Analiza datelor de marketing utilizand S.P.S.S.

- analiza diferentiala -

Analiza diferentiala a datelor

- Utilizata pentru stabilirea reprezentativitatii statistice a diferentelor constatate intre:
  - o valoare presupusa a unui indicator (ipoteza) si valoarea estimata la nivelul populatiei investigate;
  - doua sau mai multe variabile independente;
  - doua sau mai multe esantioane dependente (analiza transversala sau longitudinala).
- Utilizari frecvente:
  - testarea ipotezelor statistice;
  - testarea reprezentativitatii indicatorilor statistici;
  - testarea semnificatiei variatiei valorilor observate pentru doua sau mai multe variabile;
  - testarea semnificatiei variatiei valorilor observate pentru doua sau mai multe grupuri (esantioane);

Testarea ipotezelor statistice

- Exemple de ipoteze utilizate in marketing:
  - *In cinematografele bucurestene merg cel putin o data pe an 20% dintre locuitorii orasului;*
  - *Consumatorii frecventi si ocazionali ai unui produs (marca) au caracteristici psihografice diferite;*
  - *Imaginea publica a hotelului Howard Johnson este mai buna decat cea a hotelului Ibis.*

Testarea ipotezelor statistice

- Etape pentru testarea ipotezelor:
  1. Identificarea testelor statistice adecvate.
  2. Formularea ipotezei nule H_0 si a ipotezei alternative H_1.
  3. Alegerea unei probabilitati de garantare a rezultatelor.
  4. Calcularea indicatorului asociat testului statistic.
  5. Stabilirea ipotezei acceptate (nula sau alternative).
  6. Formularea unei concluzii logice in limbajul specific marketingului.

Testarea ipotezelor statistice

- Cunoscuta si sub denumirea de analiza diferentiala <u>univariata</u>.
  - Variabile **categoriale**: se utilizeaza testul χ^2 univariat;
  - Variabile **parametrice**: se utilizeaza testul Student univariat (in varianta t sau z, depinzand de marimea esantionului).

Testul χ^2 univariat

- Utilizat pentru variabilele categoriale.
  - *Exemplu: in Romania, 25% dintre consumatori prefera Dacia. In urma unei cercetari (sondaj) s-a constatat ca 33% dintre soferi se afla la volanul unui autoturism Dacia. Ipoteza este falsa sau corecta?*
    - H_0: NU exista diferente semnificative statistic intre cei doi parametrii.
    - H_1: exista diferente semnificative statistic intre cei doi parametrii.

Testul χ^2 univariat

- Valori asteptate (conform ipotezei):
  - *Conduc Dacia: 25%*
  - *Nu conduc Dacia: 75%*
- Valori observate (din sondaj):
  - *Conduc Dacia: 33%*
  - *Nu conduc Dacia: 67%*

Testul χ^2 univariat

- Indicatorul (calculat) al testului χ^2:

$$\chi_c^2 = \sum_{i=1}^{n} \frac{(O_i - A_i)^2}{A_i}$$

$$\chi_c^2 = \frac{(33-25)^2}{25} + \frac{(67-75)^2}{75} = 2{,}56 + 0{,}85 = 3{,}41$$

Testul χ^2 univariat

- Pentru o probabilitate de garantare a rezultatelor de 99%, valoarea tabelata a lui t univariat este de 6,635.
- Se observa ca $\chi_c^2 \leq \chi_t^2$ (3,41 < 6,635) => se accepta ipoteza nula (nu exista diferente semnificative statistic intre valorile prognozate si cele observate, deci ipoteza initiala a fost corecta!)

Testul Student univariat

- Utilizat pentru variabile parametrice (se poate calcula media), normal distribuite.
  - *Exemplu: venitul mediu in gospodariile celor care isi cumpara Dacia este de 2000 de lei lunar. In urma aceluiasi sondaj, am constatat ca venitul in cauza este de fapt de 1752 de lei. Este confirmata sau infirmata ipoteza initiala?*
    - H_0: NU exista diferente semnificative statistic intre valoarea din ipoteza si cea estimata la nivelul populatiei investigate, pe baza valorii observate in esantionul cercetat.
    - H_1: Exista diferente semnificative statistic intre valoarea din ipoteza si cea estimata la nivelul populatiei investigate, pe baza valorii observate in esantionul cercetat.

Testul Student univariat

- Valoarea calculata a testului:

$$t_c = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Testul Student univariat

- Pentru o dimensiune a esantionului de 1000 de persoane si o abaterea medie patratica de de 3315, avem t_c =2,36.
- **Gradele de libertate** asociate testului t univariat sunt n-1, in cazul de fata 999, iar probabilitate de garantare a rezultatelor α aleasa este de 95%. In acest caz gasim t_t = 1,64
- Interpretarea teoretica a testului Student:
  - $t_c \leq t_t$: se accepta ipoteza nula
  - $t_c > t_t$: se accepta ipoteza alternativa

Testul Student univariat

- t_c (2,36) > t_t (1,64) => se respinge ipoteza nula (exista diferente semnificative statistic intre valoarea ipotezei si cea estimata la nivelul populatiei, deci ipoteza formulata este gresita).

