Intervalul de încredere
Inferența statistică
Testarea distribuției unui set de date
Analiza corelației

Intervalul de încredere

Cuprins

- Definiție. Scop
- Interpretare
- Intervalul de încredere pentru medie
- Intervalul de încredere pentru frecvență

De ce intervalul de încredere?

- Estimarea punctuală
  - = o valoare pentru parametrul teoretic estimat
  - Influențată de fluctuațiilor de eşantionare
  - poate fi la o mare distanță de valoarea reală a parametrului estimat
- Este recomandabil să se estimeze un parametru teoretic nu printr-o singură valoare ci printr-un interval, numit **interval de încredere** (în care să se poată afirma că parametrul estimat se găseşte cu o probabilitate ridicată).

Definiție

- Un şi de valori al unui estimator de interes calculat astfel încât pentru o probabilitate de eroare aleasă să includă valorile adevărate ale variabilei.
- **P[valoarea critică inferioară < estimatorul < valoarea critică superioară] = 1-α**
  - unde α = nivelul de semnificație
- Intervalul definit de valorile critice va cuprinde estimatorul populației cu o probabilitate de 1-α
- Se aplică în cazul variabilelor distribuite normal!

Interpretare

- Dacă intervalul de încredere pentru diferența dintre o medie observată şi una teoretică cuprinde valoarea 0, datele sunt compatibile cu o diferență a mediei populației egală cu 0.
- Dacă intervalul de încredere pentru diferența dintre o medie observată şi una teoretică nu cuprinde valoarea 0, datele nu sunt compatibile cu egalitatea mediilor populației.

Intervalul de încredere

- Se calculează în funcție de:
  - Talia eşantionului sau a populației
  - Variabila de studiat (calitativă, cantitativă)
- Formula de calcul cuprinde 2 părți:
  - Un estimator al calității eşantionului pe baza căruia estimatorul populației s-a calculat (eroarea standard)
  - Gradul de încredere (confidență) al intervalului specificat (scorul Z_α)
- Cel mai frecvent utilizat este intervalul de încredere pentru **medie**

Intervalul de încredere pentru medie

- Eroarea standard a mediei este egală cu deviația standard împărțită la radicalul volumului eşantionului
  - Dacă deviația standard este mare, şansa de eroare în estimator este mare
  - Dacă volumul eşantionului este mare, şansa erorii în estimator este mică.

$$\left[\overline{X} - Z_\alpha \frac{s}{\sqrt{n}}, \overline{X} + Z_\alpha \frac{s}{\sqrt{n}}\right]$$

Intervalul de încredere pentru medie

- Scorul Z este scorul distribuției normale de medie 0 şi deviație standard de 1. Orice distribuție poate fi transformată în scorul Z utilizând formula:

$$Z = \left(X - \overline{X}\right)/s$$

  - Scorul pozitiv este mai mare decât media
  - Scorul negativ este mai mic decât media
  - **Pentru intervalul de confidență de 95%: $Z_{5\%} = 1,96$**
  - Pentru intervalul de confidență de 99%: $Z_{1\%} = 2,58$

$$\left[\overline{X} - Z_\alpha \frac{s}{\sqrt{n}}, \overline{X} + Z_\alpha \frac{s}{\sqrt{n}}\right]$$

Intervalul de încredere pentru medie

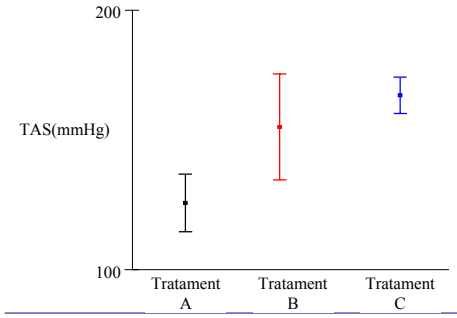
- Media glicemiei la un eşantion de 121 pacienți este de 105 iar variația de 36. Care este intervalul de încredere al mediei glicemiei în populația din care s-a extras eşantionul cu un prag de semnificație α=0,05, considerând că glicemia este normal distribuită şi pentru acest prag Z = 1,96.

- n = 121 $\quad \overline{X} = 105$
- $s^2 = 36$
- s = 6

$$\left[105 - 1{,}96\frac{6}{\sqrt{121}}; 105 + 1{,}96\frac{6}{\sqrt{121}}\right]$$

- [105-1.07, 105+1.07]
- [103.93 – 106.07]
- [104-106]

Compararea mediilor cu ajutorul intervalului de încredere



Intervalul de încredere pentru frecvențe

- Dacă $n \cdot p > 10$

$$\left[f - Z_\alpha\sqrt{\frac{f(1-f)}{n}}; f + Z_\alpha\sqrt{\frac{f(1-f)}{n}}\right]$$

Intervalul de încredere pentru frecvențe

- Suntem interesați în estimarea frecvenței cancerului de sân la femeile între 50 și 54 de ani care au antecedente familiale pozitive. Într-un studiu randomizat la care au participat 10000 de femei, s-a constatat că 400 dintre acestea au fost diagnosticate cu cancer de sân.
- Care este intervalul de încredere de 95% asociat frecvenței observate?

