta care ia în calcul mai mulţi factori se numeşte *ANOVA multi-factorială* (în engl. multi-way ANOVA).

Pentru această secţiune, să presupunem că studiem oportunităţile de acces a tinerilor pe piaţa forţei de muncă şi analizăm diferite strategii folosite pentru a atrage tinerii să se angajeze. Ipoteza pe care dorim să o testăm este următoarea: *media de vârstă a persoanelor nou angajate este influenţată de strategia de atragere utilizată.*

Setul de ipoteze generale pentru ANOVA este:

$$\begin{cases} \mathrm{H}_0: & \mu_1 = \mu_2 = \cdots = \mu_k \\ \mathrm{H}_\mathrm{A}: & \text{cel puţin două medii sunt diferite} \end{cases}$$

După cum se poate observa, ipoteza de nul se referă la mediile din populaţie (notate cu μ). Pentru simplificare, vom lua în considerare doar trei localităţi, care beneficiază de strategii diferite. Din fiecare localitate vom extrage câte un eşantion (să spunem de 10 persoane), iar ipotezele devin:

$$\begin{cases} \mathrm{H}_0: & \mu_1 = \mu_2 = \mu_3 \\ \mathrm{H}_\mathrm{A}: & \text{cel puţin două localităţi au medii de vârstă sunt diferite} \end{cases}$$

În cuvinte, ipoteza de nul susţine că nu este nici o diferenţă între rezultatele diferitelor strategii (fie sunt toate strategiile foarte bune şi atrag mulţi tineri, fie sunt toate foarte slabe şi nu atrag tineri) iar ipoteza alternativă susţine că cel puţin o strategie dă rezultate mai bune decât cel puţin una dintre celelalte (este posibil ca o strategie să aibă un rezultat de mijloc, care nu este semnificativ diferit nici faţă de strategia de succes maxim, nici faţă de strategia care atrage cei mai puţini tineri; diferenţă semnificativă în acest caz există doar între prima şi ultima strategie).

După cum se poate vedea în **Tabelul 1.1**, primul grup (cel de control, din localitatea unde nu s-a aplicat nici o strategie) are o medie a vârstei la angajare de 27,8 ani cu o abatere standard de 3,65 ani; al doilea grup (din localitatea unde s-a aplicat prima strategie) are o medie de 23,6 ani cu o abatere standard

Tabelul 1.1: Vârstele la angajare a persoanelor în cadrul a 3 eşantioane independente.

Nr.crt.	Localitate 1	Localitate 2	Localitate 3
1	22	28	20
2	27	22	28
3	32	24	31
4	30	18	26
5	29	21	26
6	27	26	30
7	33	25	21
8	24	20	25
9	24	24	29
10	30	28	27
$\bar{x}$	**27,8**	**23,6**	**26,3**
s	**3,65**	**3,34**	**3,59**

de 3,34 ani iar al treilea grup o medie de 26,3 ani cu o abatere standard de 3,59 ani.

La o primă vedere, toate cele trei grupuri conţin tineri: există vreo diferenţă semnificativă între cele trei medii? Cum testăm, mai exact, acest lucru?

În fine, dacă obiectivul principal al acestei analize este de a testa diferenţele dintre medii, de ce se numeşte "*Analiză de varianţă*"?

1.2 Funcţionarea şi logica analizei de varianţă

Privind **Tabelul 1.1**, putem extrage câteva informaţii interesante, care ne vor ajuta în cele ce vor urma.

Avem un eşantion total format din 30 de persoane, deci $n = 30$. Acest eşantion este format din trei grupuri independente de câte 10 persoane fiecare (subeşantioane din trei localităţi diferite); avem deci: $n_1 = 10$, $n_2 = 10$ şi $n_3 = 10$. Pentru fiecare dintre cele trei localităţi/grupuri putem calcula câte o medie şi câte o abatere standard; mai avem aşadar: $\bar{x}_1 = 27,8$ cu $s_1 = 3,65$, $\bar{x}_2 = 23,6$ cu $s_2 = 3,34$ şi $\bar{x}_3 = 26,3$ cu $s_3 = 3,59$.

În acelaşi timp, putem calcula o medie generală pentru eşantionul total (pentru toate cele 30 de observaţii) $\bar{\bar{x}} = 25,9$ precum şi o abatere standard totală $s = 14,714$.