Analiza diferentiala bivariata

- Testele utilizate sunt alese in functie de modul de masurare al variabilelor, numarul de esantioane (grupuri) analizate si relatiile existente intre esantioane:
  - **Variabile nominale**:
    - grupuri (esantioane) independente: χ^2
    - grupuri (esantioane) dependente: χ^2 (varianta McNemar)
  - **Variabile ordinale (sau variabile interval tratate ca variabile ordinale)**:
    - 2 grupuri (esantioane) independente: **Mann-Whitney**, **Wald-Wolfowitz**;
    - 2 grupuri (esantioane) dependente: **Wilcoxon**;
    - 3 sau mai multe grupuri (esantioane): **Kruskal-Wallis**;
  - **Variabile proportionale**:
    - 2 grupuri (esantioane) independente: **testul Student pentru esantioane independente**;
    - 2 grupuri (esantioane) dependente: **testul Student pentru variabile dependente**;
    - 3 sau mai multe grupuri (esantioane): **ANOVA**;

Testul neparametric χ^2

- In varianta clasica, testul χ^2 presupune testarea unor variabile categoriale (de regula non-parametrice) si independenta esantioanelor analizate.
- Se bazeaza pe utilizarea tabelelor de contingenta.

Testul neparametric χ^2

- Preferinta pentru imbracaminte sport, in functie de statutul marital.

Prefera pantofii sport	Statut marital		Total
	Casatoriti	Necasatoriti	
Adesea	196	104	300
Rar	58	142	200
Total	254	246	500

- Valorile din tabelul de contingenta, rezultate in urma cercetarii, sunt denumite valori observate.

Testul neparametric χ^2

- Bazat pe ipotezele:
  - H_0: NU exista diferente semnificative intre cele doua variabile.
  - H_1: Exista diferente semnificative intre cele doua variabile.
- Valoarea calculata a testului este data de:

$$\chi_c^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - A_{ij})^2}{A_{ij}}$$

- Valorile asteptate sunt determinate conform distributiei (teoretice) χ^2 de formula:

$$A_{ij} = \frac{\sum_{i=1}^{r} O_{ij} \times \sum_{j=1}^{k} O_{ij}}{\sum_{i=1}^{r} \sum_{j=1}^{k} O_{ij}}$$

Testul neparametric χ^2

- Valoarea calculata χ_c^2 a testului este comparata cu valoarea tabelata χ_t^2 a acestuia, obtinuta in functie de probabilitatea de garantare a rezultatului si gradele de libertate asociate: (r-1)(k-1).
    - $\chi_c^2 \leq \chi_t^2$: se accepta ipoteza nula
    - $\chi_c^2 > \chi_t^2$: se accepta ipoteza alternativa
- Conditie:
    - Pentru mai mult de doua subesantioane independente trebuie ca frecventele $O_{ij} > 1$ si $O_{ij} < 5$ sa nu depaseasca 20% (celulele din tabelul de contingenta cu frecvente de aparitie diferita de zero si mai mica decat 5 sa nu depaseasca 20%).

Testul Fisher

- Inlocuieste testul χ^2 atunci cand dimensiunea esantionului n<20 si k=r=2 (variabile dihotomice);
- Tabelul de contingenta pentru k=r=2:

Prefera incaltamintea sport	Statut marital		**Total**
	Casatoriti	Necasatoriti	
Adesea	A	B	A+B
Rar	C	D	C+D
Total	A+C	B+D	N

Testul Fisher

- Testul probabilitatii exacte (Fisher) are aceiasi ipoteza nula:
  - H_0: NU exista diferente semnificative intre cele doua variabile;
  - H_1: Exista diferente semnificative intre cele doua variabile.

$$p = \frac{(A+B)!\,(C+D)!\,(A+C)!\,(B+D)!}{N!\,A!\,B!\,C!\,D!}$$

- Valoarea calculata p a testului se compara cu probabilitatea de garantare a rezultatului (ex.: 95%).
  - $p \leq 0{,}05$: se accepta ipoteza alternativa
  - $p > 0{,}05$: se accepta ipoteza nula

Testul Fisher (corectia Yates)

- Atunci cand dimensiunea esantionului n>20 si k=r=2 se utilizeaza corectia lui Yates a testului Fisher:

$$\chi_c^2 = \frac{N\left(|ad - bc| - \frac{N}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Testul McNemar

- Inlocuieste testul χ^2 atunci cand cele doua esantioane investigate sunt dependente (analiza longitudinala sau transversala).
- Testul McNemar are aceiasi ipoteza nula:
  - H_0: NU exista diferente semnificative intre cele doua variabile;
  - H_1: Exista diferente semnificative intre cele doua variabile.

$$\chi_c^2 = \frac{(|a-d|-1)^2}{a+d}$$

- a si d reprezinta frecventele subesantioanelor independente.
- Interpretarea este aceiasi ca si in cazul testului χ^2 :
  - $\chi_c^2 \le \chi_t^2$: se accepta ipoteza nula
  - $\chi_c^2 > \chi_t^2$: se accepta ipoteza alternativa