- f = 400/10000 = 0.04

$$\left[0{,}04-1{,}96\sqrt{\frac{0{,}04\cdot 0{,}96}{10000}};\ 0{,}04+1{,}96\sqrt{\frac{0{,}04\cdot 0{,}96}{10000}}\right]$$

- [0,04-0,004; 0,04+0,004]
- [0,036; 0,044]

$$\left[f-Z_\alpha\sqrt{\frac{f(1-f)}{n}};\ f+Z_\alpha\sqrt{\frac{f(1-f)}{n}}\right]$$

De reținut!

- Estimarea corectă a unui parametru statistic se face cu ajutorul intervalului de încredere.
- Intervalul de încredere depinde de volumul eșantionului și de eroarea standard.
- Cu cât eroarea standard este mai mare cu atât intervalul de încredere este mai larg.
- Cu cât volumul eșantionului este mai mic cu atât intervalul de încredere este mai larg.

Inferența statistică

Cuprins

- Definiție, aplicabilitate
- Ipoteza statistică versus ipoteza clinică
- Testarea unei ipoteze statistice:
  - Etapele unui test statistic

Definiție, aplicabilitate

- Un test statistic este conceput și utilizat pentru verificarea unei ipoteze statistice.
- De regulă, ipoteza care trebuie testată (H_0, ipoteza nulă) se poate formula ca fiind una în care nu există nici o schimbare:
  - Nu există nici o diferență între mediile a două populații (media taliei la o populație de nou-născuți la termen și respectiv născuți prematur)
  - Nu există diferență semnificativă între mediile a două eșantioane extrase din aceste populații.

Termeni

- Ipoteza nulă (H_0): ipoteza care urmează a fi testată
- Ipoteza alternativă (H_1): opusul ipotezei nule
- Prag de semnificație:
  - Probabilitatea de eroare acceptată de cercetător
  - De obicei este de 5% (0,05)

Testul statistic

- Metodă de comparație a două sau mai multe populații, prin intermediul unor variabile observate ale lor.

Ipoteza statistică vs ipoteza clinică

- Scopul unui test statistic este de a defini realitatea.
- Definirea întrebării de cercetare (ipoteza clinică):
  - Tratamentul cu Nebivolol este la fel de eficient ca şi cel cu Valsartan în tratamentul hipertensiunii arteriale?
- Transpunerea întrebării de cercetare în termeni statistici (ipoteza statistică):
  - Media tensiunii arteriale a pacienților tratați cu Valsartan nu diferă semnificativ de media tensiunii arteriale a pacienților tratați cu Nebivolol

Etapele unui test statistic

1. Formularea problemei în termenii ipotezelor statistice.
2. Alegerea şi calcularea parametrului statistic al testului.
3. Regiunea critică.
4. Concluzia testului.

1. Formularea problemei în termenii ipotezelor statistice

- **Ipoteza nulă**: ipoteza care trebuie testată, testul efectuându-se sub prezumția că ipoteza nulă ar fi adevărată
- **Ipoteza alternativă**: acea ipoteză care într-un sens sau altul contrazice ipoteza nulă. Această ipoteză se mai numeşte şi ipoteza de lucru

1. Formularea problemei în termenii ipotezelor statistice

- **Ipoteza nulă**: tipuri
  - O coadă ("one-tailed" sau "one-side"):
    - Media este mai mare
    - Media este mai mică
  - Două cozi ("one-tailed" sau "one-side"):
    - Media este egală

2. Alegerea şi calcularea parametrului statistic al testului

- Parametrul statistic al testului exprimă într-o anumită formă, diferența dintre elementele comparate.
- Ținând seama de faptul că eşantionul sau eşantioanele utilizate sunt aleator extrase din populațiile care fac obiectul testului, parametrul statistic este o variabilă aleatoare de selecție, care urmează o anumită lege de probabilitate.

2. Alegerea şi calcularea parametrului statistic al testului

- Un parametru statistic al testului bun trebuie să îndeplinească două condiții:
  - Trebuie să se comporte diferit atunci când ipoteza nulă H_0 este adevărată față de situația în care ipoteza alternativă H_1 este adevărată.
  - Distribuția de probabilitate a parametrului statistic al testului sub prezumția că H_0 este adevărată, este cunoscută.

3. Regiunea critică

- Trebuie să fim capabili să decidem în funcție de valoarea parametrului statistic calculat care dintre ipoteze, cea nulă sau cea alternativă, este adevărată.
- **Dacă valoarea parametrului statistic aparține regiunii critice, ipoteza nulă H_0 va fi respinsă şi va fi acceptată ipoteza alternativă H_1.**
- **Dacă valoarea parametrului statistic nu aparține regiunii critice, ipoteza nulă H_0 va fi acceptată.**

3. Regiunea critică

- Decidem mărimea regiunii critice.
  - Pentru aceasta trebuie să specificăm mărimea riscului de eroare pe care îl acceptăm.
  - Pe scurt, definim nivelul de semnificație, notat cu α, sau mărimea riscului pe care suntem dispuşi să ni-l asumăm în respingerea ipotezei nule H_0 în cazul în care aceasta este adevărată. De obicei se alege un nivel de semnificație între 1% şi 5%.

3. Regiunea critică

- Decidem mărimea regiunii critice.
  - Probabilitatea unei erori de tipul I:
    - probabilitatea de respingere a ipotezei nule H_0 în favoarea ipotezei alternative H_1, în condițiile în care H_0 este adevărată.
    - probabilitatea unei erori de tipul I se notează cu α şi se mai numeşte nivel de semnificație al testului.