Sintetizând:

$$\begin{array}{lll} n = 30 & \bar{\bar{x}} = 25,9 & s = 14,714 \\ n_1 = 10 & \bar{x}_1 = 27,8 & s_1 = 3,65 \\ n_2 = 10 & \bar{x}_2 = 23,6 & s_2 = 3,34 \\ n_3 = 10 & \bar{x}_3 = 26,3 & s_3 = 3,59 \end{array}$$

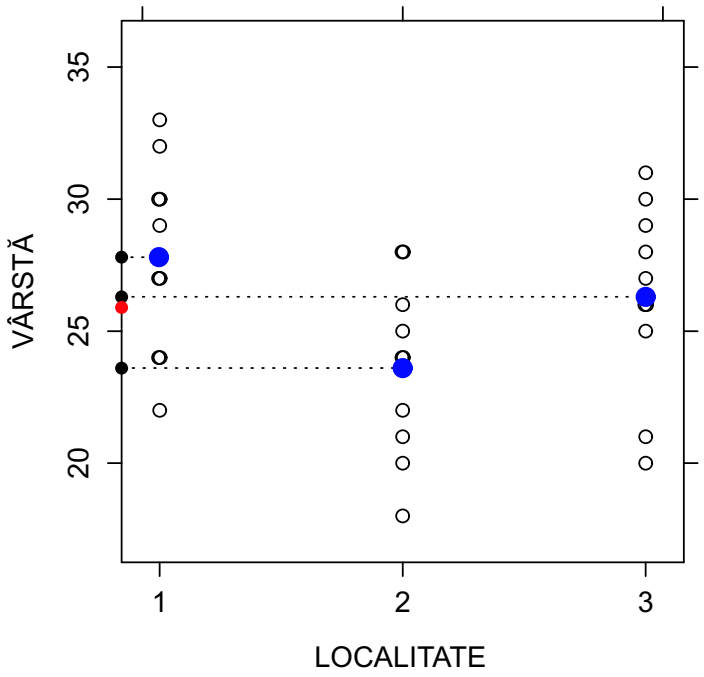
Din faptul că putem calcula abaterile standard, avem un prim indiciu că există o *variaţie internă* în cadrul fiecărui grup. În **Figura 1.1** se pot observa cele trei grupuri, ale căror observaţii variază în jurul mediilor reprezentate de punctele colorate în albastru. Această variaţie poate fi calculată uşor, *pentru fiecare grup în parte*, cu binecunoscuta formulă:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{1.1}$$

Avem aşadar trei varianţe în interiorul grupurilor (câte o varianţă pentru fiecare din cele trei grupuri).

Apoi, pentru că există trei medii diferite pentru fiecare eşantion, plus o medie generală pentru toate eşantioanele, se poate constata o variaţie a celor trei medii de grupuri în jurul mediei generale.

Figura 1.1: Scatterplot al vârstelor la angajare pentru cele trei eşantioane.



Celor trei medii le corespund punctele de culoare albastră de pe axa $0y$, iar media generală ($\bar{\bar{x}} = 25,9$) este punctul roşu de pe aceeaşi axă. Poate fi uşor observată variaţia punctelor albastre în jurul punctului roşu, cu alte cuvinte variaţia mediilor de grupuri în jurul mediei generale. În cazul nostru, deoarece avem doar trei grupuri, $k = 3$. Adaptând formula 1.1, obţinem:

$$s_{\bar{x}}^2 = \frac{\sum_{j=1}^{k}(\bar{x}_j - \bar{\bar{x}})^2}{k-1} \tag{1.2}$$

După cum ne aducem aminte de la distribuţia de eşantionare a mediei, deviaţia standard a mediilor în jurul mediei generale este denumită Eroare Standard. **Ecuaţia 1.2** indică aşadar o estimare a Erorii Standard din populaţie, de unde putem extrage foarte simplu varianţa din populaţie, deoarece:

$$\text{ES} = \frac{\sigma}{\sqrt{n}}$$

Ne confruntăm, deci, cu două tipuri de variaţii: o variaţie *între grupuri* (variaţia mediilor de grup în jurul mediei generale) şi una *în interiorul grupurilor* (variaţiile observaţiilor în jurul fiecărei medii de grup). Ambele tipuri de variaţii sunt folosite ca estimări ale *variaţiei generale* în populaţie.

IMPORTANT!
Analiza de varianţă se bazează pe comparaţia dintre două estimări ale varianţei σ^2 **pentru întreaga populaţie**.

Logica analizei este următoarea: dacă cele două estimări ale varianţei din populaţie σ^2 sunt aproximativ egale, atunci ipoteza de nul este adevărată (în populaţie, toate mediile sunt egale). Dacă ipoteza de nul nu este adevărată, atunci cele două estimări ale varianţei vor fi semnificativ diferite.

Analiza de varianţă se efectuează în trei paşi:

- Calcularea primei estimări a varianţei în populaţie: varianţa *dintre mediile grupurilor*
- Calcularea celei de a doua estimări a varianţei în populaţie: varianţa *din interiorul grupurilor*
- Se compară cele două estimări cu ajutorul statisticii test F. Dacă sunt aproximativ egale (raportul dintre cele două este aproape de valoarea 1), atunci *nu respingem* ipoteza de nul.