Testul Mann-Whitney

- Utilizat de preferinta pentru pentru identificarea diferentelor semnificative intre (doua) variabile ce provin din esantioane independente, masurate cu ajutorul scalei ordinale (*se poate utiliza insa si in cazul variabilelor proportionale*), distribuite normal.
- Ipotezele testului Mann-Whitney:
  - H_0: NU exista diferente semnificative intre cele doua variabile.
  - H_1: Cele doua variabile difera in mod semnificativ.
- Valoarea calculata a testului U este data de:

$$U_c^i = R_i - \frac{n_i\ (n_i + 1)}{2}, \text{unde } i \in \{1,2\}$$

Testul Mann-Whitney

- R_i reprezinta suma **rangurilor** asociate valorilor din esantionul i (primul sau al doilea).
- Pentru esantioane totale (n_1+n_2) mai mici de 30, valorile lui U_t sunt tabelate.
- Pentru esantioane de peste 30 de subiecti se utilizeaza testul Student pentru stabilirea semnificatiei statistice a testului U, dupa formula:

$$z_c = \frac{U - \frac{n_1 \times n_2}{2}}{\sigma_U}$$

unde:

$$\sigma_U = \sqrt{\frac{n_1 \times n_2 \ (n_1 + n_2 + 1)}{n_1 + n_2}}$$

Testul Mann-Whitney

- Interpretarea testului U pentru esantioane mai mici de 30 de subiecti:
  - $U_c \leq U_t$: se accepta ipoteza nula
  - $U_c > U_t$: se accepta ipoteza alternativa
- Interpretarea teoretica a testului U pentru esantioane mai mari de 30 de subiecti:
  - $z_c \leq z_t$: se accepta ipoteza nula
  - $z_c > z_t$: se accepta ipoteza alternativa

Testul Mann-Whitney

- Presupunand ca Esop nu a fost foarte satisfacut de experimentul sau clasic, in care o broasca testoasa intrece un iepure si repeta experiementul cu 6 iepuri si 6 broaste testoase. “Clasamentul” se afla in tabelul de mai jos:

I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
T	I	I	I	I	I	T	T	T	T	T	I

- Suma rangurilor R_1 asociate testoaselor este:

$$1+7+8+9+10+11 = 46$$

Testul Mann-Whitney

$$U_c^1 = 46 - \frac{6 \quad (6+1)}{2} = 25$$

- Din tabelul asociat testului Mann-Whitney gasim ca U_t (pentru n_1=6, n_2=6 si α=0,05) = 5, deci putem constata ca U_c > U_t => vom accepta ipoteza alternativa (exista diferente semnificative intre comportamentul in concurs al broastelor testoase si al iepurilor, dat de suma rangurilor, mai exact 46 pentru testoase si 25 pentru iepuri)

Testul Wilcoxon

- Testul Wilcoxon este un test non-parametric bivariat utilizat pentru identificarea semnificatiei statistice a diferentelor identificate pentru variabile provenite din esantioane dependente (masuratori repetate sau variabile masurate ale acelorasi respondenti), masurate cu ajutorul scalelor ordinale, indiferent de tipul distributiei.
  - *Exemplu: existenta unor diferente semnificative statistic intre perceptiile asupra a doua marci diferite (utilizand scala Likert) sau pentru perceptia asupra imaginii berii Redd's inainte si dupa realizarea unei campanii promotionale.*

Testul Wilcoxon

- Ipotezele testului Wilcoxon:
  - H_0: NU exista diferente semnificative intre cele doua variabile.
  - H_1: Cele doua variabile difera in mod semnificativ.
- Pentru calculul statisticii W^+, asociata testului Wilcoxon, se ordoneaza toate valorile observate, se calculeaza diferentele observate w_i, aceste diferente sunt ordonate in functie de marime, fiecareia fiind ulterior asociat un rang R_i pe baza pozitiei in aceasta serie de diferente:

$$w_i = y_i - x_i \qquad R_i = \text{rangul}\left|w_i\right|$$

Testul Wilcoxon

- De asemenea, pentru calculul W^+ se utilizeaza o functie indicator, Φ_i:

$$\varphi_i = I \quad (w_i > 0)$$

- Valoarea W^+ este data de:

$$W^+ = \sum_{i=1}^{n} \varphi_i R_i$$

  - Sustinerea (sau respingerea) ipotezei nule se bazeaza pe probabilitatea de aparitie a valorii W^+, data de tabele statistice asociate testului (pentru n de maxim 30 de respondenti) sau estimata cu ajutorul testului Student.

Testul Wilcoxon

- Utilizand scala Likert pentru identificarea disponibilitatii respondentilor de a cumpara berea Redd's, masurata inainte si dupa expunerea la un spot de promovare a produsului, au fost inregistrate urmatoarele valori (5 = sigur da; 4 = probabil da, 3 = indiferent, 2 = probabil nu; 1 = sigur nu):

Respondent	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Inainte	5	3	1	5	2	4	4	3	2	1	1	5	4	2	1
Dupa	5	4	2	3	5	5	4	3	1	4	4	5	3	2	5
Diferente (w_i)	0	-1	-1	2	-3	-1	0	0	1	-3	-3	0	1	0	-4
Ranguri R_i	-	3	3	6	8	3	-	-	3	8	8	-	3	-	10

Testul Wilcoxon

- Insumand rangurile pozitive R_i din tabelul anterior obtinem $W^{+}=12$, careia ii este asociata o probabilitate p(12)=0,002136 (aleasa pentru n=15 si α=0,05), mai mica decat 0,05 – pragul de sustinere al ipotezei nule in textul Wilcoxon, deci se poate concluziona ca ipoteza nula este acceptata (este respinsa ipoteza alternativa) => cele doua seturi de date NU difera in mod semnificativ (spotul publicitar NU a schimbat atitudinea respondentilor fata de marca Redd's).
- Pentru esantioane dependente de peste 30 de respondenti se utilizeaza:

$$z_c = \frac{W^{+} - 0{,}05}{\sigma_W} \qquad \sigma_W = \sqrt{\frac{n\ (n+1)\ (2n+1)}{2n}}$$