3. Regiunea critică

- Decidem mărimea regiunii critice.
  - Probabilitatea unei erori de tipul II:
    - probabilitatea acceptării ipotezei nule în condițiile în care ipoteza alternativă H_1 este adevărată.
    - această probabilitate se notează cu β.

3. Regiunea critică

- unilaterală la dreapta – valoarea parametrului statistic al testului este mai mare sau egală cu valoarea din dreapta a intervalului critic;
- unilaterală la stânga – valoarea parametrului statistic al testului este mai mică sau egală cu valoarea din stânga a intervalului critic;
- bilaterală – valoarea parametrului statistic al testului este mai mică sau egală cu valoarea extremă din stânga regiunii critice sau mai mare sau egală cu valoarea extremă din dreapta regiunii critice, valorile extreme ale regiunii critice având nivele egale de semnificație.

4. Concluzia testului

- Ipoteza nulă H_0 este respinsă dacă valoarea parametrului statistic aparţine regiunii critice.
- Regiunea critică trebuie astfel aleasă încât dacă ipoteza alternativă H_1 este adevărată, probabilitatea de respingere a ipotezei nule H_0 este mai mare decât în cazul în care ipoteza nulă H_0 ar fi adevărată.

4. Concluzia testului

- Acceptarea ipotezei nule H_0 atunci când ipoteza alternativă H_1 este adevărată, este cunoscută ca şi eroarea de tipul II.
  - probabilitatea ei se notează cu β
  - măsoară „nivelul de eroare”

4. Concluzia testului

- În testarea oricărei ipoteze statistice, există patru situaţii care determină dacă decizia noastră este corectă sau nu

		Cazuri	
		H_0 este adevărată	**H_0 este falsă**
Concluzie	**H_0 se acceptă**	**decizie corectă**	**eroare de tipul II**
	H_0 se respinge	**eroare de tipul I**	**decizie corectă**

Luarea deciziei pe baza valorii probabilităţii p de semnificaţie a testului

- În momentul în care prelucrăm statistic o serie de date dorim să ştim dacă rezultatele obţinute sunt sau nu semnificative statistic.
- Răspunsul la această întrebare este dat de valoarea lui p calculată de orice program statistic la prelucrarea unor date.
- În cazul testelor statistice, ipoteza nulă este respinsă dacă nivelul de semnificaţie este mai mic decât 0,05 iar programele de prelucrare statistică a datelor vor afişa o steluţă (*) în tabelul rezultatelor.

Luarea deciziei pe baza valorii probabilităţii p de semnificaţie a testului

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	TASmz - TASmz6	31.000	9.539	5.508	7.303	54.697	5.629	2	.030
Pair 2	TASmz - TASmz12	37.667	2.517	1.453	31.415	43.918	25.924	2	.001
Pair 3	TASmz6 - TASmz12	6.667	7.371	4.256	-11.644	24.978	1.567	2	.258
Pair 4	TADmz - TADmz6	10.333	.577	.333	8.899	11.768	31.000	2	.001
Pair 5	TADmz - TADmz12	13.333	.577	.333	11.899	14.768	40.000	2	.001
Pair 7	TASmn - TASmn6	23.667	13.577	7.839	-10.060	57.394	3.019	2	.094
Pair 8	TASmn - TASmn12	26.000	14.000	8.083	-8.778	60.778	3.217	2	.085
Pair 9	TASmn6 - TASmn12	2.333	1.528	.882	-1.461	6.128	2.646	2	.118
Pair 10	TADmn - TADmn6	10.667	5.033	2.906	-1.837	23.170	3.671	2	.067
Pair 11	TADmn - TADmn12	17.000	7.211	4.163	-.913	34.913	4.083	2	.055
Pair 12	TADmn6 - TADmn12	6.333	2.309	1.333	.596	12.070	4.750	2	.042
Pair 13	TASm24 - TASm24-6	22.000	13.856	8.000	-12.421	56.421	2.750	2	.111
Pair 14	TASm24 - TASm24-12	24.667	15.373	8.876	-13.522	62.856	2.779	2	.109
Pair 15	TASm24-6 - TASm24-12	2.667	2.082	1.202	-2.504	7.838	2.219	2	.157
Pair 16	TADm24 - TADm24-6	12.667	6.429	3.712	-3.304	28.637	3.413	2	.076
Pair 17	TADm24 - TADm24-12	18.667	7.024	4.055	1.219	36.115	4.603	2	.044
Pair 18	TADm24-6 - TADm24-12	6.000	4.000	2.309	-3.937	15.937	2.598	2	.122

Luarea deciziei pe baza valorii probabilității p de semnificație a testului

- Dacă $p \leq 0,05$: respingem ipoteza nulă şi acceptăm ipoteza alternativă (am obținut semnificația statistică)
- Dacă $p > 0,05$: acceptăm ipoteza nulă (nu am obținut semnificația statistică)

Luarea deciziei pe baza valorii probabilității p de semnificație a testului

p = 0,13
NU respingem ipoteza nulă
Risc de eroare de tip II
α = 0,05
p = 0,02
Respingem ipoteza nulă
Risc de eroare de tip I