Un lucru important de care trebuie să ne aducem aminte (de la măsurile tendinţei centrale, capitolul de descriere a variabilelor) este caracterizarea varianţei; să mai examinăm încă odată formula:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Avem în partea de sus o *sumă de pătrate*, împărţită la un număr de *grade de libertate* (gl = numărul de observaţii n minus 1, detalii la capitolul de care vorbeam). Ecuaţia de mai sus mai poate fi scrisă ca:

$$s^2 = \frac{\text{SP}}{gl}$$

După cum ştim, orice sumă împărţită la numărul de observaţii se numeşte *medie*, de unde reiese că varianţa nu este nimic altceva decât o medie a unei sume de pătrate.

De aici şi denumirile pe care le poartă, în literatura de specialitate, cele două estimări ale varianţei din populaţie σ^2:

- **MPD** (**M**edia sumei **P**ătratelor **D**intre grupuri) - varianţa dintre mediile grupurilor, unde:
$$\text{MPD} = \frac{\text{SPD}}{gl_{\text{D}}}$$
- **MPI** (**M**edia sumei **P**ătratelor din **I**nteriorul grupurilor) - varianţa din interiorul grupurilor, unde:
$$\text{MPI} = \frac{\text{SPI}}{gl_{\text{I}}}$$

IMPORTANT!
Atât MPD cât şi MPI reprezintă estimări ale varianţei în populaţie, deci pot fi notate amândouă cu $\hat{\sigma}^2$.

1.3 Calcularea varianţei dintre grupuri

Aplicăm mai întâi formula din **Ecuaţia 1.2**:

$$s^2_{\bar{x}} = \frac{\sum_{j=1}^{k}(\bar{x}_j - \bar{\bar{\text{x}}})^2}{k-1}$$

Folosind estimarea Erorii Standard, obţinem estimarea varianţei din populaţie:

$$\widehat{\text{ES}} = \frac{\hat{\sigma}}{\sqrt{n}} \quad \text{de unde reiese că} \quad \hat{\sigma}^2 = n{\cdot}\widehat{\text{ES}}^2$$

În această formulă, n este mărimea totală a eşantionului. Cum $s_{\bar{x}}$ este chiar estimarea Erorii Standard, putem înlocui pe $\widehat{\text{ES}}^2$ cu $s^2_{\bar{x}}$ şi obţinem:

$$\hat{\sigma}^2 = n{\cdot}s^2_{\bar{x}}$$

$$\Rightarrow \hat{\sigma}^2 = \text{MPD} = \frac{\sum_{j=1}^{k} n(\bar{x}_j - \bar{\bar{\text{x}}})^2}{k-1}$$

Numărul de observaţii n, fiind o constantă, poate sta oriunde: înaintea fracţiei, înaintea sumei sau chiar în interiorul sumei. Există un motiv special pentru

care n stă în interiorul sumei, pentru că trebuie să luăm în calcul şi mărimea grupurilor (în cazul nostru cele trei grupuri au mărime egală, însă de obicei nu este aşa); distanţa dintre media de grup $\bar{x}_j$ şi media generală $\bar{\bar{x}}$ va fi ponderată cu mărimea grupului respectiv n_j (rezultatul final fiind exact acelaşi):

$$\text{MPD} = \frac{\text{SPD}}{gl_D} = \frac{\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2}{k-1} \tag{1.3}$$

Cu alte cuvinte, grupurile mai mari vor avea o pondere mai mare (vor *cântări* mai mult în calcul) decât grupurile mai mici, ceea ce este absolut normal.
Se poate observa că *numărul de grade de libertate* pentru MPD este $gl = k - 1$.

Efectuând calculele pentru exemplul nostru:

$$\text{MPD} = \frac{10{\cdot}(27,8 - 25,9)^2 + 10{\cdot}(23,6 - 25,9)^2 + 10{\cdot}(26,3 - 25,9)^2}{3-1}$$

$$\Rightarrow \text{MPD} = \frac{36,1 + 52,9 + 1,6}{2} = \frac{90,6}{2}$$

$$\Rightarrow \text{MPD} = 45,3$$

Spunem că varianţa *dintre* grupuri este egală cu 45,3.

1.4 Calcularea varianţei din interiorul grupurilor

O întrebare pertinentă în acest moment este: *ce înţelegem prin **varianţa** din interiorul grupurilor*? Avem trei grupuri, deci trei varianţe (cîte una în interiorul fiecăruia); pe care dintre cele trei o folosim?

Există două răspunsuri posibile:

1. putem folosi oricare dintre cele trei varianţe, dacă ele sunt egale în populaţie (ceea ce ne duce spre una dintre asumpţiile acestei analize, prezentată în **Secţiunea 1.7**)
2. putem folosi o medie ponderată a tuturor celor trei varianţe, folosind o procedură derivată de asemenea din formula clasică a varianţei.