Testul Student bivariat

- Utilizat pentru stabilirea semnificatiei satistice a diferentelor constatate intre doua esantioane (*dependente* sau *independente*) sau variatia a doua variabile, masurate pe scala proportionala.
  - *Exemplu: persoanele de sex masculin si feminin au un comportament diferit in utilizarea Internetului (numarul de ore de utilizare saptamanale)? Persoanele cu venit mare au un procent mai ridicat de "loialisti" fata de marca decat persoanele cu venit scazut?*
- *Observatii:*
  - *Analiza este realizata diferentiat pentru* medii *si procente.*
  - Analiza este realizata *diferentiat* in cazul esantioanelor *independente*, in functie de existenta unor diferente (semnificative statistic) intre dispersiile celor doua grupuri.

Testul Student bivariat

- Bazat pe ipotezele
  - H_0: NU exista diferente semnificative statistic intre (media) celor doua esantioane investigate.
  - H_1: Exista diferente semnificative statistic intre (mediile) celor doua esantioane investigate.
- In cazul esantioanelor independente, se utilizeaza testul F (varianta Levine) pentru stabilirea asocierii dintre dispersiile celor doua grupuri (in anumite cazuri poate fi folosit si testul Kolmogorov-Smirnov).

Testul Student bivariat

- Ipotezele testului F:
  - H_0: NU exista diferente semnificative statistic intre dispersiile celor doua esantioane investigate.
  - H_1: cele doua esantioane inregistreaza diferente ale valorilor observate semnificative statistic.
- Valoarea testului F:

$$F_c = \frac{\sigma_1^2}{\sigma_2^2}$$

Testul Student bivariat

- Gradele de libertate asociate testului F sunt n_1-1 si n_2-1, iar probabilitate de garantare a rezultatelor α este aleasa, in functie de nevoile analizei.
- Daca probabilitatea asociata testului F_t (data de gradele de libertate si probabilitatea de garantare a rezultatelor) este mai mare decat cea asociata F_c atunci se accepta H_1 (cele doua esantioane au dispersii diferite), altfel se accepta H_0 (dispersiile celor doua esantioane independente sunt asemanatoare).
- Pentru esantioane independente (medii distincte) formula testului t (z in esantioane de peste 30 de respondenti) este:

$$z_c = \frac{\overline{x}_1 - \overline{x}_2}{s_{\overline{x}_1 - \overline{x}_2}}$$

Testul Student bivariat

- Abaterea standard asociata dispersiei, pentru esantioane independente, cu dispersii diferite semnificativ:

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

- Abaterea standard asociata dispersiei, pentru esantioane independente, cu dispersii asemanatoare:

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Testul Student bivariat

- Gradele de libertate asociate testului t bivariat (esantioane independente) sunt n_1+n_2-2 si probabilitate de garantare a rezultatelor α.
- Interpretarea teoretica a testului Student:
    - $t_c \leq t_t$: se accepta ipoteza nula
    - $t_c > t_t$: se accepta ipoteza alternativa
- Analiza difera in functie de dispersiile asociate celor doua esantioane utilizate

Testul Student bivariat

- Numarul de ore petrecute saptamanal utilizand resurse din Internet

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Ore Internet	14	2	3	3	13	6	2	6	6	15	3	4	9	8	5
Sex	1	2	2	2	1	2	2	2	2	1	2	2	1	1	1
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Ore Internet	3	9	4	14	6	9	5	2	15	6	13	4	2	4	3
Sex	2	1	1	1	2	1	1	2	1	2	1	2	2	1	1

Testul Student bivariat

Sex	Nr. de respondenti	Media (orelor de navigatie saptamanale)	Eroarea standard asociata mediei
Masculin	15	9.33	1.14
Feminin	15	3.87	0.44

$F_c = 15{,}507 > F_{14,14,95\%} = 2{,}46$ => se accepta ipoteza alternativa (dispersiile celor doua esantioane sunt semnificativ diferite)

$t_c = 4{,}492 > t_{28,\ 95\%} = 1{,}701$ => se accepta ipoteza alternativa (exista diferente semnificative intre gradul de utilizare a Internetului pentru barbati si femei)

Testul Student bivariat

- Testul t bivariat (pentru esantioane independente) se poate folosi si pentru alti indicatori (ex.: procente).

$$z_c = \frac{p_1 - p_2}{s_{p_1 - p_2}}$$

$$s_{p_1 - p_2} = \sqrt{\frac{p_1\ (1 - p_1\ )}{n_1} + \frac{p_2\ (1 - p_2\ )}{n_2}}$$

Testul Student bivariat (esantioane dependente)

- Testul t bivariat pentru esantioane dependente (masuratori repetate sau variabile masurate ale acelorasi respondenti).
  - *Exemplu: existenta unor diferente semnificative statistic intre perceptiile asupra a doua marci diferite (utilizand scala Stapel) sau pentru perceptia asupra unei marci la doua momente diferite (inainte si dupa efectuarea unor activitati promotionale?*

Testul Student bivariat (esantioane dependente)