Semnificația lui p

- Criteriu de luare a deciziei cu privire la o ipoteză statistică nulă
- Cuantifică şansa ca o decizie de respingere a ipotezei nule să fie greşită
- Măsură a semnificației statistice şi NU CLINICĂ

Semnificația lui p: reguli empirice

- **$0,01 \leq p < 0,05$**: rezultatul e semnificativ statistic
- **$0,001 \leq p < 0,01$**: rezultatul e înalt semnificativ statistic
- **$p < 0,001$**: rezultatul e foarte înalt semnificativ statistic
- **$p \geq 0,05$**: rezultatul e considerat nesemnificativ statistic

Limite ale valorii p

- Valoarea *p* NU ne dă informații despre:
  - Şansa de beneficiu a unui pacient individual
  - Procentul de pacienți care vor avea un beneficiu în urma instituirii procedurii medicale
  - Gradul de beneficiu expectat pentru un anumit pacient

Puterea unui test statistic

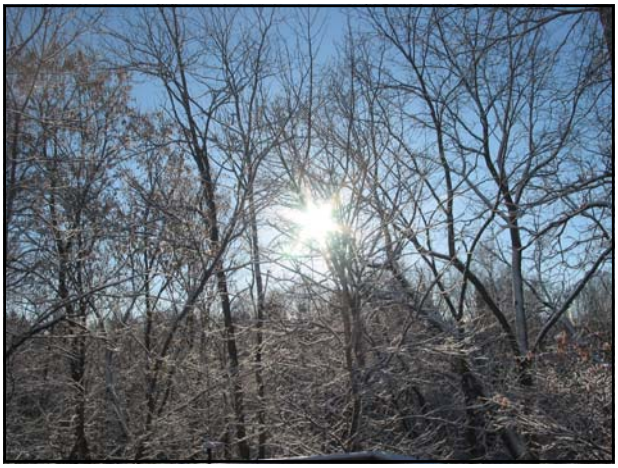
- Este capacitatea de a detecta o diferență acolo unde există
- Creşterea volumului eşantionului determină creşterea puterii testului statistic aplicat
- Valoarea este în relație directă cu eroarea de tip II:
  - Puterea = $1 - \beta$
- Cea mai utilizată modalitate de creştere a puterii unui test statistic este de a creşte volumul eşantionului

Tipul scalei de măsură – testul statistic

Denumire test	Interval	Nominal	Observații
Corelație Pearson	2	0	Există o relație liniară?
χ^2	0	2	
Student	1	1	Doar 2 grupuri
ANOVA	1	1	2 sau mai multe grupuri
Student perechi	1	1	Eşantioane perechi
Măsurători repetate (ANOVA)	1	1	Mai mult de 2 grupuri, date perechi

De reţinut!

- Paşii testului statistic sunt identici atât pentru testele parametrice cât şi pentru cele non-parametrice.
- Orice test statistic se poate interpreta din perspectiva valorii critice sau a intervalului critic şi respectiv din perspectiva valorii p.
- Orice test statistic are asociat 2 tipuri de erori. Fiecare tip de eroare are o anumită semnificaţie.
- Puterea unui test statistic este în relaţie cu eroarea de tip II şi depinde de volumul eşantionului.



Testarea distribuţiei unui set de date

Cuprins

- Obiective
- Testarea normalităţii unei distribuţii
- Testarea egalităţii a două distribuţii

Obiective

- Datele urmează o distribuţie normală?
  - Se poate aplica şi pe alte tipuri de distribuţii (Binomială, Poisson, etc.)
- Două distribuţii au aceeaşi formă?
  - Nu răspunde la întrebarea: "Care este formă de distribuţie a datelor?"
  - Ne spune dacă formele de distribuţie a două seturi de date sunt sau nu diferite.

Testarea normalității unei distribuții

- De ce normalitate?
  - Este o condiție preliminară de aplicare a unor teste statistice (test parametric vs test non-parametric)
  - Teste parametrice: aplicate pe date care urmează o distribuție normală:
    - Testul t
    - Testul z
    - Analiza variantei

Teste de normalitate

- Chi-Square goodness-of-fit: 1900
  - conservativ
- Kolmogorov-Smirnov (abreviere KS): 1933
  - conservativ
- Shapiro-Wilk: 1965

Teste de normalitate

	Test mai puțin conservativ	Test conservativ
Eşantion mic (5-50)	**Shapiro-Wilk**	**Kolmogorov-Smirnov**
Eşantion mare (> 50)	**Shapiro-Wilk**	**Chi-Square Goodness-of-Fit**

Kolmogorov-Smirnov: un eşantion

- Forma distribuției datelor e normală (nu facem asumpții asupra mediei sau deviației standard)?
  - H_0: forma distribuției este normală
- Eşantionul a fost extras dintr-o populație normal distribuită (facem asumpții asupra medie şi a deviație standard)?
  - H_0: forma distribuției populației din care a fost extras eşantionul nu este diferită de o distribuție normală specificată (de o anumită medie şi deviație standard)

Kolmogorov-Smirnov: un eşantion

- Valorile critice pentru diferite valori ale nivelului de semnificație:
  - $\alpha = 0{,}01$: $(1{,}63/\sqrt{n})-(1/3{,}5\cdot n)$
  - $\alpha = 0{,}05$: $(1{,}36/\sqrt{n})-(1/4{,}5\cdot n)$
  - $\alpha = 0{,}10$: $(1{,}22/\sqrt{n})-(1/5{,}5\cdot n)$