Având trei grupuri, adunăm trei sume de pătrate:

$$\text{MPI} = \frac{\text{SPI}}{gl} = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2 + \sum_{i=1}^{n_3}(x_{3i} - \bar{x}_3)^2}{n_1 - 1 \quad + \quad n_2 - 1 \quad + \quad n_3 - 1}$$

Partea de jos a ecuaţiei este egală cu $n_1+n_2+n_3-3 = n-3$ (pierdem trei grade de libertate pentru că avem trei puncte fixe: mediile corespunzătoare celor trei grupuri). La modul general (cu k grupuri) aceasta va fi egală cu $n-k$, iar ecuaţia poate fi scrisă sub forma unei sume duble:

$$\text{MPI} = \frac{\text{SPI}}{gl_I} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji} - \bar{x}_j)^2}{n-k} \tag{1.4}$$

Numărul de grade de libertate pentru MPI este aşadar: $gl = n - k$.

Ecuaţia 1.4 poate fi simplificată şi mai mult, folosind formula varianţei:

$$s_j^2 = \frac{\sum_{i=1}^{n_j}(x_{ji} - \bar{x}_j)^2}{n_j - 1} \quad \Rightarrow \quad \sum_{i=1}^{n_j}(x_{ji} - \bar{x}_j)^2 = (n_j - 1)s_j^2$$

pentru fiecare dintre cele $j = 1 \ldots k$ grupuri, de unde:

$$\text{MPI} = \frac{\text{SPI}}{gl} = \frac{\sum_{j=1}^{k}(n_j - 1)s_j^2}{n-k} \tag{1.5}$$

La fel ca la MPD, avem şi aici o medie ponderată a varianţelor celor k grupuri (aici însă ponderarea s-a realizat prin utilizarea gradelor de libertate ale fiecărui grup): grupurile de mărime mai mare vor avea o pondere mai mare în calcul.

Aplicând **Ecuaţia 1.5** pentru exemplul nostru cu trei eşantioane:

$$\text{MPI} = \frac{(10-1)3,65^2 + (10-1)3,34^2 + (10-1)3,59^2}{30-3}$$

$$\Rightarrow \text{MPI} = \frac{119,6 + 100,4 + 116,1}{27} = \frac{336,1}{27}$$

$$\Rightarrow \text{MPI} = 12,448$$

Spunem că varianţa *în interiorul* grupurilor este egală cu 12,448.

Ecuaţiile 1.3 şi **1.5** sunt cele folosite pentru calcularea celor două estimări ale varianţei în populaţie, în cazul general cu k grupuri şi mărimi n_k ale grupurilor. Formulele utilizate nu sunt foarte complicate (chiar dacă aşa par la prima vedere), bazându-se exclusiv pe formula clasică a varianţei. Din fericire pentru persoanele cu abilităţi matematice mai scăzute, ele nu trebuie calculate de mână; computerul ne va da automat rezultatele, singura noastră grijă fiind aceea de a le interpreta corect.

1.5 Statistica F. Tabelul ANOVA.

Cel de-al treilea pas în efectuarea analizei de varianţă este calcularea statisticii test F, ca raport între cele două estimări ale varianţei în populaţie:

$$\mathtt{F} = \frac{\text{Varianţa dintre grupuri}}{\text{Varianţa în interiorul grupurilor}}$$

sau mai simplu:

$$\mathtt{F} = \frac{\text{MPD}}{\text{MPI}} \tag{1.6}$$

După cum vom vedea, există un motiv puternic pentru faptul că MPD se află la numărător, în partea de sus a fracţiei.

Multe din informaţiile prezentate în continuare sunt explicate în detaliu la **Capitolul ??** - Regresia liniară simplă; în general, multe informaţii în statistică se bazează pe altele, formând un tot unitar. Singura metodă de a înţelege corect toate informaţiile este de a citi capitolele care fac referiri unele la altele şi de a reciti un capitol cu referinţele proaspăt citite.

Să vedem însă cum interpretăm raportul F: MPI, varianţa în interiorul grupurilor, este un bun estimator al varianţei din populaţie σ^2, *indiferent* dacă ipoteza de nul este sau nu adevărată. Aceasta deoarece MPI se bazează pe variaţiile din interiorul fiecărui grup, care luate împreună oferă o imagine destul de bună (o estimare destul de bună) a lui σ^2.

Partea care se află sub lupa testului este însă MPD; dacă ipoteza de nul este adevărată (toate mediile sunt egale) atunci şi MPD va fi un bun estimator a lui σ^2. În această situaţie MPD va avea o valoare apropiată de cea a lui MPI, iar valoarea lui F va fi aproape de 1 (raportul dintre două cantităţi egale este egal cu 1). Cu cât valoarea lui F se va apropia mai mult de 1, cu atât va creşte probabilitatea de a greşi respingând ipoteza de nul.