- Testul t bivariat pentru esantioane dependente

$$z_c = \frac{\overline{D} - \mu_D}{s_{\overline{D}}}$$

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

$$s_D = \frac{\sqrt{\frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1}}}{\sqrt{n}}$$

Analiza Variatiei (ANOVA)

- In ciuda denumirii, reprezinta tot un test statistic, utilizat pentru stabilirea semnificatiei satistice a diferentelor constatate intre trei sau mai multe esantioane (*dependente* sau *independente*), masurate pe o scala proportionala.
- Echivalentul testului Stundent pentru mai mult de doua esantioane
  - *Exemple: utilizarea Internetului (numarul de ore de utilizare saptamanale) difera in functie de nivelul de educatie al persoanelor investigate (gimnazial, liceal, universitar, post-universitar)? Categoriile (intervalele) de varsta influenteaza semnificativ nivelul salarial al respondentilor?*

Analiza Variatiei (ANOVA)

- Utilizeaza:
  - *o variabila de grupare X (ce determina subgrupurile), denumita si variabila independenta;*
  - o variabila analizata (dependenta), masurata pe scala proportionala;
- Variabila dependenta este subdivizata in c subesantioane (grupuri), de dimensiuni n_1, n_2,…n_c.
- In analiza diferentelor constatate intre mediile subgrupurilor 1…c, ANOVA utilizeaza notiunea de **descompunere a variatiei totale**, in **variatie interna** (in interiorul acestor grupuri) si **variatie externa** (diferenta constatata intre grupuri).

Analiza Variatiei (ANOVA)

- Variatia totala:

$$V_T = V_I + V_E$$

$$V_T = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{ij} - \overline{x})^2$$

$$V_E = \sum_{j=1}^{c} (\overline{x}_j - \overline{x})^2$$

$$V_I = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{ij} - \overline{x}_j)^2$$

Analiza Variatiei (ANOVA)

- Gradele de libertate asociate:
  - *variatia totala: n-1;*
  - variatia interna: n-c;
  - variatia externa: c-1;
- Magnitudinea (importanta) variatiilor se calculeaza cu ajutorul unui indicator, denumit media patratica η:
  - *Media patratica interna:*

$$\eta^2_{interna} = \frac{V_I}{n-c}$$

  - *Media patratica externa:*

$$\eta^2_{externa} = \frac{V_E}{c-1}$$

Analiza Variatiei (ANOVA)

- Ipotezele asociate ANOVA:
  - *NU exista o diferenta semnificativa statistic intre (mediile) grupurile analizate*;
  - grupurile investigate (mediile lor) difera in mod semnificativ;
- Ipotezele sunt acceptate sau respinse in functie de valoarea coeficientului F asociat ANOVA:

$$F_c = \frac{\eta^2_{externa}}{\eta^2_{interna}}$$

Analiza Variatiei (ANOVA)

- Valorile teoretice ale testului F se regasesc in tabele, indexate pe baza probabilitatii de garantare a rezultatelor (1-α) si gradele de libertate interne (n-1) si externe (c-1).
- Interpretarea teoretica a testului F (ANOVA):
  - $F_c \leq F_t$: se accepta ipoteza nula
  - $F_c > F_t$: se accepta ipoteza alternativa
  - Exemplu: Zone Records doreste sa lanseze pe piata noul album Holograf si, pentru inceput, produce 10000 de cópii. *Trimite cate 2000 de exemplare in cele 5 depozite regionale sau tine seama de vanzarile celorlalte grupuri de rock din fiecare regiune din ultimul an*?

Analiza Variatiei (ANOVA)

- Date istorice despre vanzarile de muzica rock:

Grup	Bucuresti	Constanta	Iasi	Cluj	Timisoara	Total
Iris	3000	800	1000	1500	1000	7300
Bere gratis	750	200	1200	2000	1500	5650
O.C.S.	1250	400	300	1400	1000	4350
Sarmalele reci	2000	500	600	400	800	4300
Celelalte cuvinte	1000	400	100	200	700	2400
Total	9000	2300	3200	5500	5000	25000
Medii partiale	1800	460	640	1100	1000	**1000**

Analiza Variatiei (ANOVA)

- n = 5x5=25 de observatii
- c=r=5 ($n_1=n_2=n_3=n_4=n_5=5$)

$$V_E = \sum_{j=1}^{c} (\bar{x}_j - \bar{x})^2 = 1071200$$

$$V_I = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 = 3525000 + 198800 + 852000 + 2360000 + 380000 = 39040800$$

$$F_c = \frac{V_E \ (n-c)}{V_I \ (c-1)} = \frac{1071200 \quad (25 - 5)}{39040800 \quad (4 - 1)} = 1,827$$

Analiza Variatiei (ANOVA)

- $F_c = 1{,}827 < F_{t\ (5,5,\alpha=0{,}05)} = 5{,}05$ => se accepta ipoteza nula (mediile subesantioanelor nu difera in mod semnificativ).
- Cum se distribuie CD-ul celor de la IRIS?