1. Aranjăm valorile eşantionului în ordine crescătoare
2. x este valoarea eşantionului la care datele se modifică

Kolmogorov-Smirnov: un eşantion

3. Fie k numărul de membrii cu valori mai mici de x
4. Fie $F_n(x) = k/n$ (calculat pentru fiecare valoare a lui x)
5. Valoarea expectată pentru fiecare x este dată de formula: $z=(x-m)/s$
6. Pentru fiecare z calculăm valoarea expectată $F_e(n)$ dată de aria de sub curba normală de la dreapta lui z (e nevoie de program sau de tabel standard).
7. Calculăm diferența absolută $| F_n(x) - F_e(n) |$
8. Testul statistic L este dat de cea mai mare valoarea a diferenței
9. Dacă L este mai mare decât valoarea critică se respinge ipoteza nulă (H_0).

Kolmogorov-Smirnov: un eşantion

- Variabila de interes: vârsta a zece pacienți internați cu infarct miocardic
- Valoarea critică:
- $\alpha = 0{,}05$: $(1{,}36/\sqrt{10})-(1/4{,}5\cdot 10) = 0{,}408$

Kolmogorov-Smirnov: un eşantion

Vârsta (ani)	x	k	$F_0(x)$	z	$F_e(x)$	$\lvert F_0(x)-F_e(x)\rvert$
54						
61	61	1	=1/10=0,1	(61-65,1)/7=-0,586	0,279	0,179
61						
61						
62	62	4	=4/10=0,4	(62-65,1)/7=-0,443	0,329	0,071
62						
68	68	6	=6/10=0,6	(68-65,1)/7=0,414	0,661	0,061
73	73	7	=7/10=0,7	(73-65,1)/7=1,129	0,871	0,171
74	74	8	=8/10=0,8	(74-65,1)/7=1,271	0,898	0,098
75	75	9	=9/10=0,9	(75-65,1)/7=1,414	0,921	0,021

m =(54+61+61+61+62+62+68+73+74+75)/10 = 651/10 = 65,1
s = 7
L = 0,171
1,171 ≤ 0,408: acceptăm ipoteza nulă, Distribuția populației din care s-a extras eșantionul nu este diferită de distribuția normală,

Chi-Square goodness-of-fit

- H_0: populația din care a fost extras eşantionul este normal distribuită
- Dacă valoarea calculată a testului e mai mare decât valoarea critică: respingem ipoteza nulă (H_0)

$$\chi^2 = \sum \frac{(n_i - e_i)^2}{e_i} = \sum \frac{n_i^2}{e_i}$$

α (1 coadă)	0,10	0,05	0,025	0,01
df = 6	10,64	12,59	14,45	16,81
df = 8	13,36	15,51	17,53	20,09
df = 9	14,68	16,92	19,02	21,67

Chi-Square goodness-of-fit

- $\chi^2 = 7{,}895 < 16{,}92$: acceptăm ipoteza nulă

Interval	Valoarea z standard la capătul intervalului	p	Frecvența expectată (e_i)	Frecvența observată (n_i)
< 50	-2,0691	0,0195	5,87	3
50 - < 55	-1,4519	0,0542	16,31	17
55 - < 60	-0,8346	0,1286	38,71	32
60 - < 65	-0,2473	0,2120	63,81	74
65 - < 70	0,4000	0,2407	72,45	62
70 - < 75	1,0173	0,1900	57,19	57
75 - < 80	1,6346	0,1035	31,15	37
80 - < 85	2,2519	0,0359	10,81	14
85 - < 90	2,8691	0,0145	4,36	4
≥ 90	∞	0,0019	0,57	1

De reținut!

- Normalitatea datelor trebuie testată pentru a aplica corect un test statistic.
- Testele de normalitate se fac cu ajutorul programelor (SPSS, Statistica, etc.)
- Trebuie să ştim să interpretăm un test de normalitate atât din perspectiva regiunii critice cât şi din cea a valorii p.

Analiza corelațiilor

Cuprins

- Corelația
- Semnificația corelației
- Tipuri de coeficienți de corelație
- Regresia liniară simplă
- Regresia liniară multiplă

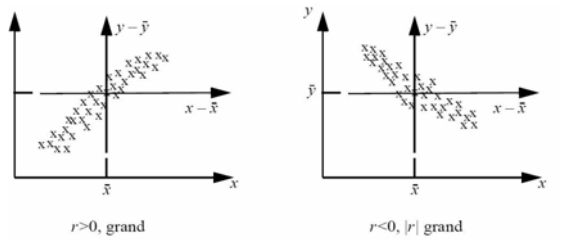
Corelație vs regresie

- Se folosesc pentru:
  - Evaluarea puterii de asociere dintre două variabile cantitative continue —› corelație
  - Prezicerea unei variabile (Y) în funcție de o altă variabilă (X) —› regresie