În cealaltă situaţie, în care ipoteza de nul nu este adevărată (adică cel puţin una dintre medii este semnificativ diferită), atunci valoarea lui F se va mări considerabil; în acelaşi timp, probabilitatea de a greşi respingând ipoteza de nul se va micşora pe măsură. Acest lucru se întâmplă deoarece diferenţele dintre grupuri tind să mărească MPD.

IMPORTANT!
Statistica F este o măsură care ne arată cât de multă variaţie se datorează diferenţelor dintre grupuri, raportată la variaţia generată de selecţia aleatoare a eşantionului.

Pentru a înţelege şi mai bine aceste lucruri, vom introduce încă o sumă de pătrate, ignorată până acum: STP (**S**uma **T**otală a **P**ătratelor); este vorba despre distanţa dintre toate observaţiile din eşantionul general în jurul mediei generale $\bar{\bar{x}}$:

$$\text{STP} = \sum_{i=1}^{n} (x_i - \bar{\bar{x}})^2$$

Dacă împărţim STP la $n-1$ vom obţine varianţa pentru eşantionul general, a tuturor observaţiilor. Deoarece varianţa este o măsură a variaţiei iar STP face parte din formula acesteia, rezultă că STP este de asemenea o *bună măsură* a variaţiei totale. Se poate arăta că:

$$\text{STP} = \text{SPD} + \text{SPI} \tag{1.7}$$

Cu alte cuvinte, cantitatea totală de variaţie este egală cu cantitatea de variaţie explicată de diferenţele dintre grupuri plus cantitatea de variaţie rămasă neexplicată (erorile aleatoare în jurul mediei); pe scurt, variaţia totală este egală cu variaţia explicată plus variaţia neexplicată, iar valoarea lui `F` se poate defini ca raport între cele două:

$$\texttt{F} = \frac{\text{VE}}{\text{VN}} \tag{1.8}$$

Cu cât variaţia explicată va fi mai mare, cu atât va scădea variaţia neexplicată, iar valoarea lui `F` va creşte spectaculos; invers, cu cât variaţia explicată va fi mai mică (ceea ce înseamnă că diferenţele dintre grupuri sunt foarte mici), cu atât va creşte variaţia neexplicată (datorată erorilor aleatoare) iar valoarea lui `F` va tinde spre zero (la limită, când grupurile seamănă perfect şi nu există absolut nici o diferenţă între ele, variaţia explicată va fi egală cu zero).

Diverse programe de analiză statistică pot să difere foarte puţin în modul de prezentare a rezultatelor, însă toate se vor referi la exact acelaşi lucru; în general, orice tabel de rezultate va conţine următoarele lucruri:

	gl	Suma Pătratelor	Varianţa	Valoare `F`	`p`
între grupuri	k - 1	SPD	MPD	$\texttt{F} = \frac{\text{MPD}}{\text{MPI}}$	Pr(>`F`)
în interiorul grupurilor	n - k	SPI	MPI		
Total	n - 1	STP			

Evident că aceasta este o variantă în limba română a tabelului; pentru că cele mai bune programe statistice sunt în limba engleză, rezultatele afişate de calculator vor avea denumirile ca în următorul tabel:

	df	Sum of Squares	Mean Square	`F` value	`p`
Between	k - 1	SSB	MSB	$\texttt{F} = \frac{\text{MSB}}{\text{MSW}}$	Pr(>`F`)
Within	n - k	SSW	MSW		
Total	n - 1	SST			

Uneori pot să apară prescurtări de genul “Sum Sq” sau “Mean Sq”, iar în loc de `p` se poate găsi frecvent “Sig.” (vom vedea semnificaţia acestora în următoarea secţiune). Mai departe:

- “df” se referă la *gl* - gradele de libertate
- “Sum of Squares” înseamnă “Suma Pătratelor”
- SSB (în engl. **S**um of **S**quares **B**etween) este echivalent cu SPD
- SSW (în engl. **S**um of **S**quares **W**ithin) este echivalent cu SPI

- SST (în engl. **S**um of **S**quares **T**otal) este echivalent cu STP
- MSB (în engl. **M**ean **S**quare **B**etween) este echivalent cu MPD
- MSW (în engl. **M**ean **S**quare **W**ithin) este echivalent cu MPI

Efectuând calculele pentru exemplul nostru, am obţinut următorul tabel:

	gl	Suma Pătratelor	Varianţa	Valoare F	p
între grupuri	2	90,6	45,3	3,639	0,040
în interiorul grupurilor	27	336,1	12,448		
Total	29	426,7			

Se poate vedea foarte clar modul cum a fost calculat F:

$$\texttt{F} = \frac{45,3}{12,45} = 3,639$$

Valoarea de 3,639 (mult mai mare decât 1) ne sugerează că ipoteza de nul este pe cale de a fi respinsă, pentru că variaţia explicată de diferenţele dintre grupuri este mult mai mare decât variaţia datorată erorilor aleatoare; existând diferenţe majore între grupuri, vor exista cu siguranţă şi diferenţe între mediile acestora.