Testul Levene

- Un test bivariat, pentru stabilirea gradului de asemanare intre variatiile a doua esantioane (*dependente* sau *independente*), masurate pe o scala categoriala sau continua, normal distribuite.
- Ipotezele asociate testului Levene:
    - *NU exista o diferenta semnificativa statistic intre dispersiile grupurilor analizate (dispersiile sunt asemanatoare – avem o relatie de **homoscedasticitate**);*
    - Disersiile grupurilor investigate sunt semnificativ diferite (prezinta o relatie de **heteroscedasticitate**);

Testul Levene

- Indicatorul testului este denumit Levene F sau W si se calculeaza conform formulei:

$$W_c = \frac{(n-c)\sum_{j=1}^{c} n_j (\overline{D}_j - \overline{D})^2}{(c-1)\sum_{j=1}^{c}\sum_{i=1}^{n_j} (D_{ij} - \overline{D}_i)^2}$$

- unde:

$$D_{ij} = \left| y_{ij} - \overline{y}_j \right|$$

Testul Levene

- Valorile teoretice ale testului Levene se regasesc in tabele, indexate pe baza probabilitatii de garantare a rezultatelor (1-α) si gradele de libertate (n-c).
- Interpretarea teoretica a testului Levene:
  - $F_c \leq F_t$: se accepta ipoteza nula (relatia este homoscedastiva)
  - $F_c > F_t$: se accepta ipoteza alternativa (relatia este heteroscedastiva)

Testul Kruskal-Wallis

- utilizat pentru stabilirea semnificatiei satistice a diferentelor constatate intre trei sau mai multe esantioane (*dependente* sau *independente*), masurate pe o scala ordinala, normal distribuite si homoscedastice.
- Kruskal-Wallis este echivalentul testelor Mann-Whitney si Wilcoxon pentru mai mult de doua esantioane.
  - *Exemple: identificarea gradului in care nivelul de educatie influenteaza preferinta pentru un anumit produs, masurat pe o scala categoriala; Stabilirea gradului in care gradul de loialitate al respondentilor este influentat de perceptia imaginii unui produs?*

Testul Kruskal-Wallis

- Ipotezele asociate testului Kruskal-Wallis:
  - *NU exista o diferenta semnificativa statistic intre (medianele) grupurile analizate;*
  - Grupurile investigate (medianele) difera in mod semnificativ;
- Gradele de libertate asociate K sunt *c-1* (c reprezentand numarul de grupuri determinat de variabila de grupare asupra variabilei independente).

Testul Kruskal-Wallis

- Ipotezele sunt acceptate sau respinse in functie de valoarea coeficientului K asociat testului:

$$K_c = (n-1)\frac{\sum_{j=1}^{c} n_j (\bar{r}_j - \bar{r})^2}{\sum_{j=1}^{c}\sum_{i=1}^{n_j} (r_{ij} - \bar{r})^2}$$

- unde: - r_{ij} reprezinta rangul observatiei i din grupul j;
  - r_j – media subesantionului j;
  - n_j – dimensiunea subesantionului j;
  - c – numarul de grupuri (c>2)

Testul Kruskal-Wallis

- Interpretarea teoretica a testului Kruskal-Wallis se bazeaza pe valorile tabelate ale **testului χ^2**, pentru *c-1* grade de libertate si o probabilitate de garantare a rezultatelor de α:
  - $K_c \leq \chi^2_t$: se accepta ipoteza nula (grupurile nu sunt semnificativ diferite);
  - $K_c > \chi^2_t$: se accepta ipoteza alternativa (grupurile au comportamente diferite).
- Observatii:
  – In cazul variabilelor nominale se utilizeaza testul χ^2 , indiferent de numarul subesantioanelor;
  – Testul K este mai exact decat χ^2 in cazul variabilelor ordinale, utilizand rangurile, spre deosebire de χ^2, care utilizeaza frecvente de aparitie.

Testul Kruskal-Wallis

- Exemplu: In urma unor focus grupuri realizate pentru identificarea perceptiei consumatorilor potentiali pentru berea Redd's, inainte de lansarea acesteia pe piata, au fost stranse date despre nivelul de educatie (liceu, universitar, post-universitar) al respondentilor, ca si asupra perceptiei asupra gustului, pretului si imaginii produsului, folosindu-se scala Stapel (note de la 1 la 10, 10 reprezentand valoarea maxima). Datele stranse se regasesc in tabelul urmator.

Testul Kruskal-Wallis

- Pentru fiecare respondent, valorile celor 3 indicatori ai perceptiei (gust, pret si imagine) sunt agregati utilizandu-se media algebrica.

	Liceu	Facultate	Master/Doctor
1	6.4	2.5	1.3
2	6.8	3.7	4.1
3	7.2	4.9	4.9
4	8.3	5.4	5.2
5	8.4	5.9	5.5
6	9.1	8.1	8.2
7	9.4	8.2	
8	9.7		
Medie	8.2	5.5	4.9

Testul Kruskal-Wallis

- Valorile sunt agregate intr-o singura variabila, de dimensiunea n=21, iar apoi sunt atribuite ranguri, dupa sistemul explicat pentru testul Mann-Whitney:

	Liceu	Facultate	Master/Doctor
1	11	2	1
2	12	3	4
3	13	5.5	5.5
4	17	8	7
5	18	10	9
6	19	14	15.5
7	20	15.5	
8	21		
Suma rangurilor	131	58	42
Medie	16.4	8.3	7

Testul Kruskal-Wallis

- Suma tuturor rangurilor este 231, cu o medie de 11 (231/21). Tabelul patratelor diferentelor de rang este:

	Liceu	Facultate	Master/Doctor
1	0	81	100
2	1	64	49
3	4	30.25	30.25
4	36	9	16
5	49	1	4
6	64	9	20.25
7	81	20.25	
8	100		
Suma rangurilor	29.16	7.29	16

Testul Kruskal-Wallis

- Suma patratelor diferentelor intre rangurile observate si media rangurilor este 769, in timp ce patratul diferentelor dintre rangurile mediie ale grupurilor si media generala a rangurilor este 52,45. In acest fel, putem calcula:

$$K_c = (n-1)\frac{\sum_{j=1}^{c} n_j (\bar{r}_j - \bar{r})^2}{\sum_{j=1}^{c}\sum_{i=1}^{n_j} (r_{ij} - \bar{r})^2} = 20\frac{769}{52.45} = 293{,}23$$

  - Observam ca $K_c = 293{,}23 > \chi^2_t = 5{,}991$, calculat pentru 3-1 grade de libertate si un α=0,05, deci acceptam impoteza alternativa, concluzionand ca nivelul de educatie influenteaza semnificativ modul in care este perceputa marca de bere Redd's

Analiza CoVariatiei (ANCOVA)

- Reprezinta un test statistic, utilizat pentru stabilirea semnificatiei satistice a diferentelor constatate intre trei sau mai multe esantioane (*dependente* sau *independente*), masurate pe o scala categoriala sau continua, normal distribuite si homoscedastice.
  - *Exemple: utilizarea Internetului (tipuri de abonament) difera in functie de nivelul de educatie al persoanelor investigate (gimnazial, liceal, universitar, post-universitar)? Cum este influentata intentia de cumparare pentru un produs, la nivelul unor grupuri distincte, de catre expunerea la instrumente promotionale distincte, in conditiile in care respondentii cunosteau deja produsul?*

Analiza CoVariatiei (ANCOVA)

- ANCOVA testeaza in plus (fata de ANOVA) efecte ale covariantei (influenta unor variabile independente suplimentare) variabilei dependente.
- CoVarianta este utilizata pentru izolarea efectelor altor variabile indepentente (covariante) asupra variabilei dependente investigate.
- Variabilele independente suplimentare sunt denumite variabile de control.

Analiza CoVariatiei (ANCOVA)

- Variabila dependenta este subdivizata in c subesantioane (grupuri), de dimensiuni n_1, $n_2, \ldots n_c$.
- Covariatia totala a subesantioanelor este descompusa in covariatie interna (in interiorul acestor grupuri) si covariatie externa (diferenta constatata intre grupuri).

Analiza CoVariatiei (ANCOVA)

- Variatia totala: $V_T = V_I + V_E$

$$V_T = \sum_{j=1}^{c}\sum_{i=1}^{n_j} y_{ij}^2 - \frac{\left(\sum_{j=1}^{c}\sum_{i=1}^{n_j} y_{ij}\right)^2}{n}$$

$$V_E = \sum_{j=1}^{c}\sum_{i=1}^{n}(\bar{y}_{ij} - \bar{y}_j)(\bar{x}_{ij} - \bar{x}_j)$$

$$V_I = n\sum_{j=1}^{c}(\bar{y}_j - \bar{y})(\bar{x}_j - \bar{x})$$

Analiza CoVariatiei (ANCOVA)

- CoVariatia este data de:

$$COV_E = \sum_{j=1}^{c}\sum_{i=1}^{n_j} x_{ij}^2 y_{ij}^2 - \frac{\sum_{j=1}^{c}\sum_{i=1}^{n_j} x_{ij}^2 \times \sum_{j=1}^{c}\sum_{i=1}^{n_j} y_{ij}^2}{n}$$

$$COV_I = \sum_{j=1}^{c} \left(\sum_{i=1}^{n_j} x_{ij} y_{ij} - \frac{\sum_{i=1}^{n_j} x_{ij} y_{ij}}{n_j} \right)$$

Analiza CoVariatiei (ANCOVA)

- Gradele de libertate asociate (fiecare variabila de control suplimentara duce la pierderea unui grad de libertate):
  - variatia interna: n-c-1;
  - variatia externa: c-1;
- Coeficientul de determinare (indica in ce masura variatia din interiorul/exteriorul grupurilor identificate la nivelul variabilei dependente este explicata de variabila de grupare):
  - *externa (intre grupuri):* $r^2_{extern} = \frac{COV_E^2}{V_T V_E}$
  - *interna (in interiorul grupurilor):* $r^2_{intern} = \frac{COV_I^2}{V_T V_I}$

Analiza CoVariatiei (ANCOVA)

- Ipotezele asociate ANCOVA:
  - *NU exista o diferenta semnificativa statistic intre (mediile) grupurile analizate*;
  - grupurile investigate (mediile lor) difera in mod semnificativ;
- Ipotezele sunt acceptate sau respinse in functie de valoarea coeficientului F asociat ANCOVA:

$$F_c = \frac{V_E \,(c-1)}{V_I \,(n-c-1)}$$

Analiza CoVariatiei (ANCOVA)

- Interpretarea testului F se face la fel ca in cazul ANOVA, prin identificarea valorilor tabelate, indexate pe baza probabilitatii de garantare a rezultatelor (1-α) si gradele de libertate interne (n-1) si gradele de libertate externe (c-1).
- Interpretarea teoretica a testului F (ANCOVA):
    - $F_c \leq F_t$: se accepta ipoteza nula
    - $F_c > F_t$: se accepta ipoteza alternativa

Analiza CoVariatiei (ANCOVA)

- Exemplu: Pentru cursul de Analiza Datelor de Marketing utilizand SPSS avem 4 manuale alternative. Pentru a testa care dintre ele este mai util studentilor, am oferit cate un manual fiecarei grupe. Am administrat un examen comun, cu 25 de intrebari, tuturor celor 4 grupe, iar apoi am prelevat esantioane formate din 10 studenti din fiecare grupa, pentru a determina daca exista diferente semnificative in pregatirea acestora.