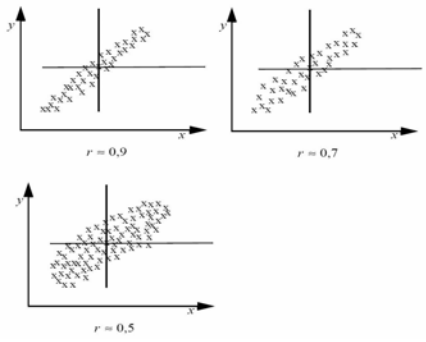
Coeficient de corelație

- Puterea asocierii dintre două variabile prin măsurarea gradului în care punctele unui grafic de tip scatter (nor de puncte) se întind de-a lungul unei linii.
  - *Să se stabilească dacă există o legătură între variabilele X şi Y (cantitative continue) şi să se determine o modalitate de a măsura intensitatea acestei legături.*
    - *Coeficientul de corelatie*

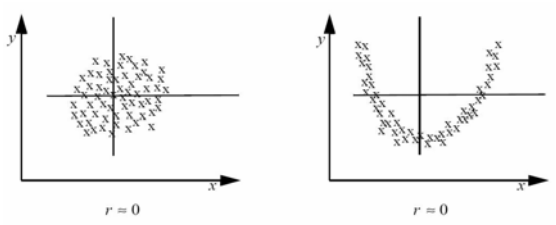
Coeficient de corelație



Coeficient de corelație



Coeficient de corelație



Tipuri de coeficienți de corelație

- Pearson: 2 variabile cantitative continue (relație de liniaritate, variabile normal distribuite)
- Spearman: 2 variabile cantitative (relație de non-liniaritate sau date nedistribuite normal); 1 variabilă calitativă + 1 variabilă cantitativă
- Kendall tau a, b, şi c: similar cu Spearman
- Gamma: Similar cu Spearman

Corelația Pearson (r)

- Scop: cuantifică puterea şi direcția legăturii liniare dintre două variabile prin descrierea direcției şi a gradului în care o variabilă este în relație de liniaritate cu cealaltă variabilă de interes (Pearson, 1896).
- Condiții de aplicare:
  - Ambele caractere sunt de tip interval sau rație
  - Ambele variabile urmează o distribuție normală şi distribuția lor comună este bivariată normală

Corelația Pearson (r)

- H_0: coeficientul de corelație = 0
- H_1: coeficientul de corelație ≠ 0
- Testul statistic aplicat pentru obținerea semnificației coeficientului de corelație: Student

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}}$$

- unde X, Y = valori ale caracterului pentru fiecare măsurătoare i (i = 1, 2, …, n); X_m, Y_m = medii ale măsurătorilor celor două caractere.

Interpretarea coeficientului de corelație

- $r \subset [-1, +1]$, $r = +1 \rightarrow$ există o relație de liniaritate între cele două caractere; $r = -1 \rightarrow$ există o relație inversă de liniaritate între cele două caractere.
- Clasificarea (regulile) lui Colton (Colton, 1974):
  - $r \subset [-0.25, +0.25] \rightarrow$ nu există relație
  - $r \subset (0.25, +0.50] \cup (-0.25, -0.50] \rightarrow$ relație slabă
  - $r \subset (0.50, +0.75] \cup (-0.50, -0.75] \rightarrow$ relație moderată
  - $r \subset (0.75, +1) \cup (-0.75, -1) \rightarrow$ relație foarte bună

Coeficientul de corelație al rangurilor Spearman (ρ)

- Scop:
  - Măsură non-parametrică de cuantificare a relației dintre două caractere (evaluează cât de bine o funcție monotonă poate descrie relație dintre cele două caractere) (Spearman, 1904).
- Metoda este satisfăcătoare pentru testarea ipotezei nule (nu există relație între cele două caractere) dar nu se recomandă ca şi instrument de cuantificare a relației (Bland, 1995).

Coeficientul de corelație al rangurilor Spearman (ρ)

- Condiții de aplicare:
  - Nu necesită nici un fel de asumpție asupra distribuției de frecvență a măsurătorilor;
  - Nu necesită asumpția relației de liniaritate dintre caractere;
  - Caracterele nu trebuie să fie cantitative de tip rație sau interval.
- Testul statistic aplicat pentru obținerea semnificației coeficientului de corelație:
  - Testul Student

Coeficientul de corelație al rangurilor Spearman (ρ)

- unde R_X, R_Y = rangurile atribuite valorilor măsurate ale caracterelor; R_{Xm}, R_{Ym} = media rangurilor asociate celor două caractere
- unde D = diferența dintre două perechi de ranguri ($R_X - R_Y$); n = volumul eşantionului

$$r = \frac{\sum (R_X - R_{\bar{X}})(R_Y - R_{\bar{Y}})}{\sqrt{\sum (R_X - R_{\bar{X}})^2 \sum (R_Y - R_{\bar{Y}})^2}}$$

$$\rho = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

Coeficienții de corelație Kendall tau (τ)

- Scop:
  - coeficienți de corelație non-parametrici utilizați pentru evaluarea şi testarea corelației dintre date non-interval ordinale (Kendall, 1938; 1942).
  - Este considerat a fi echivalent cu coeficientul de corelație al rangurilor Spearman.
- Se cunosc trei coeficienți de corelație notați τ_a, τ_b, şi τ_c.