La fel ca la testele t şi χ^2, decizia se ia după compararea acestei valori cu una critică. Modalitatea alternativă este de a compara valoarea lui p cu pragul de semnificaţie ales; cum p este mai mic decât $\alpha = 5\%$ (un prag generic, pentru un nivel de încredere de 95%), vom respinge ipoteza de nul: cel puţin una dintre strategii a dat rezultate.

1.6 Distribuţia de eşantionare F

În mod similar cu testele t şi χ^2, analiza de varianţă foloseşte o distribuţie de eşantionare numită pe scurt "Distribuţia F". Modul de testare a ipotezei de nul este de asemenea similar, prin calcularea unei valori F critice cu care se compară valoare F obţinută în test, sau prin compararea valorii p obţinute cu pragul de semnificaţie α.

Ca la orice distribuţie de eşantionare, putem distinge o primă caracteristică a distribuţiei F: este *continuă* (existând o infinitate de eşantioane posibile), cu un interval de valori care poate varia între 0 şi $+\infty$.

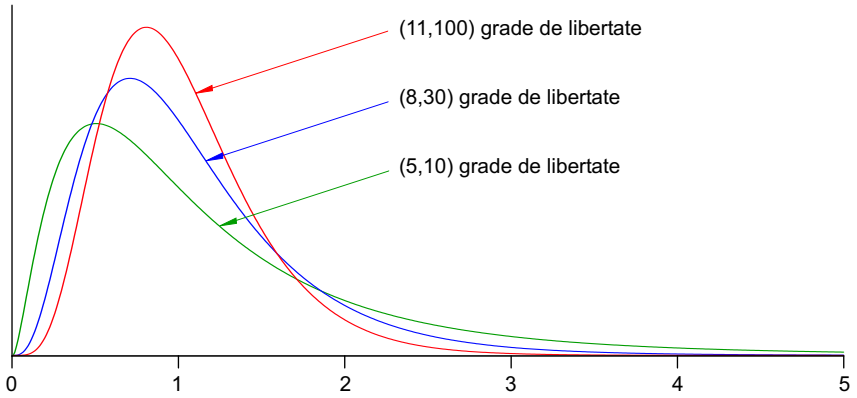
După cum am mai arătat, când eşantioanele sunt perfect similare (media lor este exact aceeaşi) atunci MPD va fi egal cu 0 iar F va fi egal de asemenea egal cu 0; la celălalt pol, când eşantioanele sunt total diferite (diferenţa dintre medii este maximă) atunci MPI va fi egal cu 0 iar F va fi egal cu $+\infty$. Astfel, o altă caracteristică a distribuţiei este aceea că valorile lui F sunt *non-negative* (mai mari sau egale cu zero).

Tot similar cu distribuţiile t şi χ^2, nu există o singură distribuţie F, ci o întreagă familie de distribuţii; dacă însă până acum forma distribuţiilor depindea de un

singur număr de grade de libertate, la analiza de varianţă forma distribuţiilor depinde de *o pereche* de grade de libertate.

Figura 1.2 arată trei asemenea distribuţii, unde primul număr reprezintă numărul de grade de libertate de la numărător (din MPD), iar cel de al doilea număr reprezintă numărul de grade de libertate de la numitor (din MPI).

Figura 1.2: Trei distribuţii F, cu:



După cum se poate vedea toate curbele sunt mai mult sau mai puţin alungite la dreapta, fiecare având un singur mod. Acestea sunt alte două caracteristici ale distribuţiei F: este *unimodală* şi *alungită la dreapta*. La număr mic al gradelor de libertate pentru numărător (cu alte cuvinte, pentru număr mic de grupuri), curba se apropie din ce în ce mai mult de axa verticală; spre exemplu, la o pereche (1,100) curba va fi chiar lipită de axa $0y$. Odată cu creşterea numărului de grade de libertate, cozile distribuţiilor se vor apropia din ce în ce mai mult de axa orizontală $0x$, însă nu o vor atinge decât la infinit.

Forma distribuţiei se modifică deci odată cu creşterea numărului de grade de libertate (atât la numărător cât şi la numitor), fiind din ce în ce mai puţin alungită la dreapta. În principiu, creşterea volumului total al eşantionului modifică distribuţia până la o formă relativ apropiată de distribuţia normală.