Analiza CoVariatiei (ANCOVA)

- Raspunsuri corecte la examen, pe baza unor manuale diferite

Nota la SPSS	1	2	3	4	5	6	7	8	9	10	Total	Medii partiale
Grupa 1	12	15	14	14	18	18	16	14	19	19	159	15,9
Grupa 2	13	16	15	16	19	17	19	23	19	22	179	17,9
Grupa 3	14	16	18	20	18	19	22	21	23	20	191	19,1
Grupa 4	15	16	13	15	19	17	20	18	20	21	174	17,4

- Media generala a raspunsurilor corecte: 17,57

Analiza CoVariatiei (ANCOVA)

- n = 4 x 10 = 40 de observatii
- c = 4, iar r = 10

$$V_E = \sum_{j=1}^{c} (\bar{x}_j - \bar{x})^2 = 5,2675$$

$$V_I = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 = 54,9 + 86,9 + 66,9 + 62,4 = 271,1$$

$$F_c = \frac{V_E \ (n - c)}{V_I \ (c - 1)} = \frac{5,2675 \ (40 - 10)}{271,1 \ (10 - 1)} = 0,0648$$

Analiza CoVariatiei (ANCOVA)

- $F_c = 0{,}0648 < F_{t\ (39,9,\alpha=0,05)} = 2{,}84$ => se accepta ipoteza nula (mediile subesantioanelor NU difera in mod semnificativ) => dintre cele 4 grupe, nu exista cel putin doua ale caror masteranzi au o pregatire semnificativ diferita la Analiza Datelor de Marketing Utilizand SPSS (ex.: grupa 1 a raspuns corect, in medie, la 16 intrebari, iar membrii grupei 3 au raspuns corect, in medie, la 19 intrebari, insa aceasta diferenta nu este semnificativa statistic, data fiind dimensiunea esantioanelor utilizate).
- Putem concluziona ca nu conteaza ce manual voi recomanda anul viitor?

Analiza CoVariatiei (ANCOVA)

- Dupa cum stiti, la Marketing Strategic studentii sunt ordonati in diferite grupe in functie de facultatile absolvite, deci este teoretic posibil ca unii dintre ei sa aiba o pregatire anterioare in domeniul analizei datelor, ceea ce ar afecta acuratetea testului efectuat.
- Pregatirea anterioare poate fi estimata prin intermediul notei la Metode si Modele in Marketing, de pe primul semestru, care presupunea cunostinte in aproximativ acelasi domeniu.

Analiza CoVariatiei (ANCOVA)

- Raspunsuri corecte la examen, pentru grupe care s-au pregatit cu manuale diferite, incluzand nota la Metode si Modele in Marketing.

		1	2	3	4	5	6	7	8	9	10	Total	Medii partiale
Grupa 1	SPSS	12	15	14	14	18	18	16	14	19	19	159	15,9
	Modelare	*5*	*5*	*6*	*7*	*7*	*8*	*8*	*9*	*9*	*10*	*74*	*7,4*
Grupa 2	SPSS	13	16	15	16	19	17	19	23	19	22	179	17,9
	Modelare	*4*	*4*	*5*	*6*	*6*	*8*	*8*	*9*	*10*	*10*	*70*	*7*
Grupa 3	SPSS	14	16	18	20	18	19	22	21	23	20	191	19,1
	Modelare	*4*	*4*	*6*	*6*	*7*	*8*	*8*	*9*	*10*	*10*	*72*	*7,2*
Grupa 4	SPSS	15	16	13	15	19	17	20	18	20	21	174	17,4
	Modelare	*4*	*5*	*5*	*6*	*6*	*7*	*7*	*9*	*9*	*10*	*68*	*6,8*

Analiza CoVariatiei (ANCOVA)

- Analiza covariatiei:

$$V_E = \sum_{j=1}^{c}\sum_{i=1}^{n} (y_{ij} - \overline{y}_j)(x_{ij} - \overline{x}_j) = 161$$

$$V_I = n\sum_{j=1}^{c} (\overline{y}_j - \overline{y}) \quad (\overline{x}_j - \overline{x}) = 3,3$$

$$F_c = \frac{V_E\ (c-1)}{V_I\ (n-c-1)} = \frac{161 \times\ (9-1)}{3,3\ \ (40-9-1)} = 13,1$$

Analiza CoVariatiei (ANCOVA)

- $F_c = 13{,}1 > F_{t\ (39,9,\alpha=0{,}05)} = 2{,}84$ => se accepta ipoteza alternativa (mediile subesantioanelor difera in mod semnificativ) => exista diferente semnificative intre contributiile la pregatirea studentilor a celor 4 manuale utilizate!

ACADEMIA DE STUDII ECONOMICE
DOCENDO DISCIMUS
CATEDRA DE MARKETING

Academia de Studii Economice