Coeficientul de corelație Gamma (Γ)

- Scop: Metodă de determinare a coeficientului de corelație care în comparație cu Kendall e mai rezistent la existența perechilor de date cu ranguri egale (Goodman şi Kruskal, 1963); este utilizat când datele de analizat conțin multe date perechi cu ranguri egale (Siegel şi Castellan, 1988).
- Parametrul statistic:

$$\Gamma = (C-D)/(C+D)$$

- unde C = concordanță şi D = discordanța dintre perechile de caractere cantitative de interes.
- Testul statistic aplicat pentru obținerea semnificației coeficientului de corelație: Testul Z

Coeficienți de corelație: exemplu

Metoda	Valoarea coeficientului	Valoarea parametrului statistic (p)	
Pearson (r)	r = 0,9986	t = 272	$3,13 \cdot 10^{-266}$
Spearman (ρ)	ρ = 0,9984	t = 256	$8,81 \cdot 10^{-261}$
Semi-cantitativ (r_{sQ})	r_{sQ} = 0,9985	t = 263	$2,01 \cdot 10^{-263}$
Kendall Tau-a (τ_a)	τ_a = 0,9724	Z = 21	$4,13 \cdot 10^{-97}$
Kendall tau-b (τ_b)	τ_b = 0,9724	Z = 21	$4,13 \cdot 10^{-97}$
Kendall Tau-b (τ_c)	τ_c = 0,9678	Z = 21	$3,34 \cdot 10^{-96}$
Gamma (Γ)	Γ = 0,9732	Z = 20	$3,22 \cdot 10^{-92}$

Testele statistice au fost aplicate pentru un prag de semnificație de 5%
t = parametrul statistic al testului Student; Z = parametrul testului Z

Coeficient de determinare r^2

- Măsura în care variația unei variabile poate fi explicată variației celei de a doua variabile
- Proporția prin care variația unei variabile poate fi explicată de relația liniară cu cealaltă variabilă.
  - Defineşte mărimea asocierii
  - Nu defineşte direcția asocierii

Coeficient de determinare r^2

- r^2=0 variaţia lui Y nu poate fi atribuită modificărilor lui X
- r^2=1 variaţia lui Y este atribuită relaţiei liniare dintre Y şi X
- când r este semnificativ statistic şi r^2 este semnificativ
  - Într-un studiu de asociere dintre psihoza indusă ce consumul de amfetamine şi nivelul plasmatic de amfetamine s-a determinat un r=0,94 —› r^2=0,94^2 = 0.8836.
  - —› 88% din variaţia psihozei poate fi atribuită variaţiei nivelului plasmatic al amfetaminei.

De reţinut! Coeficientul de corelaţie

- Coeficientul de corelaţie:
  - Identificarea legăturii dintre două variabile
  - Cuantificarea legăturii
  - Direcţia legăturii
- Coeficientul de determinare: este pătratul coeficientului de corelaţie

Regresia

- Să se stabilească dacă Y depinde de X şi dacă da în ce formă se realizează această dependenţă.
  - Funcţia de regresie

Regresii liniare

- Descrie relaţia dintre două variabile
- Este posibilă determinarea unei variabile dependente în funcţie de o variabilă independentă folosind ecuaţia:

$$Y = a + bX$$

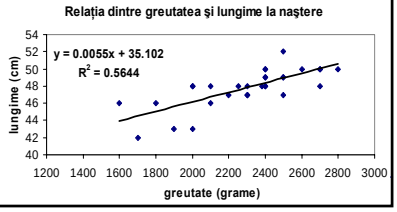
Y= variabila dependentă
X= variabila independentă
a = deplasarea de origine pe axa 0Y
b = panta liniei de regresie

Regresii liniare

- X = variabila independentă
- Y = variabila dependentă
  - Epidemiologie: variabila independentă = factor de risc, variabila dependentă = apariţia unei anumite patologii
  - Studii experimentale: variabilele independente sunt fixate de cercetător (doze ale unui nou medicament)

Diagrama scatter

- Alegerea axelor:
  - Variabile observate —› alegere arbitrară
  - Regresia este folosită pentru a prezice o variabilă în funcţie de alta —› variabila care trebuie prezisă (variabila dependentă) va fi pe axa 0Y



Regresii multiple

- Două tehnici:
  - Regresii lineare multiple
  - Regresii logistice multiple
- Relația dintre o variabilă dependentă şi una sau mai multe variabile independente
  - Presiunea arterială ‹— vârstă, greutate, fumat, antecedente heredo-colaterale
- Variabila dependentă = continuă, normal distribuită ... dacă nu se foloseşte regresia logistică

Regresii multiple

- Regresia logistică:
  - Variabila de răspuns nu este normal distribuită —› transformarea măsurii răspunsului în rata şansei (probabilitatea de a avea boala/probabilitatea de a fi indemn la boală) —› logaritmare
  - Permite prezicerea probabilității ca o patologie să apară (cancer pulmonar) folosind mai multe variabile ca predictori (fumatul, vârsta, sexul)

Indicatori statistici în evaluarea modelelor de regresie

- Coeficientul de corelație (r): exprimă cantitativ puterea relației liniare dintre activitatea de interes şi variabila sau variabilele independente
- Coeficientul de determinare (r^2): pătratul coeficientului de corelație.
  - Cuantifică proporția variației lui Y care poate fi explicată de relația de liniaritate dintre aceasta şi variabilele X.

Indicatori statistici în evaluarea modelelor de regresie

- Coeficientul de determinare ajustat: valoarea ajustată a coeficientului de determinare pentru numărul variabilelor independente din model (X).
  - Valoarea lui creşte dacă noul termen introdus în model determină o îmbunătățire a acestuia mai mare decât cea aşteptată prin şansa.
  - Spre deosebire de coeficientul de determinare care este un estimator pentru eşantion, coeficientul de determinare ajustat este un estimator pentru populație.
- Eroarea standard a estimatului: media erorii în estimarea lui Ŷ obținută pe baza ecuației de regresie.