Rezumând, proprietăţile distribuţiei F sunt următoarele:

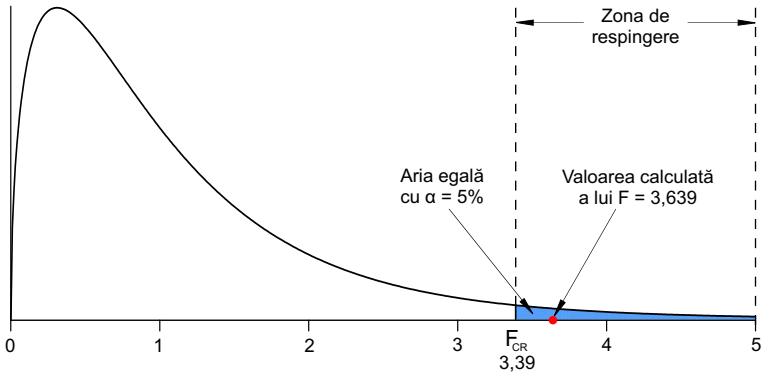
1. este continuă
2. este non-negativă
3. este uni-modală
4. este alungită la dreapta
5. aria de sub curbă este egală cu 1

Testul F este uni-direcţional, *doar* pe coada din dreapta; asta înseamnă că probabilitatea de eroare calculată va fi reprezentată ca o arie sub curbă numai în partea dreaptă. Există o singură valoare critică a lui F, în dreapta căreia se află aria de sub curbă corespunzătoare nivelului de semnificaţie ales.

Valoarea critică a lui F poate fi găsită cu ajutorul tabelelor de valori care pot fi găsite în anexele oricărui manual de statistică. Există mai multe tabele de valori, câte unul pentru fiecare nivel de semnificaţie clasic: 10%, 5%, 2,5%, 1% şi uneori chiar 0,1%.

Modul de citire a tabelelor este foarte simplu: a) se alege tabelul corespunzător nivelului de semnificaţie ales; b) se localizează numărul de grade de libertate de la numărător ($k-1$, de la MPD) pe orizontală, în partea de sus a tabelului; c) se localizează numărul de grade de libertate de la numitor ($n-k$, de la MPI) pe verticală, în partea din stânga a tabelului; d) la intersecţia dintre coloana şi linia identificate se află valoarea critică a lui F.

Figura 1.3: Probabilitatea de eroare de tipul I, valoarea critică şi valoarea calculată a lui F, pe o distribuţie cu o pereche (2, 27) grade de libertate



Pentru exemplul nostru, la un nivel de semnificaţie $\alpha = 5\%$, valoarea critică a lui F este aproximativ egală cu 3,35. Valoarea calculată a lui F este egală cu 3,639 şi este mai mare decât valoarea critică, intrând în zona de respingere a ipotezei de nul (colorată cu albastru).

1.7 Asumpţiile analizei de varianţă

Pentru a putea utiliza această analiză, trebuie să ne asigurăm că sunt îndeplinite următoarele condiţii/asumpţii:

1. Fiecare eşantion este extras dintr-o populaţie cu o distribuţie normală.
2. Populaţiile din care au fost extrase eşantioanele au aceeaşi varianţă (cu alte cuvinte, toate variază în aceeaşi măsură).
3. Eşantioanele sunt extrase în mod aleator şi independent.

Ca şi la testele t şi χ^2, prima asumpţie este legată de normalitatea distribuţiilor: populaţiile din care au fost extrase eşantioanele trebuie să aibă o distribuţie normală. Aşa cum se întâmplă mai întotdeauna în practică, însă, asumpţia de normalitate este rareori satisfăcută (cazurile în care populaţia are o distribuţie perfect normală sunt foarte rare, dacă nu inexistente).

Există totuşi o soluţie: violarea acestei asumpţi poate fi tolerată, dacă eşantionul este *suficient de mare* (deoarece în acest caz distribuţia de eşantionare nu mai depinde de forma distribuţiei în populaţie).

Pentru a verifica dacă distribuţiile sunt normale, se construieşte câte o histogramă a variabilei metrice pentru fiecare grup (atenţie însă: histogramele sunt relevante doar pentru eşantioane mari).

Asumpţia de bază a analizei de varianţă este cea a *omogenităţii varianţelor*: această a doua asumpţie este cea mai dezbătută de către specialişti. Unii dintre ei afirmă că, dacă varianţele în populaţiile din care provin eşantioanele nu sunt egale, atunci ANOVA nu poate fi aplicată. Alţii afirmă că acest test este irelevant, deoarece rezultatele lui sunt foarte puternic influenţate de forma distribuţiei în populaţie (testarea egalităţii dintre varianţe nu poate fi realizată *decât dacă* distribuţiile în populaţie sunt normale).

ANOVA este o analiză destul de robustă, chiar şi în cazul în care varianţele nu sunt egale; totul este ca diferenţa dintre varianţe să nu fie foarte mare (adică o varianţă să nu fie de câteva ori mai mare decât alta). Mai mult decât atât, analiza este şi mai robustă la încălcarea acestei asumpţii dacă eşantioanele sunt de mărime egală ($n_1 = n_2 = n_3$). În concluzie, este bine să evităm aplicarea acestei analize dacă eşantioanele sunt mici, au distribuţii puternic deplasate de la normalitate şi au varianţe în populaţie inegale; dacă eşantioanele au însă mărime egală, cu distribuţii moderat deplasate şi varianţe în populaţie moderat inegale, atunci putem aplica analiza cu încredere.

În ceea ce ne priveşte, vom proceda în mod similar cu testul `t`, unde există o variantă de formulă pentru cazul în care varianţele sunt egale (este vorba de cea clasică, predefinită în orice program de analiză statistică) şi o altă variantă de formulă pentru cazul în care varianţele nu sunt egale (testul Welch, care mai este denumit şi testul "robust" al egalităţii mediilor); decizia folosirii uneia sau alteia din variante se ia pe baza valorii lui `p` din testul Levene de testare a omogenităţii varianţelor.

Setul de ipoteze din acest test (pentru exemplul nostru particular cu trei eşantioane) este:

$$\begin{cases} \mathrm{H}_0: & \sigma_1 = \sigma_2 = \sigma_3 \\ \mathrm{H}_\mathrm{A}: & \text{cel puţin două varianţe sunt diferite} \end{cases}$$

În urma efectuării testului cu datele noastre, a fost obţinută o valoare a statisticii test `F` = 0,108 şi un `p` = 0,898. După cum se poate judeca din valoarea lui `p`, dovezile sunt zdrobitoare că varianţele sunt omogene (sunt aproape 90% şanse de a greşi afirmând contrariul), drept pentru care vom utiliza testul clasic.

În fine, a treia asumpţie arată că toate elementele eşantioanelor trebuie extrase în mod independent, utilizând o tehnică aleatoare. Un rol major îl are metodologia utilizată în cercetare, claritatea cu care a fost făcut instructajul dinaintea cercetării, corectitudinea cu care operatorii de teren aplică instrucţiunile primite etc. Cu cât controlăm mai bine toate aceste detalii, cu atât putem fi mai siguri pe rezultatele noastre. A extrage elemente în mod independent unele de altele înseamnă că între orice pereche de elemente din eşantion nu trebuie să fie nici o legătură (spre exemplu, doi respondenţi să nu fie rude).

1.8 Glosar de termeni

Analiza de varianţă - ANOVA (în engl. **AN**alisys **O**f **VA**riance). O tehnică statistică utilizată pentru a testa egalitatea dintre trei sau mai multe medii.

Distribuţia F (în engl. F Distribution). O familie de distribuţii de eşantionare folosite pentru a testa diferenţele dintre medii sau varianţe, a căror formă depinde de doi parametri (gradele de libertate de la numărător şi de la numitor).

Grade de libertate (în engl. Degrees of freedom). Număr de observaţii care pot fi alese în mod liber.

MPD - Varianţa dintre grupuri (în engl. MSB - Mean Square Between sau Between Group Variance). Medie a sumei pătratelor dintre grupuri, este o estimare a varianţei din populaţie care calculează variaţia mediilor de grupuri în jurul mediei generale, împărţind SPD la un număr de grade de libertate.

MPI - Varianţa în interiorul grupurilor (în engl. MSW - Mean Square Within sau Within Group Variance). Medie a sumei pătratelor din interiorul grupurilor, este o estimare a varianţei din populaţie care calculează variaţia din interiorul tuturor grupurilor (unde grupurile mai mari vor avea o pondere mai mare), împărţind SPI la un număr de grade de libertate.

SPD - Suma Pătratelor Dintre grupuri (în engl. SSB - Sum of Squares Between). O măsură a variaţiei dintre grupuri, calculată prin însumarea pătratelor distanţelor de la fiecare medie de grup la media generală.

SPI - Suma Pătratelor în Interiorul grupurilor (în engl. SSW - Sum of Squares Within). O măsură a variaţiei din interiorul tuturor grupurilor, calculată prin însumarea pătratelor distanţelor de la fiecare observaţie la media grupului de care aparţine.

STP - Suma Totală a Pătratelor (în engl. SST - Sum of Squares Total). Măsură a variaţiei totale, calculată ca sumă dintre SPD şi SPI.

Statistica F sau **Raportul F** (în engl. F statistic sau F ratio). Este un raport între cele două estimări ale varianţei din populaţie (MPD - variaţia explicată şi MPI - variaţia neexplicată). Dacă cele două estimări sunt aproximativ egale atunci raportul va fi egal cu 1 iar ipoteza de nul nu poate fi respinsă.

Variaţie explicată (în engl. Explained variation). Parte a variaţiei totale, explicată de diferenţele dintre grupuri.

Variaţie neexplicată sau **Eroare** (în engl. Unexplained variation sau Error). Parte a variaţiei totale care nu poate fi explicată şi care se datorează erorilor aleatoare ale observaţiilor în jurul mediilor de grup.

Variaţie totală (în engl. Total variation). Variaţia generală a tuturor observaţiilor din eşantion, egală cu variaţia explicată plus variaţia neexplicată.