Indicatori statistici în evaluarea modelelor de regresie

	Grade de libertate	∑ pătratelor	Media pătratelor	F	p
Regresia	1	5,83	5,83	46,62	4.21E-08
Reziduu	38	4,75	0,12		
Total	39	10,58			

$\hat{Y} = 1.43 - 2.49 \cdot 10^{-3} \cdot IPmrSMg$ (Ŷ – activitate de inhibiție a anhitrazei carbonice II)

- Ipoteza statistică în analiza de regresie: H_0: $b_1 = 0$ vs H_a: $b_1 \neq 0$. Dacă coeficienții asociați variabilelor independente sunt semnificativ diferiți de zero se poate concluziona că există o relație liniară semnificativă statistic între activitatea de interes şi descriptorii moleculari.
- Regresia: variația în activitate explicată de relație liniară dintre aceasta şi descriptorul sau descriptorii moleculari.
- Reziduu: variația în activitatea de interes care nu poate fi explicată de descriptorul sau descriptorii moleculari ai modelului.
- Total: variația totală a activității de interes.
- F: valoarea parametrului testului Fisher.
- p: semnificația parametrului F.

Indicatori statistici în evaluarea modelelor de regresie

	Valoare [IC95%]	Eroare standard	Parametrul t	p_t
b_0	1,43 [1,13 – 1,74]	0,1513	9,48	1.47E-11
X_1 (IPmrSMg)	-2,49E-03 [-3,23E-03 – -1,75E-03]	0,0004	-6,83	4.21E-08

Parametrul t = parametrul testului Student
p_t = semnificația parametrului Student

- Parametrul t: testează ipoteza diferenței semnificativ statistic de zero a valorilor asociate lui b_0 şi b_1.

Exemplu 1

S-a studiat scăderea numărului de neuroni odată cu înaintarea în vârstă, prin examinarea creierului unui eşantion de 38 pacienţi cu vârste cuprinse între 13 - 101 ani care au decedat fără nici un fel de istoric de boală sau demenţă. S-au numărat neuronii din hipocamp, pe mai multe secţiuni ale fiecărei regiuni a hipocampusului. Cercetătorii implicaţi în numărarea neuronilor nu cunoşteau vârstele pacienţilor.

Exemplu 1

S-au obţinut următoarele rezultate*:

Subdiviziuni hipocamp	Panta număr neuroni/vârstă	p
Dental granule cell layer	-54000	0,1700
Dental hilus	-9000	0,0120
Pyramidal cell layer CA3-2	-6000	0,1800
Pyramidal cell layer CA1	-29000	0,2600
Subiculum	-36000	0,0013

* West et al. 1994

Exemplu 1

1. Ce se înţelege prin panta număr de neuroni versus vârstă?
   - Estimarea modificării numărului neuronilor pe an
   - Este panta liniei de regresie
   - Panta este negativă deoarece numărul de neuroni scade odată cu înaintarea în vârstă

Exemplu 1

2. De ce cercetătorii nu au cunoscut vârsta pacienţilor?
   - Cercetătorii se pot aştepta ca persoanele mai în etate să aibă mai puţini neuroni în comparaţie cu persoanele mai tinere. Astfel, dacă cercetătorii cunosc vârsta pacienţilor se aşteaptă ca numărul de neuroni să fie mai mic la persoanele mai în etate —› numărul de neuroni poate fi astfel subestimat

Exemplu 2

Greutatea la naştere a unui eşantion de 1333 bărbaţi suedezi în vârstă de 50 ani a fost extrasă din registrele de evidenţă a naşterilor. S-a descoperit o corelaţie semnificativă între vârsta la naştere şi înălţimea acestor persoane adulte ($r = 0.22$, $p < 0.001$)

(Leon et al. 1996)

Exemplu 2

- Ce se înţelege prin "corelaţie" şi "r=0,22"?
  - Corelaţie pozitivă:
    - greutatea la naştere ↑ – înălţime ↑
    - greutatea la naştere ↓ – înălţime ↓
  - Corelaţie negativă:
    - greutatea la naştere ↑ – înălţime ↓
    - greutate la naştere ↓– înălţime ↑
  - r = coeficientul de corelaţie care măsoară puterea relaţiei liniare între cele două variabile continue
    - r = 0,22 —› corelaţia este pozitivă; înălţimea adulţilor tinde să fie mai mare pentru subiecţii cu greutate mai mare la naştere dar corelaţia este slabă.

Exemplu 2

- Ce concluzie putem trage din relația dintre înălțimea adultului şi greutatea la naştere?
  - Pentru populația din care acest eşantion a fost extras , înălțimea adultului este în relație cu greutatea la naştere, dar relația este slabă. Bărbații înalți se pare că au avut o greutate la naştere mai mare. Din aceste date nu putem trage concluzia că relația este una de cauzalitate.

De reținut! Regresia

- Se utilizează pentru a estima şi ulterior a prezice o variabilă în funcție de altă variabilă.
- Parametrii de interpretare ai modelului de regresie!
- Relația de liniaritate nu se întâlneşte frecvent în studiile medicale.